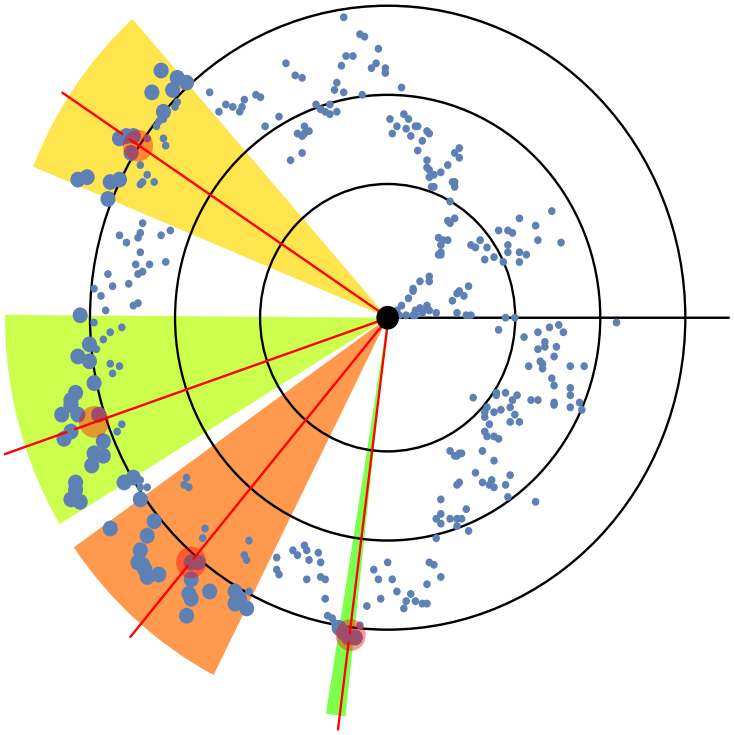
Sveučilište Josipa Jurja Strossmayera u Osijeku
Odjel za matematiku

Rudolf Scitovski Martina Briš Alić

G R U P I R A N J E P O D A T A K A



Osijek, 2016.

Prof. dr. sc. Rudolf Scitovski
Sveučilište u Osijeku
Odjel za matematiku
Trg Ljudevita Gaja 6
HR-31 000 Osijek

Doc. dr. sc. Martina Briš Alić
Ekonomski fakultet u Osijeku
Sveučilište u Osijeku
Trg Ljudevita Gaja 7
HR-31 000 Osijek

Izdavač:
Sveučilište Josipa Jurja Strossmayera u Osijeku – Odjel za matematiku

Recenzenti:

prof. dr. sc. Robert Manger, PMF–Matematički odsjek, Sveučilište u Zagrebu

prof. dr. sc. Kristian Sabo, Odjel za matematiku,
Sveučilište Josipa Jurja Strossmayera u Osijeku

prof. dr. sc. Marijana Zekić-Sušac, Ekonomski fakultet u Osijeku,
Sveučilište Josipa Jurja Strossmayera u Osijeku

Lektorica:
Ivanka Ferčec, Fakultet elektrotehnike, računarstva i informacijskih tehnologija
Osijek, Sveučilište Josipa Jurja Strossmayera u Osijeku

CIP zapis dostupan je u računalnom katalogu Gradske
i sveučilišne knjižnice Osijek pod brojem 140604093.

ISBN 978-953-253-141-1

Ovaj udžbenik objavljuje se uz suglasnost Senata Sveučilišta Josipa Jurja Strossmayera u Osijeku pod brojem 21/16.

Ovaj udžbenik objavljuje se uz financijsku pomoć Ministarstva znanosti, obrazovanja i sporta Republike Hrvatske.

Tisak: STUDIO HS Internet d.o.o., Osijek

PREDGOVOR

Veći dio sadržaja ovog udžbenika izvodi se u okviru predmeta *Algoritmi i strukture podataka* na sveučilišnom diplomskom studijskom programu *Poslovna informatika* Ekonomskog fakulteta u Osijeku. Tekst je napisan tako da podrazumijeva osnovna predznanja iz vektorskog računa koja su kratko navedena u *Dodatku* na kraju udžbenika. Udžbenik je namijenjen prvenstveno studentima diplomskih i poslijediplomskih studijskih programa društvenohumanističkog, ali i STEM područja. Pored stvaranja osnovnih pretpostavki za ispunjavanje odgovarajućih ishoda učenja spomenutih predmeta, sadržaj udžbenika može korisno poslužiti i u nekim praktičnim istraživanjima i samostalnom usavršavanju. U novije vrijeme razvojem *Big Data platformi* za upravljanje velikim količinama podataka i sve raspoloživijim programskim alatima za obradu tih podataka, kao najveće ograničenje u literaturi ističe se nedostatak obrazovanih stručnjaka koji će moći kvalitetno obraditi te podatke i protumačiti rezultate. Upravo ovaj udžbenik svojim gradivom doprinosi osposobljavanju takvih stručnjaka koji će moći raditi kao podatkovni analitičari ili podatkovni znanstvenici (eng. Data Scientists). Upravo ta zanimanja navedena su prema brojnim portalima kao najtraženija zanimanja budućnosti.

Opsežna recentna literatura, navedena na kraju udžbenika, koja obuhvaća brojne knjige i odgovarajuće članke iz renomiranih međunarodnih znanstvenih časopisa, daje neophodan pregled najvažnijih spoznaja, a također može poslužiti i za nastavak samostalnog rada u ovom znanstvenom području.

Pojmovi i metode koje se obrađuju u udžbeniku najprije su uvedene za najjednostavnije slučajeve, što je kasnije poslužilo kao motivacija za općenitije definiranje važnih pojmova i metoda. Tekst sadrži puno primjera koji mogu doprinijeti razumijevanju izložene materije. Također, u okviru svakog poglavlja nalaze se brojni zadaci: od sasvim jednostavnih, do onih koji mogu poslužiti kao teme seminarskih i sličnih radova. Gdje god je to moguće, dana su i rješenja zadataka, a kod nekih i odgovarajuće upute za rješavanje. Izrada većine zadataka povezana je s mogućnošću korištenja programskog sustava *Mathematica*, kao i *Mathematica*-programa navedenih u ovom udžbeniku. U tom smislu čitatelja se sustavno uvodi u mogućnosti ovog programskog sustava za koji Sveučilište u Osijeku već godinama raspolaže odgovarajućim brojem licenci.

Za sve važnije metode koje se navode u udžbeniku napisani su odgovarajući algoritmi, a na osnovi njih i odgovarajući *Mathematica*-moduli koji su navedeni u devetom poglavlju (programski kodovi dostupni su na `http://www.mathos.unios.hr/images/homepages/scitowsk/Programi.zip`). U brojnim primjerima i zadacima pokazano je korištenje tih modula. Na taj način, uz usvajanje osnovnih znanja, studenti imaju mogućnost proširiti svoje znanje najnovijim mogućnostima koje se nude u literaturi.

Na kraju udžbenika navodimo jedan praktični primjer sa stvarnim podacima o kretanju prosječnih dnevnih temperatura u Osijeku od 1985. godine. Pri tome nemamo ambiciju davati zaključke i prijedloge vezano uz klimatske procese, već samo opisati jednu mogućnost korištenja teorije i metoda navedenih u ovom udžbeniku.

Zbog specifičnih zahtjeva matematičkog teksta i sličnih zahtjeva u programskim sustavima `Mathematica, Matlab, FORTRAN`, koji se koriste u tekstu, cijeli dio od decimalnog dijela decimalnog broja odvajan je decimalnom točkom (.), a ne decimalnim zarezom (,).

Svi teoremi, leme, definicije, slike, tablice, primjedbe, primjeri i zadaci u tekstu imaju svoju jedinstvenu oznaku i na taj način pozivaju se u cijelom tekstu. Zbog toga su pisani velikim početnim slovom[1].

Zahvaljujemo recenzentima *Robertu Mangeru, Kristianu Sabi* i *Marijani Zekić-Sušac* te lektorici *Ivanki Ferčec*, koji su svojim primjedbama i prijedlozima značajno pomogli da ovaj tekst bude bolji. Također zahvaljujemo kolegi *Ivanu Soldi* koji je u tehničkom smislu pomogao podizanju kvalitete ovog udžbenika.

Osijek, 1. rujna, 2016.

Rudolf Scitovski
Martina Briš Alić

[1] Za pregledavanje ovog udžbenika Adobe Readerom možete koristiti sljedeće pogodnosti:

- klikom na naslov nekog poglavlja u Sadržaju dolazite na to poglavlje. Povratak (na isto mjesto odakle ste krenuli) je držanjem tipke Alt pa nakon toga klik na <;
- klikom na oznaku nekog teorema, leme, definicije, slike, tablice, primjedbe, primjera ili zadatka u tekstu odlazite na taj objekt. Povratak je na prethodno opisani način;
- klikom na oznaku neke reference u tekstu odlazite na tu referencu u Literaturi na kraju knjige. Povratak je na prethodno opisani način;
- klikom na oznaku stranice u Indeksu na kraju knjige odlazite na taj pojam u knjizi. Povratak je na prethodno opisani način;

Sadržaj

Poglavlje 1

Uvod

U ovom udžbeniku razmatra se problem prirodnog grupiranja podataka u dvije ili više logičkih skupina. Podaci mogu imati jedno ili više obilježja (atributa). U nekim situacijama broj obilježja može biti vrlo velik: u problemu prepoznavanja razdoblja toplih i hladnih dana u godini za neki duži vremenski period (koji se razmatra u t. 7.3, str. 134) podaci imaju 28 obilježja, u problemu ocjenjivanja kvalitete vina [49] (vidi također `https://archive.ics.uci.edu/ml/datasets/Wine`) broj obilježja kreće se od 13 do 30, a kod istraživanja ljudskog gena [47] taj broj penje se na nekoliko tisuća.

Matematički gledano, podaci se interpretiraju kao vektori u apstraktnom n-dimenzionalnom vektorskom prostoru. Zbog toga je važno poznavati osnove vektorskog računa, uključujući pojam norme vektora i pojam udaljenosti. Prilikom traženja optimalne particije s najprikladnijim brojem klastera koriste se različite statističke metode i metode numeričke matematike. Treba naglasiti da praktični problemi iz primjena obično podrazumijevaju izuzetno veliki broj podataka (od nekoliko stotina do više stotina tisuća) [22, 43, 55]. Za analizu i grupiranje ovakvih podataka potrebno je imati prilagođene metode i odgovarajući software.

U tom smislu studentima će se ukazati na grafičke i računske mogućnosti korištenja programskog sustava *Mathematica* kroz pisanje malih i sasvim jednostavnih programa ali i složenih programskih struktura.

Iako se problem grupiranja podataka u posljednje vrijeme pojavljuje gotovo u svim znanstvenim disciplinama, spomenimo samo nekoliko važnih primjena počevši od onih u ekonomskim istraživanjima:

- U poslovnoj praksi puno je slučajeva u kojima postoji potreba za grupiranjem podataka. U maloprodaji trgovački lanci imaju potrebu za

grupiranjem (segmentiranjem) svojih kupaca prema zajedničkim obilježjima, primjerice prema vrsti artikala koje kupuju, lojalnosti, imovinskom stanju, učestalosti kupovine i drugim obilježjima. U marketingu je često istraživanjem tržišta potrebno segmentirati potencijalne kupce, odnosno izdvojiti ih od onih za koje je velika vjerojatnost da to neće postati. Slični problemi pojavljuju se i kod analize kreditne sposobnosti [17, 19, 55, 67, 71];

- Donošenje raznih odluka u tijelima državne i lokalne uprave [17];
- Grupiranje i rangiranje projekata, institucija, sveučilišta [75];
- Određivanje najprihvatljivijeg broja izbornih jedinica, njihovog sastava i alokacije [38, 52];
- Predviđanje satne potrošnje energenata [55];
- Analiza i pretraživanje teksta, računalni pretraživači baza podataka [5, 31, 55];
- Detekcija seizmogenih zona i primjene u potresnom inženjerstvu [43, 63];
- Prepoznavanje opasnih mjesta na cesti [26, 54];
- Prepoznavanje oblika na medicinskim slikama [17, 22, 71];
- Prepoznavanje kružnica i elipsi na slici uz korištenje u prometu, medicini, robotici [22, 39, 60];
- Razumijevanje klimatskih kretanja i problema alokacije objekata [17, 71];
- Razne primjene u poljoprivredi, primjerice prepoznavanje redova zasijanja [80], razvrstavanje oranica prema plodnosti zemljišta [19, 30], klasifikacija kukaca u skupine [10, 19, 30, 71];
- Identifikacija fenotipske sličnosti vina i grožđa [49].

Poglavlje 2

Reprezentant

Često je u primijenjenim istraživanjima potrebno dani skup podataka reprezentirati jednim podatkom koji na neki način obuhvaća većinu svojstava promatranog skupa. Najčešće korištena veličina u tom smislu je dobro poznata aritmetička sredina podataka. Primjerice, prosječna ocjena svih položenih predmeta nekog studenta može se izraziti aritmetičkom sredinom, ali stopu ekonomskog rasta tijekom nekoliko godina ne bi bilo dobro predstaviti na takav način (vidi [64]).

U ovom poglavlju razmotrit ćemo dva najčešće korištena reprezentanta nekog skupa podataka: *aritmetičku sredinu* i *medijan skupa*.

2.1 Reprezentant podataka iz $\mathbb{R}$

Zadan je skup podataka $\mathcal{A} = \{y_1, y_2, \ldots, y_m\} \subset \mathbb{R}$. Treba odrediti realni broj $c^\star \in \mathbb{R}$ (reprezentant skupa $\mathcal{A}$) koji će što bolje reprezentirati skup podataka $\mathcal{A}$.

Definicija 2.1. Funkciju $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, koja ima svojstvo *pozitivne definitnosti*[1]

$$\begin{aligned} &\text{(i)} \quad d(x,y) \geq 0 \quad \forall x, y \in \mathbb{R}, \\ &\text{(ii)} \quad d(x,y) = 0 \quad \Leftrightarrow \quad x = y, \end{aligned}$$

zovemo kvazimetrička funkcija (funkcija sličnosti, funkcija različitosti).

[1] Koristimo sljedeće oznake:
$\mathbb{R}_+ = \{x \in \mathbb{R}\colon x \geq 0\}$ – skup svih nenegativnih realnih brojeva;
$\mathbb{R}_{++} = \{x \in \mathbb{R}\colon x > 0\}$ – skup svih pozitivnih realnih brojeva.

Primjer 2.1. *Dvije najčešće korištene kvazimetričke funkcije na* $\mathbb{R}$ *su* kvazimetrička funkcija najmanjih kvadrata *(engl.: Least Squares distance like function) i* ℓ_1*-metrička funkcija koja se u literaturi (vidi primjerice [17, 19, 30, 54]) često naziva* Manhattan metrička funkcija*:*

$$d_{LS}(x,y) = (x-y)^2 \qquad \textit{[Least Squares (LS) kvazimetrička funkcija]}$$
$$d_1(x,y) = |x-y| \qquad \textit{[}\ell_1\textit{-metrička funkcija (Manhattan metrika)]}$$

U Dodatku*, str. 143, definirane su najčešće korištene metričke funkcije:* d_1*,* d_2*,* d_∞ *i poopćena* d_p*,* $p \geq 1$ *Minkowsky udaljenost (vidi također [2, 13, 18, 32, 74]. Provjerite da na skupu realnih brojeva* $\mathbb{R}$ *vrijedi*

$$d_1(x,y) = d_2(x,y) = d_\infty(x,y) = d_p(x,y), \quad p \geq 1, \quad \forall x, y \in \mathbb{R},$$

gdje su d_1, d_2, d_∞, d_p *najčešće korištene metričke funkcije (vidi t. ,str. 158).*

Zadatak 2.1. Pokažite da funkcija d_{LS} iz prethodnog primjera nije metrika na $\mathbb{R}$, a da je funkcija d_1 metrika na $\mathbb{R}$.

Definicija 2.2. Neka je $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ kvazimetrička funkcija. Najbolji reprezentant skupa podataka $\mathcal{A} = \{y_1, y_2, \ldots, y_m\} \subset \mathbb{R}$ u odnosu na kvazimetričku funkciju d je točka $c^\star \in \mathbb{R}$ u kojoj se postiže globalni minimum funkcije $F\colon \mathbb{R} \to \mathbb{R}_+$,

$$F(x) = \sum_{i=1}^{m} d(x, y_i), \tag{2.1}$$

što formalno zapisujemo na sljedeći način

$$c^\star \in \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{i=1}^{m} d(x, y_i). \tag{2.2}$$

Primijetite da za točku globalnog minimuma $c^\star \in \mathbb{R}$ vrijedi

$$F(x) = \sum_{i=1}^{m} d(x, y_i) \geq \sum_{i=1}^{m} d(c^\star, y_i) = F(c^\star), \tag{2.3}$$

pri čemu jednakost vrijedi onda i samo onda ako je $x = c^\star$.

2.1.1 Najbolji LS-reprezentant

Za LS-kvazimetričku funkciju vrijedi (vidi primjerice [50, 59])

$$F_{LS}(x) := \sum_{i=1}^{m}(x - y_i)^2 \geq \sum_{i=1}^{m}(\bar{y} - y_i)^2, \tag{2.4}$$

gdje je $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ aritmetička sredina skupa $\mathcal{A}$. Zato je najbolji LS-reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}$ obična aritmetička sredina[2]

$$c^\star_{LS} = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^{m} d_{LS}(x, y_i) = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

Dakle, aritmetička sredina skupa podataka $\mathcal{A} \subset \mathbb{R}$ je broj koji ima svojstvo da je suma kvadrata odstupanja do svih podataka minimalna.

Kao mjera raspršenosti skupa podataka $\mathcal{A}$ oko aritmetičke sredine $c^\star_{LS}$ u statističkoj literaturi [3] koristi se varijanca podataka (prosječno kvadratno odstupanje)

$$s_m^2 = \frac{1}{m}\sum_{i=1}^{m}(c^\star_{LS} - y_i)^2. \tag{2.5}$$

Broj s_m zovemo standardna devijacija.

Primjer 2.2. *Zadan je skup podataka* $\mathcal{A} = \{2, 1.5, 1, 3, 10\}$. *Njegova aritmetička sredina je* $c^\star_{LS} = 3.5$.

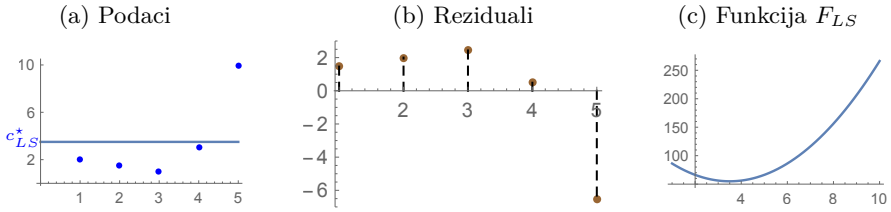
Na Slici 2.1a vidljivi su podaci i aritmetička sredina $c^\star_{LS}$, na Slici 2.1b tzv. „reziduali" podataka (brojevi $c^\star_{LS} - y_i$), a na Slici 2.1c graf funkcija F_{LS}. Primijetite da je njezin graf parabola i da je $F_{LS}(c^\star_{LS}) = 55$. Kolika je varijanca, a kolika standardna devijacija ovog skupa?

Primijetite također da među podacima ima i jedan „outlier" (jako stršeći podatak), koji je značajno utjecao na veličinu aritmetičke sredine. Kolika bi bila aritmetička sredina ovog skupa da umjesto podatka 10 stoji broj 2.5?

Zadatak 2.2. Neka je $\mathcal{A} = \{y_1, \ldots, y_m\} \subset \mathbb{R}$ skup podataka, a $c^\star_{LS}$ njegova aritmetička sredina. Pokažite da tada vrijedi

$$\sum_{i=1}^{m}(c^\star_{LS} - y_i) = 0.$$

[2] Problem traženja najboljeg LS-reprezentanta skupa podataka u literaturi se pojavljuje kao poznati princip najmanjih kvadrata, koji je 1795. godine postavio njemački matematičar Carl Friedrich Gauss (1777.–1855.) prilikom izučavanja kretanja nebeskih tijela, što je objavio 1809. godine u radu *Teoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, Perthes and Besser, Hamburg. Treba također reći da je 1805. godine francuski matematičar Adrien-Marie Legendre (1752-1833) prvi objavio algebarski postupak metode najmanjih kvadrata.



Slika 2.1: Aritmetička sredina skupa $\mathcal{A} = \{2, 1.5, 1, 3, 10\}$

Provjerite ovo svojstvo na podacima iz Primjera 2.2.

Zadatak 2.3. Neka su $\mathcal{A} = \{a_1, \ldots, a_p\}$, $\mathcal{B} = \{b_1, \ldots, b_q\} \subset \mathbb{R}$ dva skupa podataka, a $a^\star_{LS}$ i $b^\star_{LS}$ njihove aritmetičke sredine. Pokažite da je tada aritmetička sredina skupa $\mathcal{C} = \mathcal{A} \bigcup \mathcal{B}$ jednaka

$$c^\star_{LS} = \tfrac{p}{p+q} a^\star_{LS} + \tfrac{q}{p+q} b^\star_{LS}.$$

Provjerite ovu formulu na skupovima:

$$\mathcal{A} = \{1, 4, 6, 6, 8, 8, 8, 9, 10, 10\}, \quad \mathcal{B} = \{1, 1, 4, 4, 5, 6, 6, 7, 8, 8\}.$$

Kako bi glasila generalizacija ove formule za n skupova podataka $\mathcal{A}_1, \ldots, \mathcal{A}_n$ s po $p_1, \ldots, p_n$ elemenata?

2.1.2 Najbolji ℓ_1-reprezentant

Za ℓ_1-metričku funkciju vrijedi (vidi primjerice [50, 59])

$$F_1(x) := \sum_{i=1}^{m} |x - y_i| \geq \sum_{i=1}^{m} |\operatorname*{med}_k y_k - y_i|, \tag{2.6}$$

gdje je $\operatorname*{med}_k y_k$ medijan skupa $\mathcal{A}$. Dakle, najbolji reprezentant skupa $\mathcal{A} \subset \mathbb{R}$ u ovom je slučaju medijan skupa $\mathcal{A}$. To je broj koji ima svojstvo da je suma apsolutnih odstupanja do svih podataka minimalna.[3]

[3] Problem traženja najboljeg ℓ_1-reprezentanta skupa podataka u literaturi se pojavljuje kao poznati princip najmanje sume apsolutnih odstupanja i pripisuje se hrvatskom znanstveniku Josipu Ruđeru Boškoviću (1711.–1787.), koji je ovaj princip iznio još 1757. godine u radu [8]. Dugo je vremena ovaj princip zapostavljan u odnosu na Gaussov princip najmanjih kvadrata zbog složenosti računskih procesa. Tek dolaskom modernih računala ovaj princip ponovo je zauzeo važno mjesto u znanstvenim istraživanjima, posebno zbog svojstva svoje robusnosti: ovaj princip, za razliku od Gaussovog principa najmanjih kvadrata, u značajnoj mjeri ignorira jako stršeće podatke („outliers") u skupu podataka. U švicarskom gradu Neuchâtelu još uvijek se redovito održavaju znanstvene konferencije posvećene ℓ_1 metodama i primjenama, a na naslovnici zbornika radova nalazi se slika hrvatske novčanice s likom Josipa Ruđera Boškovića [16].

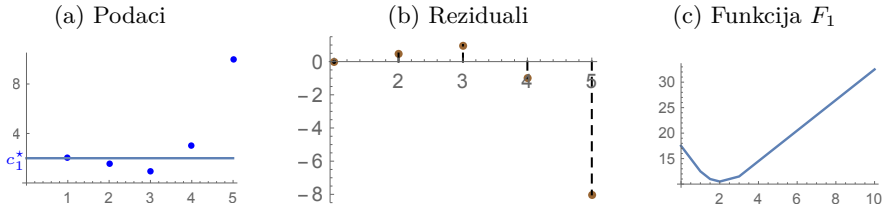
Medijan skupa $\mathcal{A}$ dobije se tako da se njegovi elementi najprije sortiraju. Tada, ako skup $\mathcal{A}$ ima neparan broj elemenata, medijan je srednji element, a ako skup $\mathcal{A}$ ima paran broj elemenata, medijan je bilo koji broj između dva srednja elementa. Primjerice[4],

$$\texttt{Median}\,\{3,1,4,5,9\} = 4,$$
$$\texttt{Median}\,\{-1,1,-2,2,-5,5,-9,9\} = [-1,1].$$

Primjedba 2.1. Primijetite da se medijan skupa podataka $\mathcal{A}$ uvijek može izabrati iz samog skupa $\mathcal{A}$. To znači da je medijan kao reprezentant skupa ujedno i element tog skupa, što nije slučaj kod aritmetičke sredine kao najboljeg LS-reprezentanta. To može biti korisno u nekim primjenama.

Primijetite također da se po 50% elemenata skupa $\mathcal{A}$ nalazi lijevo, odnosno desno od medijana skupa $\mathcal{A}$.

Primjer 2.3. *Zadan je skup podataka* $\mathcal{A} = \{2, 1.5, 1, 3, 10\}$. *Njegov je medijan* $c_1^\star = 2$.



Slika 2.2: Medijan skupa $\mathcal{A} = \{2, 1.5, 1, 3, 10\}$

Na Slici 2.2a vidljivi su podaci i medijan $c_1^\star$, na Slici 2.2b tzv. „reziduali" podataka (brojevi $c_1^\star - y_i$), a na Slici 2.2c graf funkcija F_1. Primijetite da je F_1 konveksna po dijelovima linearna funkcija i da je $F_1(c_1^\star) = 10.5$.

Primijetite također da među podacima ima i jedan „outlier", koji gotovo da i nije utjecao na veličinu medijana. Koliki bi bio medijan ovog skupa da umjesto podatka 10 stoji broj 1, a koliki da umjesto podatka 10 stoji broj 100? Usporedite ove rezultate s onima iz Primjera 2.2.

U statističkoj literaturi [3] uz pojam medijana vežu se i pojmovi donjeg kvartila (element skupa $\mathcal{A}$ koji se nalazi na mjestu 1/4 sortiranih podataka) i gornjeg kvartila (element skupa $\mathcal{A}$ koji se nalazi na mjestu 3/4 sortiranih podataka). Koliki bi bio donji i gornji kvartil skupa podataka iz prethodnog primjera?

[4]Medijan skupa $\mathcal{A}$ može se izračunati primjenom *Mathematica*–naredbe: `Median[]`. Pri tome, ako je medijan podataka interval, naredba `Median[]` daje polovište intervala

Zadatak 2.4. Neka je $\mathcal{A} = \{y_1, y_2, \ldots, y_m\} \subset \mathbb{R}$ skup podataka. Pokažite da funkcija F_{LS} postiže svoj minimum za $c_{LS}^\star = \frac{1}{m}\sum_{i=1}^m y_i$, a funkcija F_1 za $c_1^\star = \operatorname{med}_k y_k$.

2.1.3 Najbolji reprezentant skupa podataka s težinama

Ponekad je u praktičnim primjenama podacima potrebno dodijeliti težine (pondere). Na taj način svakom podatku pridružujemo faktor utjecaja ili učestalost njegova pojavljivanja. Primjerice, u Primjeru 2.13, str. 18, promatraju se podaci koji predstavljaju lokacije potresa koji su se dogodili u okolini Osijeka od 1880. godine, a svakom podatku pridružena je njegova težina: magnituda tog potresa.

Definicija 2.3. Neka je $d\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ kvazimetrička funkcija. Najbolji reprezentant skupa podataka $\mathcal{A} = \{y_1, y_2, \ldots, y_m\} \subset \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$ u odnosu na kvazimetričku funkciju d je točka $c^\star \in \mathbb{R}$ u kojoj se postiže globalni minimum funkcije $F\colon \mathbb{R} \to \mathbb{R}_+$,

$$F(x) = \sum_{i=1}^m w_i d(x, y_i), \tag{2.7}$$

što formalno zapisujemo na sljedeći način

$$c^\star \in \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^m w_i d(x, y_i). \tag{2.8}$$

Primijetite da za točku globalnog minimuma $c^\star \in \mathbb{R}$ vrijedi

$$F(x) = \sum_{i=1}^m w_i d(x, y_i) \geq \sum_{i=1}^m w_i d(c^\star, y_i) = F(c^\star), \tag{2.9}$$

pri čemu jednakost vrijedi onda i samo onda ako je $x = c^\star$.

Za LS-kvazimetričku funkciju najbolji reprezentant skupa $\mathcal{A} \subset \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$ je *težinska aritmetička sredina* [50]

$$c_{LS}^\star = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^m w_i d_{LS}(x, y_i) = \frac{1}{W}\sum_{i=1}^m w_i y_i, \quad W = \sum_{i=1}^m w_i,$$

a odgovarajuća minimizirajuća funkcija glasi

$$F_{LS}(x) = \sum_{i=1}^m w_i (x - y_i)^2.$$

Za ℓ_1-metričku funkciju najbolji reprezentant skupa podataka $\mathcal{A} = \{y_1, \ldots, y_m\}$ s težinama $w_1, \ldots, w_m > 0$ je *težinski medijan*

$$c_1^\star \in \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{i=1}^{m} w_i d_1(x, y_i) = \operatorname*{med}_i (w_i, y_i),$$

a odgovarajuća minimizirajuća funkcija glasi

$$F_1(x) = \sum_{i=1}^{m} w_i |x - y_i|.$$

Određivanje težinskog medijana u općem slučaju vrlo je složena numerička procedura [24, 50]. U stručnoj literaturi postoje brojni algoritmi za njegovo određivanje [24].[5]

Primjer 2.4. *Težinski medijan skupa $\mathcal{A} \subset \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$, u slučaju kada su težine w_i cijeli brojevi, određuje se slično kao i medijan skupa podataka bez težina. Medijan ovakvog skupa dobije se tako da najprije sortiramo elemente skupa $\mathcal{A}$ s odgovarajućom frekvencijom pojavljivanja i nakon toga odredimo srednji element. Primjerice, medijan skupa $\mathcal{A} = \{3, 1, 4, 5, 9\}$ s težinama $3, 1, 3, 2, 2$ je srednji element u nizu*

$$1, 3, 3, 3, 4, 4, 4, 5, 5, 9, 9.$$

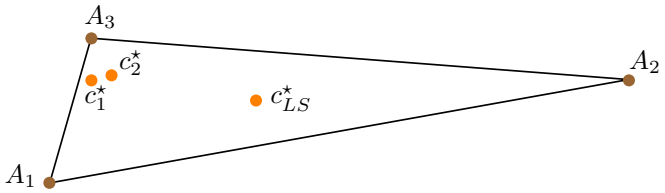
U ovom slučaju, $\operatorname{med} \mathcal{A} = 4$. *Koliki su donji i gornji kvartil skupa $\mathcal{A}$?*

2.2 Reprezentant podataka iz $\mathbb{R}^2$

2.2.1 Fermat–Torricelli–Weberov problem

Neka su $A_1, A_2, A_3 \in \mathbb{R}^2$ tri nekolinearne točke u ravnini (vidi Sliku 2.3). Točka $c^\star \in \mathbb{R}^2$, za koju je suma udaljenosti do vrhova trokuta minimalna, zove se centar skupa točaka A_1, A_2, A_3, a problem određivanja točke $c^\star \in \mathbb{R}^2$ u stručnoj literaturi naziva se Fermatov problem. Problem se može promatrati i u fizikalnom smislu (Torricellijev problem) ili u ekonometrijskom smislu (Weberov problem) [17].

[5] U programskom sustavu *Mathematica* težinska aritmetička sredina, odnosno težinski medijan, skupa podataka $\mathcal{A}$ s težinama $W : w_1, \ldots, w_m > 0$ izračunava se naredbama:
`Mean[WeightedData[A,W]];`
`Median[WeightedData[A,W]].`



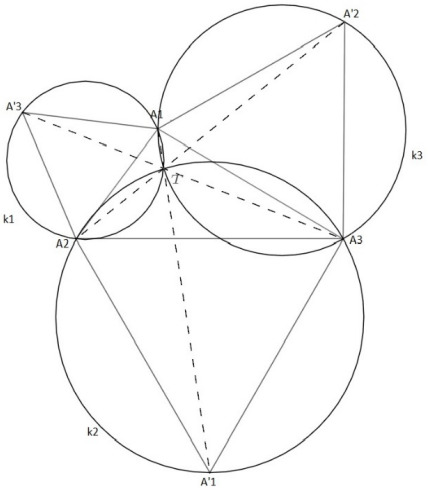
Slika 2.3: Fermat–Torricelli–Weberov problem ($c_2^\star$ - geometrijski medijan, $c_{LS}^\star$ - centroid, $c_1^\star$ - medijan)

Specijalno, točka $c_{LS}^\star \in \mathbb{R}^2$, za koju je suma LS-udaljenosti (suma kvadrata euklidskih udaljenosti) do vrhova trokuta minimalna, zove se centroid ili Steinerova točka (povezano s pojmom centra masa u fizici). Geometrijski gledano, to je težište trokuta koje se dobije na presjeku težišnica trokuta (spojnice vrhova trokuta s polovištima nasuprotnih stranica).

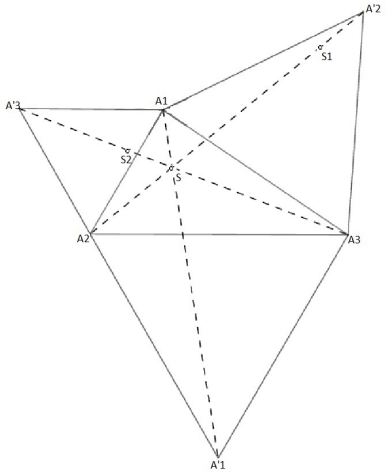
Točka $c_1^\star \in \mathbb{R}^2$, za koju je suma ℓ_1 udaljenosti do vrhova trokuta A_1, A_2, A_3 minimalna, zove se medijan skupa točaka $\{A_1, A_2, A_3\}$ (vidi Sliku 2.3).

Točka $c_2^\star \in \mathbb{R}^2$, za koju je suma ℓ_2 (euklidskih) udaljenosti do vrhova trokuta A_1, A_2, A_3 minimalna (vidi Sliku 2.3), zove se geometrijski medijan skupa točaka $\{A_1, A_2, A_3\}$. Geometrijski medijan trokuta može se dobiti [41] na presjeku tzv. Torricellijevih kružnica (vidi Sliku 2.4a) ili na presjeku tzv. Simpsonovih pravaca (vidi Sliku 2.4b).

(a) Torricellijeve kružnice



(b) Simpsonovi pravci



Slika 2.4: Fermat–Torricelli–Weberov problem: geometrijski medijan

Općenito može se promatrati konačni skup točaka iz $\mathbb{R}^n$ i proizvoljna kvazimetrička funkcija d. Problem određivanja najboljeg d-reprezentanta ima brojne primjene u najrazličitijim područjima kao primjerice: telekomunikacije (problem optimalnog antenskog pokrivanja, problem diskretne mreže), javni sektor (problem optimalnog pokrivanja), ekonomija (optimalna lokacija potrošačkih centara), problem lokacije hubova, robotika, problem optimalne asignacije, problem satne prognoze potrošnje energenata, itd. [17, 55, 57].

2.2.2 Centroid skupa točaka u ravnini

Neka je $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, m\} \subset \mathbb{R}^2$ skup točaka u ravnini[6]. Centroid $c^\star_{LS}$ skupa $\mathcal{A}$ rješenje je optimizacijskog problema

$$\operatorname*{argmin}_{T \in \mathbb{R}^2} \sum_{i=1}^{m} d_{LS}(T, a^i), \tag{2.10}$$

gdje je $d_{LS}(a, b) = d_2^2(a, b) = \|a - b\|_2^2$. Točka $c^\star_{LS}$ je točka u kojoj se postiže globalni minimum funkcije

$$F_{LS}(x, y) = \sum_{i=1}^{m} [(x - x_i)^2 + (y - y_i)^2].$$

Funkcija F_{LS} predstavlja sumu kvadrata euklidskih ℓ_2 udaljenosti točaka $a^i \in \mathcal{A}$ do neke točke $T = (x, y)$ u ravnini $\mathbb{R}^2$. Prema (2.4) vrijedi

$$F_{LS}(x, y) = \sum_{i=1}^{m} [(x - x_i)^2 + (y - y_i)^2] \geq \sum_{i=1}^{m} (\bar{x} - x_i)^2 + \sum_{i=1}^{m} (\bar{y} - y_i)^2, \tag{2.11}$$

gdje je

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i, \qquad \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i,$$

pri čemu jednakost u (2.11) vrijedi onda i samo onda ako je $x = \bar{x}$ i $y = \bar{y}$. Zato je rješenje globalno optimizacijskog problema (2.10) centroid skupa točaka $\mathcal{A}$ koji se može eksplicitno zapisati

$$c^\star_{LS} = (\bar{x}, \bar{y}). \tag{2.12}$$

[6] U cijelom tekstu elementi a^i skupa $\mathcal{A}$ označavaju se gornjim indeksom jer je donji indeks rezerviran za komponente elementa a^i.

Dakle, centroid skupa točaka $\mathcal{A}$ u ravnini je točka čija je apscisa aritmetička sredina svih apscisa točaka iz $\mathcal{A}$, a ordinata aritmetička sredina svih ordinata točaka iz $\mathcal{A}$.

Primjer 2.5. *Zadane su tri točke:* $a^1 = (0,0)$, $a^2 = (1, 3.5)$, $a^3 = (14, 2.5)$. *Centroid skupa* $\{a^1, a^2, a^3\}$ *je točka* $c_{LS}^\star = (5, 2)$ *(vidi Sliku 2.3).*

2.2.3 Medijan skupa točaka u ravnini

Medijan $c_1^\star$ skupa točaka $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, m\} \subset \mathbb{R}^2$ rješenje je optimizacijskog problema

$$\operatorname*{argmin}_{T \in \mathbb{R}^2} \sum_{i=1}^{m} d_1(T, a^i). \tag{2.13}$$

Točka $c_1^\star$ je točka u kojoj se postiže globalni minimum funkcije

$$F_1(x, y) = \sum_{i=1}^{m} (|x - x_i| + |y - y_i|).$$

Funkcija F_1 predstavlja sumu ℓ_1 udaljenosti točaka $a^i \in \mathcal{A}$ do neke točke $T = (x, y)$ u ravnini $\mathbb{R}^2$. Prema (2.6) vrijedi

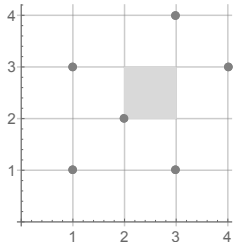
$$F_1(x, y) = \sum_{i=1}^{m} (|x - x_i| + |y - y_i|) \geq \sum_{i=1}^{m} |\operatorname*{med}_k x_k - x_i| + \sum_{i=1}^{m} |\operatorname*{med}_k y_k - y_i|, \tag{2.14}$$

pri čemu jednakost u (2.14) vrijedi onda i samo onda ako je $x = \operatorname*{med}_k x_k$ i $y = \operatorname*{med}_k y_k$. Zato je rješenje globalno optimizacijskog problema (2.13) medijan skupa točaka $\mathcal{A}$ koji se može eksplicitno zapisati

$$c_1^\star \in (\operatorname*{med}_k x_k, \operatorname*{med}_k y_k). \tag{2.15}$$

Dakle, medijan skupa točaka $\mathcal{A}$ u ravnini je točka čija je apscisa medijan svih apscisa točaka iz $\mathcal{A}$, a ordinata medijan svih ordinata točaka iz $\mathcal{A}$. Primijetite da medijan skupa $\mathcal{A}$ može biti jedna točka, segment ili čitavi pravokutnik.

Primjer 2.6. *Zadane su tri točke:* $A_1 = (0,0)$, $A_2 = (1, 3.5)$, $A_3 = (14, 2.5)$. *Medijan skupa* $\{A_1, A_2, A_3\}$ *je točka* $c_1^\star = (1, 2.5)$ *(vidi Sliku 2.3).*



Slika 2.5: Medijan skupa $\mathcal{A} = \{(1,1),(1,3),(2,2),(3,1),(3,4),(4,3)\}$

Primjer 2.7. *Medijan skupa* $\mathcal{A} = \{(1,1),(1,3),(2,2),(3,1),(3,4),(4,3)\}$ *u ravnini je kvadrat* $[2,3] \times [2,3]$ *(vidi Sliku 2.5) jer je medijan apscisa* $\text{med}\{1,1,2,3,3,4\} = [2,3]$ *i medijan ordinata* $\text{med}\{1,3,2,1,4,3\} = [2,3]$.

Zadatak 2.5. Promijenite poziciju položaja samo jedne točke tako da medijan skupa točaka $\mathcal{A}$ iz prethodnog primjera bude segment ili pravokutnik.

2.2.4 Geometrijski medijan skupa točaka u ravnini

Geometrijski medijan $c^\star$ skupa točaka $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, m\} \subset \mathbb{R}^2$ rješenje je optimizacijskog problema

$$c^\star = \underset{T \in \mathbb{R}^2}{\text{argmin}} \sum_{i=1}^{m} d_2(T, a^i). \tag{2.16}$$

Točka $c^\star$ je točka u kojoj se postiže globalni minimum funkcije

$$F_2(x,y) = \sum_{i=1}^{m} \sqrt{(x-x_i)^2 + (y-y_i)^2}.$$

Funkcija F_2 predstavlja sumu ℓ_2 udaljenosti točaka $a^i \in \mathcal{A}$ do neke točke $T = (x,y)$ u ravnini $\mathbb{R}^2$ i ne može se separirati po varijablama x i y kao u prethodnim slučajevima. Zato se rješenje globalno optimizacijskog problema (2.16) ne može eksplicitno zapisati.

Primjer 2.8. *Zadane su tri točke* $a^1 = (0,0)$, $a^2 = (1,3.5)$, $a^3 = (14,2.5)$. *U cilju pronalaženja geometrijskog medijana treba riješiti sljedeći optimizacijski problem*

$$\underset{(x,y)\in\mathbb{R}^2}{\text{argmin}}\, F_2(x,y),$$

$$F_2(x,y) = \sqrt{x^2+y^2} + \sqrt{(x-1)^2+(y-3.5)^2} + \sqrt{(x-14)^2+(y-2.5)^2}.$$

Ovaj optimizacijski problem možemo riješiti primjenom programskog sustava *Mathematica*. Najprije definiramo funkciju

```
In[1]:= F2[x_, y_]:= Sqrt[x^2 + y^2] + Sqrt[(x-1)^2 + (y-3.5)^2]
                     + Sqrt[(x-14)^2 + (y-2.5)^2]
```

Problem možemo pokušati riješiti kao problem globalne optimizacije pozivanjem `Mathematica`-modula

```
In[2]:= NMinimize[F2[x, y], {x, y}]
```

Ako to ne uspije, problem možemo pokušati riješiti kao problem lokalne optimizacije, pozivanjem `Mathematica`-modula

```
In[2]:= FindMinimum[F2[x, y], {x, 1}, {y, 2}]
```

ali u tom slučaju moramo imati dobru početnu aproksimaciju blisku rješenju. U ovom primjeru dobivamo $c_2^\star = (1.51827, 2.5876)$ (vidi Sliku 2.3).

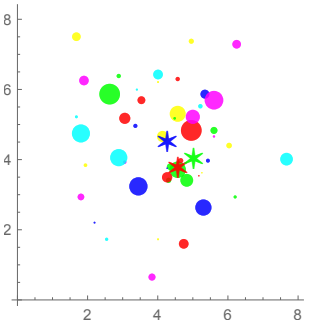
Primjedba 2.2. Optimizacijski problem (2.16) ne može se eksplicitno riješiti, već se mora primijeniti neka numerička metoda za globalnu optimizaciju [25, 27] ili uz poznavanje dobre početne aproksimacije neka metoda za lokalnu optimizaciju [61]. Najpoznatiji algoritam za traženje geometrijskog medijana je tzv. Weiszfeldov algoritam (vidi primjerice [26, 59]). To je iterativni postupak koji je nastao kao specijalni slučaj metode jednostavnih iteracija za rješavanje sustava nelinearnih jednadžbi.

Analogno Definiciji 2.3 i u slučaju podataka iz $\mathbb{R}^2$ mogu se definirati težinska aritmetička sredina, težinski medijan i težinski geometriski medijan skupa podataka $\mathcal{A}$ (vidi [50, 54]).

Primjer 2.9. *Skup podataka $\mathcal{A} \subset \mathbb{R}^2$ u ravnini definiran je* Mathematica*-programom. Svakom podatku prikazanom na Slici 2.6 pridružena je težina proporcionalna veličini kružića.*

```
In[1]:= SeedRandom[13]
        sig1 = 1.5; m1 = 50; cen = {4, 5};
        podT = Table[cen + RandomReal[NormalDistribution[0, sig1], {2}],
                     {i, m1}];
        podW = RandomReal[{0, 1}, m1];
```

Na slici je zelenom zvjezdicom označen centroid, crvenom medijan, a plavom geometrijski medijan podataka.



Slika 2.6: Težinski reprezentanti skupa $\mathcal{A}$

Zadatak 2.6. Zadan je skup $\mathcal{A} = \{(x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 10\}$, gdje je

i	1	2	3	4	5	6	7	8	9	10
x_i	9	6	8	1	1	4	4	3	9	10
y_i	5	5	5	2	5	8	1	8	8	4

Nacrtajte skup $\mathcal{A}$ u koordinatnoj ravnini i odredite centroid, medijan i geometrijski medijan ovog skupa.

Uputa: Poslužite se niže navedenim *Mathematica*-programom.

```
In[1]:= SeedRandom[2]
        A = RandomInteger[{1, 10}, {10, 2}]
        ListPlot[A, ImageSize -> 200]
        Print["Centroid = ", Mean[A]]
        Print["Medijan = ", Median[A]]
        Psi[x_, y_] := Sum[Norm[{x, y} - A[[i]]], {i, Length[A]}]
        Print["Geometrijski medijan:"]
        NMinimize[Psi[x, y], {x, y}]
```

Rješenje: $c_{LS}^\star = (5.5, 5.1)$, $c_1^\star = (5, 5)$, $c_2^\star = (6, 5)$.

2.2.5 Skup točaka na kružnici

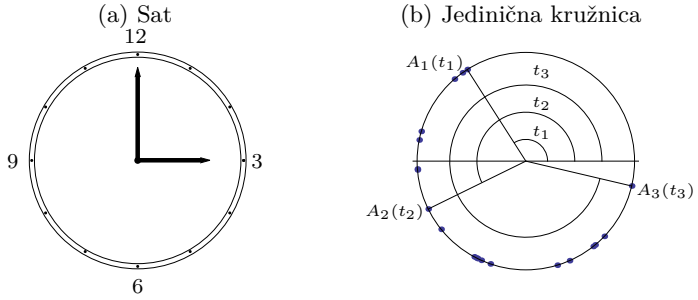
Problem određivanja najboljeg reprezentanta skupa podataka u slučaju pojava koje pokazuju periodičnost u ponašanju također je često prisutan u literaturi [43]. Temperatura zraka na nekom mjernom mjestu tijekom godine, vodostaj rijeke na nekom mjernom mjestu, seizmičke aktivnosti tijekom više godina na nekom području, količina svjetla tijekom dana, itd. primjeri su takvih pojava. Matematički gledano, treba promatrati skup točaka $\mathcal{A}$ na

kružnici. Naime, ako bismo takav skup podataka prikazali kao i ranije na brojevnom pravcu, onda bi primjerice podaci s početka i kraja iste godine bili međusobno daleko, a zapravo pripadaju istom godišnjem dobu. I za takav skup treba definirati kvazimetričku funkciju i odrediti centar podataka.

Primjer 2.10. *Neka točke $t_i \in \mathcal{A}$ predstavljaju pozicije položaja satnih kazaljki na satu s* 12 *oznaka (vidi Sliku 2.7a). Razdaljinsku (metričku) funkciju na ovom skupu definirat ćemo kao* proteklo vrijeme na satu s 12 oznaka:

$$d(t_1, t_2) = \begin{cases} t_2 - t_1, & ako\ t_1 \leq t_2 \\ 12 + (t_2 - t_1), & ako\ t_1 > t_2 \end{cases}.$$

Udaljenost $d(t_1, t_2)$ predstavlja proteklo vrijeme na satu od trenutka "t_1" do trenutka "t_2".
Primjerice: $d(2, 7) = 5$, ali $d(7, 2) = 12 + (-5) = 7$
Primijetite da ova funkcija nema svojstvo simetričnosti.



Slika 2.7: Skup podataka na kružnici

Primjer 2.11. *Neka točke $t_i \in \mathcal{A}$ predstavljaju pozicije položaja satnih kazaljki na satu s* 12 *oznaka (vidi Sliku 2.7a). Razdaljinsku (metričku) funkciju na ovom skupu definirat ćemo kao* duljinu vremenskog intervala na satu s 12 oznaka:

$$d(t_1, t_2) = \begin{cases} |t_2 - t_1|, & ako\ |t_2 - t_1| \leq 6 \\ 12 - |t_2 - t_1|, & ako\ |t_2 - t_1| > 6 \end{cases}.$$

Udaljenost $d(t_1, t_2)$ predstavlja duljinu vremenskog intervala na satu s 12 oznaka od trenutka "t_1" do trenutka "t_2".
Primjerice: $d(2, 9) = 12 - 7 = 5$ i $d(2, 7) = 7 - 2 = 5$
Primijetite da ova funkcija ima svojstvo simetričnosti.

Primjer 2.12. *Općenito, neka je $\mathcal{A}$ skup točaka na jediničnoj kružnici*

$$\mathcal{A} = \{a^i(t_i) = (\cos t_i, \sin t_i) \in \mathbb{R}^2 \colon t_i \in [0, 2\pi],\ i = 1, \ldots, m\}.$$

Udaljenost $d_K(a^1, a^2)$ dvije točke $a^1(t_1) = (\cos t_1, \sin t_1)$, $a^2(t_2) = (\cos t_2, \sin t_2)$ na jediničnoj kružnici definira se kao duljina luka između točaka a^1 i a^2

$$d_K(a^1(t_1), a^2(t_2)) = \begin{cases} |t_2 - t_1|, & ako\ |t_2 - t_1| \leq \pi \\ 2\pi - |t_2 - t_1|, & ako\ |t_2 - t_1| > \pi \end{cases}.$$

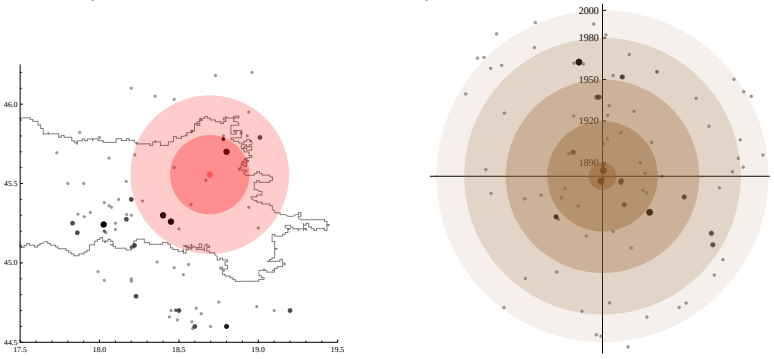
Primjerice: $d_K(a^1(0), a^2(\frac{\pi}{4})) = \frac{\pi}{4}$, $d_K(a^2(\frac{\pi}{4}), a^3(\frac{3\pi}{2})) = 2\pi - \left(\frac{6\pi}{4} - \frac{\pi}{4}\right) = \frac{3\pi}{4}$. d_K je metrička funkcija. Centar $c^\star(\tau^\star)$ skupa točaka $\mathcal{A}$ na kružnici definira se sukladno Definiciji 2.2, str. 4, kao

$$\tau^\star \in \operatorname*{argmin}_{\tau \in [0,2\pi]} \sum_{i=1}^{m} d_K(a(\tau), a^i(t_i)), \qquad c^\star(\tau^\star) = (\cos \tau^\star, \sin \tau^\star).$$

Burnov dijagram

Za grafičko prikazivanje periodičnih pojava korisno je poznavati tzv. Burnov dijagram (vidi primjerice [63]). Neka točka T na Burnovom dijagramu prikazana je kao $T = r(\cos t, \sin t)$. Parametar t je mjera kuta u radijanima kojeg zatvara radij vektor točke T s pozitivnim smjerom apscise, a parametar r predstavlja udaljenost točke T do ishodišta koordinatnog sustava.

a) Lokacije potresa b) Vremenska distribucija potresa



Slika 2.8: Potresi u okolici Osijeka od 1880. godine

Primjer 2.13. *Na Slici 2.8 a prikazane su lokacije potresa u okolici Osijeka od 1880. godine. Veće tamnije točkice predstavljaju potrese veće magnitude. Odgovarajuće točke na Burnovom dijagramu (Slika 2.8b) prikazuju godišnje trenutke potresa i njihove magnitude kao mjeru udaljenosti točke do ishodišta. Na Slici 2.8 vidi se da je posljednji jači potres u neposrednoj okolici Osijeka bio krajem zime 1922. godine na geografskoj poziciji* $(18.8, 45.7)$ *(u blizini mjesta Lug, dvadesetak kilometara sjeveroistočno od Osijeka) i da je imao magnitudu* 5.1.

2.3 Reprezentant podataka iz $\mathbb{R}^n$

U praktičnim primjenama podaci mogu imati veći broj obilježja kao što je spomenuto na početku Uvoda, str. 1, gdje je i navedeno nekoliko takvih primjera. Budući da broj obilježja podataka predstavlja njihovu dimenziju $n \geq 1$, bit će potrebno znati odrediti reprezentant i za podatke visoke dimenzije.

Pretpostavimo da je zadan skup točaka $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \in \mathbb{R}^n : i = 1, \ldots, m\}$. Treba odrediti točku koja će što bolje reprezentirati taj skup.

2.3.1 Kvazimetričke funkcije u $\mathbb{R}^n$

Definicija 2.4. Funkciju $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, koja ima svojstvo *pozitivne definitnosti*

$$\begin{aligned} &\text{(i)} \quad d(x,y) \geq 0 \quad \forall x, y \in \mathbb{R}^n, \\ &\text{(ii)} \quad d(x,y) = 0 \quad \Leftrightarrow \quad x = y, \end{aligned}$$

zovemo *kvazimetrička funkcija* na $\mathbb{R}^n$.

Primjer 2.14. *Najčešći primjeri kvazimetričkih funkcija su (vidi [31, 68]):*

$d_{LS}(x,y) = \|x-y\|_2^2 = (x-y)^T(x-y)$, *[LS-kvazimetrička funkcija]*
$d_2(x,y) = \|x-y\|_2 = \sqrt{(x-y)^T(x-y)}$, *[$\ell_2$-euklidska metrička funkcija]*
$d_1(x,y) = \|x-y\|_1$, *[ℓ_1-metrička funkcija (Manhattan metrika)]*
$d_\infty(x,y) = \|x-y\|_\infty$, *[$\ell_\infty$-Čebiševljeva metrička funkcija]*
$d_p(x,y) = \|x-y\|_p$, $p \geq 1$, *[ℓ_p-Minkowsky metrička funkcija]*
$d_M(x,y) = (x-y)^T S^{-1}(x-y)$, *[Mahalanobis kvazimetrička funkcija]*
($S \in \mathbb{R}^{n\times n}$ je simetrična pozitivno definitna matrica)

Definicija 2.5. Neka je $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ kvazimetrička funkcija. Najbolji reprezentant (centar $c^\star \in \mathbb{R}^n$) skupa $\mathcal{A} \subset \mathbb{R}^n$ u odnosu na kvazimetričku funkciju d je

$$c^\star \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{i=1}^m d(x,a^i). \tag{2.17}$$

Točka $c^\star \in \mathbb{R}^n$ je točka globalnog minimuma funkcije $F\colon \mathbb{R}^n \to \mathbb{R}_+$

$$F(x) = \sum_{i=1}^m d(x,a^i). \tag{2.18}$$

Specijalno,

(a) u slučaju LS-kvazimetričke funkcije, najbolji reprezentant skupa $\mathcal{A}$ je centroid (težište) skupa

$$c_{LS}^\star = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{i=1}^m d_{LS}(x,a^i) = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{i=1}^m \|x-a^i\|_2^2 = \frac{1}{m}\sum_{i=1}^m a^i,$$

a odgovarajuća minimizirajuća funkcija glasi

$$F_{LS}(x) = \sum_{i=1}^m \|x-a^i\|_2^2;$$

(b) u slučaju ℓ_1-metričke funkcije, najbolji reprezentant skupa $\mathcal{A}$ je medijan skupa

$$\begin{aligned} c_1^\star &\in \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{i=1}^m d_1(x,a^i) \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{i=1}^m \|x-a^i\|_1 = \operatorname*{med}_i a^i = (\operatorname*{med}_i a_1^i, \ldots, \operatorname*{med}_i a_n^i), \end{aligned}$$

a odgovarajuća minimizirajuća funkcija glasi

$$F_1(x) = \sum_{i=1}^{m} \|x - a^i\|_1.$$

Definicija 2.6. Neka je $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ kvazimetrička funkcija. Najbolji reprezentant skupa $\mathcal{A} \subset \mathbb{R}^n$ s težinama $w_1, \ldots, w_m > 0$ u odnosu na kvazimetričku funkciju d je

$$c^\star \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i d(x, a^i). \tag{2.19}$$

Točka $c^\star \in \mathbb{R}^n$ je točka globalnog minimuma funkcije $F\colon \mathbb{R}^n \to \mathbb{R}_+$

$$F(x) = \sum_{i=1}^{m} w_i d(x, a^i). \tag{2.20}$$

(a) Specijalno, ako je d LS-kvazimetrička funkcija, najbolji reprezentant skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ je težinski centroid (težište) skupa

$$c^\star_{LS} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i d_{LS}(x, a^i) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \|x - a^i\|_2^2 = \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i, \text{ tj.}$$

$$c^\star_{LS} = \Big(\tfrac{1}{W} \sum_{i=1}^{m} w_i a^i_1, \ldots, \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i_n \Big) \quad \text{[po koordinatama]},$$

gdje je $W = \sum_{i=1}^{m} w_i$, a odgovarajuća minimizirajuća funkcija glasi

$$F_{LS}(x) = \sum_{i=1}^{m} w_i \|x - a^i\|_2^2;$$

(b) Ako je d, ℓ_1-metrička funkcija, najbolji reprezentant skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ je težinski medijan skupa

$$c^\star_1 \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i d_1(x, a^i) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i \|x - a^i\|_1,$$

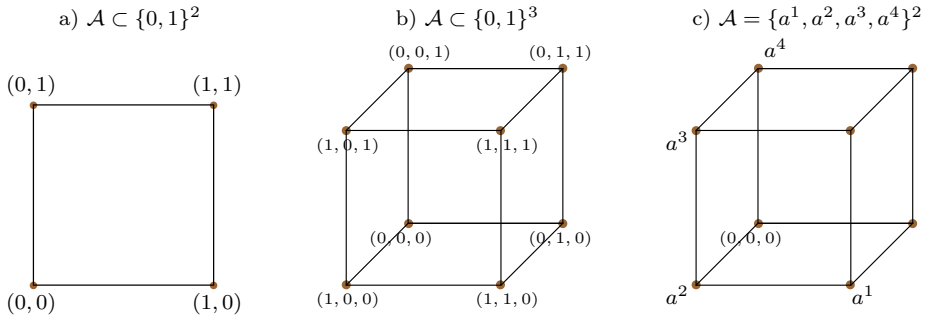
$$c^\star_1 \in \operatorname*{med}_i (w_i, a^i) = \big(\operatorname*{med}_i (w_i, a^i_1), \ldots, \operatorname*{med}_i (w_i, a^i_n) \big),$$

a odgovarajuća minimizirajuća funkcija glasi

$$F_1(x) = \sum_{i=1}^{m} w_i \|x - a^i\|_1.$$

2.3.2 Jedna primjena: Prepoznavanje riječi u tekstu

U nekom tekstu prisutnost neke riječi kodira se s 1, a odsutnost te riječi s 0. Postavlja se pitanje o sličnosti/različitosti dva teksta s obzirom na prisutnost/odsutnost promatranih riječi. Tekst u kome je prisutno/odsutno $n \geq 1$ izabranih riječi prikazat ćemo vektorom iz $\mathbb{R}^n$ s komponentama 0 ili 1.



Slika 2.9: Skup $\mathcal{A}$ za $n = 2$ i $n = 3$

Promatramo dakle skup vektora, čije su komponente brojevi 0 ili 1, tj. $\mathcal{A} = \{a^i = (x_1, \ldots, x_n) \in \{0,1\}^n \colon i = 1, \ldots, m\} \subset \mathbb{R}^n$ (vidi Sliku 2.9a-b). Na tom skupu također možemo definirati kvazimetričku funkciju $d \colon \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$, kao primjerice

$$
\begin{aligned}
&d_{LS}(x,y) = \|x-y\|_2^2, && \text{[LS-kvazimetrička funkcija]}\\
&d_1(x,y) = \|x-y\|_1, && \text{[}\ell_1\text{-metrička funkcija]}\\
&d_\infty(x,y) = \|x-y\|_\infty. && \text{[}\ell_\infty\text{-metrička funkcija]}
\end{aligned}
$$

Za ovakve probleme posebno je pogodna tzv. kosinus kvazimetrička funkcija [4–6, 10]

$$d_c(x,y) = 1 - \frac{\langle x, y\rangle}{\|x\|_2 \cdot \|y\|_2}, \tag{2.21}$$

gdje je $\langle\cdot\rangle$ uobičajeni skalarni produkt. Primijetite da međusobno bliski (skoro linearno zavisni) vektori imaju d_c-udaljenost skoro 0, a međusobno različiti (skoro „okomiti") vektori imaju d_c-udaljenost skoro 1.

Primjer 2.15. *Promatramo tekstove u kojima se mogu pojaviti riječi: A,B,C. Neka je primjerice (vidi Sliku 2.9c):*

$a^1 = (1,1,0)$: *tekst u kojemu se pojavljuju riječi* A, B*, a ne pojavljuje se riječ* C

$a^2 = (1,0,0)$: *tekst u kojemu se pojavljuje riječ* A*, a ne pojavljuju se riječi* B, C

$a^3 = (1,0,1)$: *tekst u kojemu se pojavljuju riječi* A, C*, a ne pojavljuje se riječ* B

$a^4 = (0,0,1)$: *tekst u kojemu se pojavljuje riječ* C*, a ne pojavljuju se riječi* A, B

U svrhu ispitivanja sličnosti/različitosti tekstova obzirom na prisutnost/ odsutnost nekih riječi možemo pokušati iskoristiti ranije spomenute kvazimetričke funkcije. Za prethodno spomenuti primjer dobivamo

	$d(a^1,a^2)$	$d(a^1,a^3)$	$d(a^1,a^4)$	$d(a^2,a^3)$	$d(a^2,a^4)$	$d(a^3,a^4)$
$d := d_{LS}$	1	2	3	1	2	1
$d := d_1$	1	2	3	1	2	1
$d := d_c$	.29	.5	1	.29	1	.29

Prema svim korištenim funkcijama tekstovi (a^1, a^2), (a^2, a^3) i (a^3, a^4) su najsličniji (najbliži), a tekstovi (a^1, a^4) najrazličitiji (najudaljeniji) obzirom na pojavu riječi A, B, C. Razmotrite ovaj rezultat geometrijski na Slici 2.9.

Primjer 2.16. *Promatramo tekstove u kojima se mogu pojaviti riječi:* A, B, C, D, E*. Neka je primjerice:*

$a^1 = (1,0,0,0,1)$: tekst u kojemu se pojavljuju riječi A, E, a ne pojavljuju se riječi B, C, D

$a^2 = (0,1,1,0,0)$: tekst u kojemu se pojavljuju riječi B, C, a ne pojavljuju se riječi A, D, E

$a^3 = (1,0,0,0,0)$: tekst u kojemu se pojavljuje riječ A, a ne pojavljuju se riječi B, C, D, E

Dobivamo

	$d(a^1,a^2)$	$d(a^1,a^3)$	$d(a^2,a^3)$
$d := d_{LS}$	4	1	3
$d := d_1$	4	1	3
$d := d_c$	1	.29	1

Koji su tekstovi najsličniji, a koji najrazličitiji u odnosu na pojedinu kvazimetričku funkciju?

Zadatak 2.7. Konstruirajte skup tekstova $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ tako da neki izabrani tekst $b \in \mathbb{R}^n$ bude najsličniji tekstu a^1 prema udaljenosti d_c, ali da to nije tako prema udaljenostima d_{LS} i d_1.

Zadatak 2.8. Konstruirajte skup tekstova $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$, $m \gg 2$ i neki izabrani tekst $b \in \mathbb{R}^n$. Koristeći d_c-udaljenost pronađite dva teksta iz $\mathcal{A}$ koji su najbliži tekstu b.

Poglavlje 3

Grupiranje podataka

3.1 Problem grupiranja podataka

Definicija 3.1. Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup s $m \geq 2$ elemenata. Rastav skupa $\mathcal{A}$ na $1 \leq k \leq m$ disjunktnih nepraznih podskupova $\pi_1, \ldots, \pi_k$, takvih da bude

$$
\begin{aligned}
&(i) \quad \bigcup_{j=1}^{k} \pi_j = \mathcal{A}, \\
&(ii) \quad \pi_r \bigcap \pi_s = \emptyset, \quad r \neq s, \\
&(iii) \quad m_j := |\pi_j| \geq 1, \quad j = 1, \ldots, k.
\end{aligned}
$$

zovemo *particija* Π *skupa* $\mathcal{A}$. Elemente particije $\Pi = \{\pi_1, \ldots, \pi_k\}$ zovemo *klasteri*. Skup svih particija skupa $\mathcal{A}$ sastavljenih od k klastera koje zadovoljavaju *(i)-(iii)* označavamo s $\mathcal{P}(\mathcal{A}; k)$.

Nadalje, kad god budemo govorili o particiji skupa $\mathcal{A}$, podrazumijevat ćemo da je ona sastavljena od ovakvih podskupova skupa $\mathcal{A}$. Na taj način svjesno smo iz razmatranja isključili particije koje sadržavaju prazan skup ili skup $\mathcal{A}$.

Može se pokazati [77] da je broj svih particija skupa $\mathcal{A}$ iz Definicije 3.1 jednak Stirlingovom broju druge vrste

$$
|\mathcal{P}(\mathcal{A}; k)| = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^m. \tag{3.1}
$$

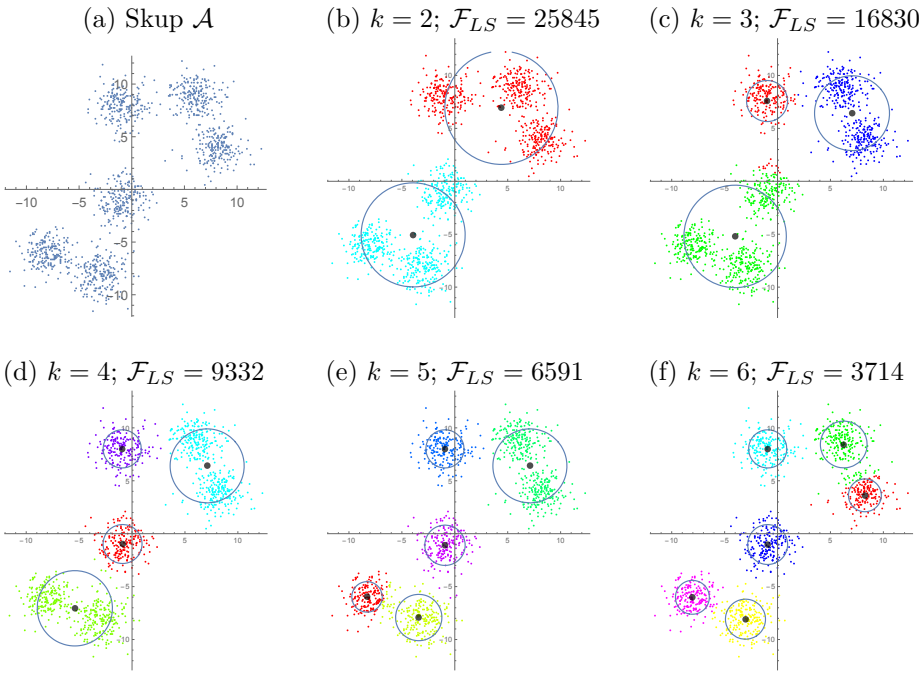
Specijalno,

$$\text{za } k=2: \quad |\mathcal{P}(\mathcal{A};2)| = \tfrac{1}{2}(2^m-2) = 2^{m-1}-1,$$
$$\text{za } k=3: \quad |\mathcal{P}(\mathcal{A};3)| = \tfrac{1}{2}\left(1-2^m+3^{m-1}\right).$$

Primjer 3.1. *Broj svih particija skupa $\mathcal{A}$ koje zadovoljavaju Definiciju 3.1 može biti ogroman. Za $m = 5, 10, 50, 1200, 10^6$ i $k = 2, 3, 4, 5, 6, 8, 10$ broj svih particija skupa $\mathcal{A}$ vidljiv je u Tablici 3.1*

$\lvert\mathcal{P}(\mathcal{A};k)\rvert$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=8$	$k=10$
$m=5$	15	25	10	1	–	–	–
$m=10$	511	9330	34105	42525	22827	750	1
$m=50$	10^{15}	10^{23}	10^{29}	10^{33}	10^{36}	10^{41}	10^{44}
$m=1200$	10^{361}	10^{572}	10^{721}	10^{837}	10^{931}	10^{1079}	10^{1193}
$m=10^6$	10^{301030}	10^{477120}	10^{602058}	10^{698968}	10^{778148}	10^{903085}	10^{10^6}

Tablica 3.1: Približni broj particija u ovisnosti o broju elemenata i broju klastera



Slika 3.1: Neke particije skupa $\mathcal{A}$

Primjer 3.2. *Zadan je skup $\mathcal{A} \subset \mathbb{R}^2$ prikazan na Slici 3.1a koji sadržava $m = 1200$ elemenata. U Tablici 3.1 može se vidjeti približan broj svih njegovih particija od $k = 2, 3, 4, 5, 6, 8$ i 10 klastera.*

Ako uvedemo kriterij da je bolja particija ona čiji su klasteri kompaktniji i bolje razdvojeni, onda bismo mogli postaviti pitanje optimalne (najbolje) particije.

Mjeru kompaktnosti i dobre razdvojenosti klastera u nekoj particiji Π s k klastera $\pi_1, \ldots, \pi_k$ mogli bismo definirati na sljedeći način:

1. Udaljenost elemenata skupa $\mathcal{A}$ mjerit ćemo LS-kvazimetričkom funkcijom $d_{LS}(a, b) = \|a - b\|_2^2$;

2. U svakom klasteru π_j odredimo centroid $c_j = \frac{1}{|\pi_j|} \sum_{a^i \in \pi_j} a^i$;

3. Za svaki klaster π_j odredimo ukupno „rasipanje” (suma kvadrata udaljenosti točaka klastera π_j do centra c_j) $F_{LS}(\pi_j) = \sum_{a^i \in \pi_j} \|c_j - a^i\|_2^2$;

4. Mjera kompaktnosti i dobre razdvojenosti klastera u particiji iskazuje se funkcijom cilja $F_{LS}(\pi_1, \ldots, \pi_k) = \sum_{j=1}^{k} F_{LS}(\pi_j) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} \|c_j - a^i\|_2^2$.

Na Slici 3.1 možemo vidjeti po jednu particija s $k = 2, 3, 4, 5$ i 6 klastera i odgovarajuće vrijednosti funkcije cilja. Primijetimo da se povećanjem broja klastera, smanjuje vrijednost kriterijske funkcije cilja. Primjerice, na Slici 3.1c prikazana je jedna od mnogobrojnih (vidi Tablicu 3.1) 3-particija skupa $\mathcal{A}$. Na njoj funkcija cilja $\mathcal{F}_{LS}$ postiže vrijednost 16830. Opravdano je postaviti pitanje je li to i najbolja 3-particija, tj. može li se pronaći neka druga 3-particija s nižom vrijednosti funkcije cilja?

Općenito bismo mogli postaviti barem nekoliko sljedećih pitanja:

1. Je li navedena funkcija cilja najprikladnija za ovaj primjer?
2. Koliki je najprikladniji broj klastera?
3. Imaju li particije prikazane na Slici 3.1 najniže vrijednosti funkcije cilja od svih mogućih particija s tim brojem klastera?

Iz navedenog primjera vidi se da odgovori na postavljena pitanja neće biti jednostavni. Pitanje izbora funkcije cilja kao i pitanje najprikladnijeg broja klastera u particiji zadire u prethodnu statističku analizu podataka. U ovom ćemo udžbeniku za definiranje funkcije cilja uglavnom koristiti LS-kvazimetričku funkciju i ℓ_1-metričku funkciju. O izboru najprikladnijeg broja klastera u particiji bavit ćemo se u Poglavlju 6.

Traženje optimalne particije općenito neće biti moguće provesti pretraživanjem čitavog skupa $\mathcal{P}(\mathcal{A}; k)$. Odmah treba reći da problem traženja optimalne particije spada u NP-teške probleme [40] nekonveksne optimizacije općenito nediferencijabilne funkcije više varijabli, koja najčešće posjeduje značajan broj stacionarnih točaka. U ovom udžbeniku bavit ćemo se traženjem optimalne particije u Poglavlju 4 i Poglavlju 5.

Zadatak 3.1. Zadan je skup $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 13\}$, gdje je

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	2	3	3	4	4	5	5	6	7	7	7	8	9
y_i	2	4	7	2	8	6	7	9	3	4	7	1	3

Pronađite globalno optimalnu 2-particiju $\Pi^\star = \{\pi_1^\star, \pi_2^\star\}$ primjenom ℓ_1-metričke funkcije. Ucrtajte točke u koordinatni sustav i naznačite rješenje.

3.1.1 Neke primjene

Primjenu klaster analize moguće je pronaći u svim područjima znanstvenih i primijenjenih istraživanja. Navedimo nekoliko ilustrativnih primjera.

Primjer 3.3. *Na Slici 3.2a crnim točkicama prikazane su geografske lokacije u širem području Republike Hrvatske [43, 63] na kojima se od 1900. godine dogodio potres magnitude veće ili jednake od* 3*. Težine (ponderi) ovih dvodimenzionalnih podataka su magnitude potresa. Geografske pozicije potresa grupirane su u zone seizmičkih aktivnosti, koje su također prikazane na Slici 3.2a.*

(a) Zone potresa (b) Izborne jedinice (c) Klasteri projekata

Slika 3.2: Neke primjene grupiranja podataka

Primjer 3.4. *Problem optimalnog definiranja izbornih jedinica u Republici Hrvatskoj [38, 52, 56] također je jedan problem optimalnog grupiranja podataka. Izborne jednice trebale bi se sastojati od približno jednakog broja birača (do na* 5%*), koje povezuje zajednički interes kroz ekonomsku, prometnu, povijesnu i drugu povezanost. Problem se može promatrati kao problem određivanja broja i geografske povezanosti izbornih jedinica s jednakim brojem zastupnika, ali se može promatrati i tako da izborne jedinice ne budu jednake veličine i da imaju različiti broj zastupnika (vidi Sliku 3.2b).*

Primjer 3.5. *U radu [75] razmatra se problem grupiranja i rangiranja istraživačkih projekata prijavljenih na neki raspisani natječaj. Projekti se grupiraju u klastere na bazi ocjena dobivenih u recenzentskom postupku primjenom adaptivne Mahalanobis kvazimetričke funkcije. Klaster najbolje ocijenjenih projekata posebno se analizira. Navedeno je nekoliko mogućnosti korištenja podataka dobivenih recenzentskim postupkom, a predložena metoda ilustrirana je na primjeru internih istraživačkih projekata na Sveučilištu u Osijeku. Na Slici 3.2c prikazana je jedna mogućnost razvrstavanja projekata.*

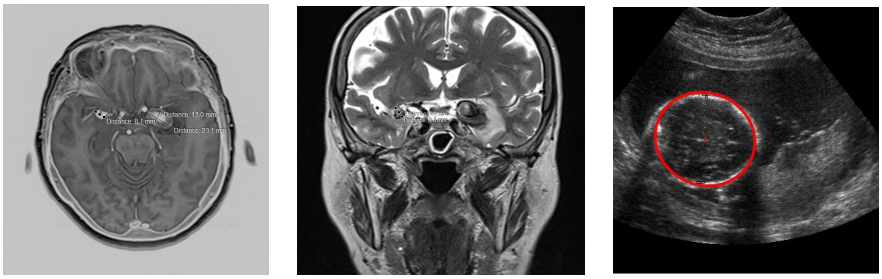
Primjer 3.6. *Na Slici 3.3b prikazana je „crno-bijela" 512×512 slika „Elaine" i njena segmentacija u 2 (Slika 3.3a) i 8 (Slika 3.3c) klastera (nijansi). U ovom slučaju podaci $\mathcal{A} = \{a^i \in \mathbb{R} : i = 1, \ldots, 262\,144\}$ imaju samo jedan atribut (gray level). Redni broj (indeks) podatka a^i definira njegovu poziciju na slici.*

(a) Reprezentacija s 2 nijanse (b) Originalna fotografija (c) Reprezentacija s 8 nijansi

Slika 3.3: Segmentacija crno-bijele slike „Elaine"

Primjer 3.7. *Na Slici 3.4a-b prikazana je crno-bijela slika ljudskog mozga*[1]*, a na Slici 3.4c slika ljudskog embrija s detektiranom glavom (vidi [22]). Primjenom klaster analize cilj je prepoznati anomalije i tendenciju njihovih povećanja ili smanjivanja.*

(a) Slika ljudske glave (b) Slika ljudske glave (c) Detekcija glave fetusa

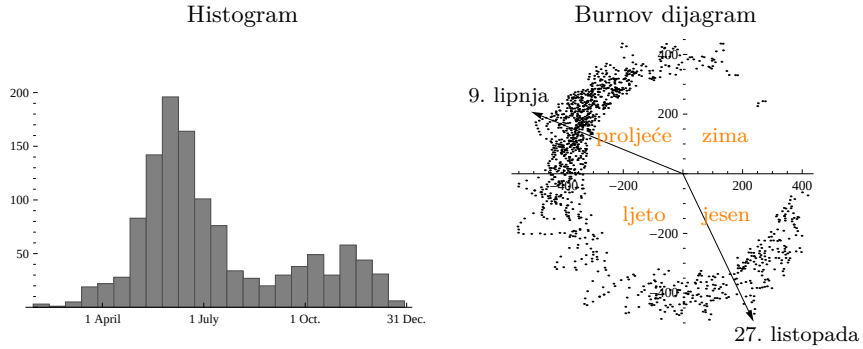


Slika 3.4: Prepoznavanje objekata na medicinskim slikama

Primjer 3.8. *Na Slici 3.5a prikazan je histogram visokog vodostaja Drave kod Donjeg Miholjca od 1900. godine, a na Slici 3.5b odgovarajući Burnov*

[1] Slike pripremio dr. Salha Tamer, KBC Osijek, Klinički zavod za dijagnosticku i intervencijsku radiologiju

dijagram. Točke na Burnovom dijagramu prikazuju godišnje trenutke visokog vodostaja i njihove vrijednosti kao mjeru udaljenosti točke do ishodišta. Točke na Burnovom dijagramu grupirane su u dva klastera primjenom kvazimetričke funkcije iz Primjera 2.12, str. 17. Primijetite da se najviši vodostaj Drave kod D. Miholjca može očekivati početkom lipnja i krajem listopada (centri klastera!).



Slika 3.5: Vodostaj Drave kod D. Miholjca od 1900. godine

3.1.2 Programska podrška

Postoje različiti izvori programske podrške za traženje i grafičko prikazivanje grupiranja podataka. Navedimo neke koje ćemo koristiti:

- Programski sustav *Mathematica*: naredba `FindClusters` s kvazimetričkim funkcijama:
  $DistanceFunction \to SquaredEuclideanDistance$ (default),
  $DistanceFunction \to ManhattanDistance$, ...
- *Mathematica*-moduli uz ovaj udžbenik:
  `http://www.mathos.unios.hr/images/homepages/scitowsk/Programi.zip`
  odnosno
  `http://www.efos.unios.hr/algoritmi-strukture-podataka/nastavni-materijali/`

3.2 Motivacija: grupiranje u dva klastera na osnovi jednog obilježja

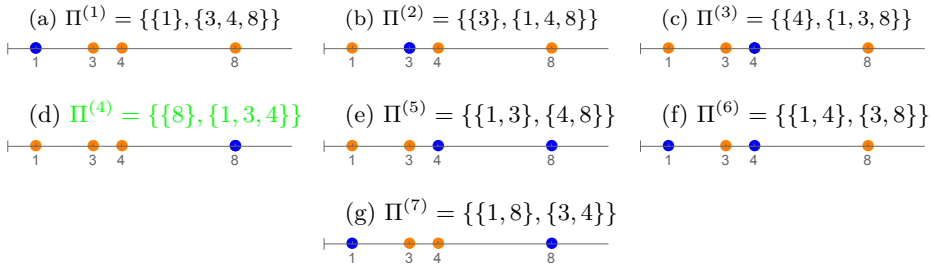
Neka je $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ podskup skupa realnih brojeva. Sukladno

Definiciji 3.1, 2-particija $\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$ skupa $\mathcal{A}$ sastoji se od dva klastera π_1, π_2, takvih da je

$$\pi_1 \cup \pi_2 = \mathcal{A}, \qquad \pi_1 \cap \pi_2 = \emptyset, \qquad m_1 = |\pi_1| \geq 1, \quad m_2 = |\pi_2| \geq 1.$$

Prema (3.1), broj svih ovakvih 2-particija skupa $\mathcal{A}$ je $|\mathcal{P}(\mathcal{A}; k)| = 2^{m-1} - 1$.

Primjer 3.9. *Neka je $\mathcal{A} = \{1, 3, 4, 8\}$. U Tablici 3.2 i na Slici 3.6 prikazano je svih 7 mogućih particija.*



Slika 3.6: Sve particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$

Postavlja se pitanje koja je od njih „najbolja" u smislu:

- *interne kompaktnosti, tj. u smislu da su svi slični/bliski elementi što više na okupu i*
- *eksterne razdvojenosti, tj. u smislu da su elementi pojedinih klastera što više razdvojeni jedni od drugih.*

Vizualno, ta svojstva najbolje ispunjava particija $\Pi^{(4)} = \{\{1, 3, 4\}, \{8\}\}$.

Pokušajmo kvantificirati navedene kriterije. Neka je $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ neka kvazimetrička funkcija. Definirajmo centre klastera

$$c_1 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_1} d(x, a), \qquad c_2 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_2} d(x, a),$$

i uvedimo sljedeću kriterijsku funkciju cilja

$$\mathcal{F}(\Pi) = \sum_{a \in \pi_1} d(c_1, a) + \sum_{a \in \pi_2} d(c_2, a). \tag{3.2}$$

Geometrijski, funkcija $\mathcal{F}$ predstavlja ukupno „rasipanje", tj. zbroj suma udaljenosti elemenata klastera do njegovih centara. Jasno je da što je vrijednost

kriterijske funkcije $\mathcal{F}$ manja, time je „rasipanje" manje, a time su bolje ispunjeni i prethodno navedeni kriteriji kompaktnosti i razdvojenosti. Problem traženja optimalne 2-particije definira se dakle kao sljedeći optimizacijski problem

$$\underset{\Pi \in \mathcal{P}(\mathcal{A};2)}{\operatorname{argmin}} \mathcal{F}(\Pi) \tag{3.3}$$

3.2.1 Princip najmanjih kvadrata

Promatrajmo LS-kvazimetričku funkciju $d_{LS}(x, y) = (x - y)^2$. U ovom slučaju centroidi klastera π_1, π_2 su

$$c_1 = \frac{1}{m_1} \sum_{a \in \pi_1} a, \qquad c_2 = \frac{1}{m_2} \sum_{a \in \pi_2} a, \tag{3.4}$$

a odgovarajuća kriterijska funkcija cilja je

$$\mathcal{F}_{LS}(\Pi) = \sum_{a \in \pi_1} (c_1 - a)^2 + \sum_{a \in \pi_2} (c_2 - a)^2. \tag{3.5}$$

Njezina vrijednost na particiji $\Pi = \{\pi_1, \pi_2\}$ je suma kvadrata udaljenosti točaka klastera π_1 do centroida c_1 i točaka klastera π_2 do centroida c_2.

Primjer 3.10. *Za skup $\mathcal{A} = \{1, 3, 4, 8\}$ iz Primjera 3.9 treba odrediti sve njegove particije, pripadne centroide i vrijednosti funkcije cilja $\mathcal{F}_{LS}$ uz primjenu LS-kvazimetričke funkcije.*

π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$		$\mathcal{G}(\Pi)$	
$\{1\}$	$\{3,4,8\}$	1	5	0+14=	14	9+3=	12
$\{3\}$	$\{1,4,8\}$	3	13/3	0+74/3=	24.67	1+1/3=	1.33
$\{4\}$	$\{1,3,8\}$	4	4	0+26=	26	0+0=	0
$\{8\}$	$\{1,3,4\}$	8	8/3	0+14/3=	4.67	16+16/3=	21.33
$\{1,3\}$	$\{4,8\}$	2	6	2+8=	10	8+8=	16
$\{1,4\}$	$\{3,8\}$	5/2	11/2	9/2+25/2=	17	9/2+9/2=	9
$\{1,8\}$	$\{3,4\}$	9/2	7/2	49/2+1/2=	25	1/2+1/2=	1

Tablica 3.2: Biranje optimalne particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$

U Tablici 3.2 vidi se da particija na kojoj funkcija cilja $\mathcal{F}_{LS}$ postiže najmanju vrijednost odgovara particiji koju smo i ranije vizualno identificirali kao najbolju. To je LS-optimalna 2-particija

$$\Pi^\star = \{\pi_1^\star, \pi_2^\star\} = \{\{8\}, \{1, 3, 4\}\},$$

na kojoj funkcija cilja prima najmanju moguću vrijednost $\mathcal{F}_{LS}(\Pi^\star) = 4.67$.

Dualni problem

Problem traženja LS-optimalne 2-particije još ćemo dodatno proanalizirati. Neka je:

$$\begin{aligned}
&\mathcal{A} = \{a_i \in \mathbb{R} \colon i = 1, \ldots, m\},\ m \geq 2 \quad \text{[skup]};\\
&c = \tfrac{1}{m} \sum_{i=1}^{m} a_i \quad \text{[centroid skupa } \mathcal{A}\text{]};\\
&\Pi = \{\pi_1, \pi_2\} \quad \text{[2-particija s klasterima } \pi_1, \pi_2\text{]};\\
&m_1 = |\pi_1|, \quad m_2 = |\pi_2| \quad \text{[broj elemenata klastera]};\\
&c_1 = \tfrac{1}{m_1} \sum_{a \in \pi_1} a \quad \text{[centroid klastera } \pi_1\text{]};\\
&c_2 = \tfrac{1}{m_2} \sum_{a \in \pi_2} a \quad \text{[centroid klastera } \pi_2\text{]}.
\end{aligned}$$

Sukladno Lemi 5.1, str. 98, vrijedi:

$$\sum_{a \in \pi_1} (c_1 - a) = 0, \tag{3.6}$$

$$\sum_{a \in \pi_2} (c_2 - a) = 0, \tag{3.7}$$

$$\sum_{a \in \pi_1} (c - a)^2 = \sum_{\pi_1} (c_1 - a)^2 + m_1 (c_1 - c)^2, \tag{3.8}$$

$$\sum_{a \in \pi_2} (c - a)^2 = \sum_{\pi_2} (c_2 - a)^2 + m_2 (c_2 - c)^2. \tag{3.9}$$

Zbrajanjem jednakosti (3.8) i (3.9) dobivamo

$$\begin{aligned}
&\sum_{a \in \pi_1} (c - a)^2 + \sum_{a \in \pi_2} (c - a)^2\\
&\quad = \sum_{a \in \pi_1} (c_1 - a)^2 + \sum_{a \in \pi_2} (c_2 - a)^2 + m_1 (c - c_1)^2 + m_2 (c - c_2)^2,
\end{aligned}$$

odnosno

$$\sum_{i=1}^{m} (c - a_i)^2 = \mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi), \tag{3.10}$$

gdje je

$$\mathcal{F}_{LS}(\Pi) = \sum_{a\in\pi_1} (c_1 - a)^2 + \sum_{a\in\pi_2} (c_2 - a)^2,$$
$$\mathcal{G}(\Pi) = m_1(c - c_1)^2 + m_2(c - c_2)^2.$$

U izrazu (3.10) prirodno se pojavila funkcija cilja $\mathcal{F}_{LS}$. Izraz (3.10) pokazuje da se ukupno rasipanje elemenata skupa $\mathcal{A}$ oko njegovog centroida c može prikazati kao zbroj dviju kriterijskih funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$. Specijalno, ako je $\Pi^\star = \{\pi_1^\star, \pi_2^\star\}$ LS-optimalna 2-particija, onda se najmanja moguća vrijednost funkcije cilja $\mathcal{F}_{LS}$ postiže na $\Pi^\star$, tj. $\mathcal{F}_{LS}(\Pi^\star) = \min\limits_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{F}_{LS}(\Pi)$.

Postavlja se pitanje: što je u tom slučaju $\mathcal{G}(\Pi^\star)$?

Kako bismo odgovorili na to pitanje, najprije dopunimo Tablicu 3.2 vrijednostima koje funkcija $\mathcal{G}$ prima na pojedinim particijama (plavi dio tablice). Primijetite da se najveća vrijednost funkcije $\mathcal{G}$ postiže baš na LS-optimalnoj 2-particiji $\Pi^\star$, na kojoj je funkcija cilja F_{LS} primila najmanju vrijednost.

Je li to slučajno?

Kako bismo odgovorili na ovo pitanje, pokušajmo najprije riješiti sljedeći zadatak, gdje se razmatra sličan problem. Za rješavanje ovog zadatka bit će nam potrebno predznanje *Matematike I* (vidi primjerice [29]).

Zadatak 3.2. Neka su $\varphi, \psi\colon \mathbb{R} \to \mathbb{R}$ dvije neprekidne funkcije za koje vrijedi $\varphi(x) + \psi(x) = const.$ Pokažite da ako postoji $x_0 \in \mathbb{R}$ takav da je

$$\varphi'(x_0) = 0 \quad \& \quad \varphi''(x_0) > 0, \quad \text{onda vrijedi} \quad \psi'(x_0) = 0 \quad \& \quad \psi''(x_0) < 0;$$

odnosno

$$\text{ako je } x_0 \in \operatorname*{argmin}_{x\in\mathbb{R}} \varphi(x), \quad \text{onda je} \quad x_0 \in \operatorname*{argmax}_{x\in\mathbb{R}} \psi(x);$$
$$\text{ako je } \min_{x\in\mathbb{R}} \varphi(x) = \varphi(x_0), \quad \text{onda je} \quad \max_{x\in\mathbb{R}} \psi(x) = \psi(x_0).$$

Provjerite imaju li funkcije $\varphi(x) = x^2 - 1$ i $\psi(x) = -x^2 + 1$ navedena svojstva. Nacrtajte njihove grafove u jednom koordinatnom sustavu. Pokušajte sami konstruirati još jedan primjer para funkcija φ, ψ koje zadovoljavaju navedena svojstva.

Sličnim razmatranjem može se pokazati (vidi primjerice [31]) da vrijedi

$$(i) \quad \operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{F}_{LS}(\Pi) = \operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{G}(\Pi) =: \Pi^\star, \tag{3.11}$$

$$(ii) \quad \min_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{F}_{LS}(\Pi) = \mathcal{F}_{LS}(\Pi^\star) \quad \& \quad \max_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{G}(\Pi) = \mathcal{G}(\Pi^\star). \tag{3.12}$$

Problem određivanja LS-optimalne particije rješavanjem optimizacijskog problema

$$\operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{G}(\Pi), \tag{3.13}$$

naziva se dualni problem u odnosu na optimizacijski problem $\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{F}_{LS}(\Pi)$.

Zadatak 3.3. Zadan je skup $\mathcal{A} = \{0, 3, 6, 9\}$. Primjenom LS-kvazimetričke funkcije odredite sve njegove dvočlane particije koje zadovoljavaju Definiciju 3.1, pripadne centroide i vrijednosti kriterijskih funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$. Odredite LS-optimalnu 2-particiju skupa $\mathcal{A}$.

Rješenje: $\Pi^\star = \{\{0, 3\}, \{6, 9\}\}$

3.2.2 Princip najmanjih apsolutnih odstupanja

Neka je $\mathcal{A} = \{a_i \in \mathbb{R} \colon i = 1, \ldots, m\}$ skup, a $\Pi = \{\pi_1, \pi_2\}$ neka njegova particija. Primjenom ℓ_1-metričke funkcije $d_1(x, y) = |x - y|$ (vidi primjerice [31, 53, 54]) određeni su centri

$$c_1 = \operatorname{med}(\pi_1) = \operatorname*{med}_{a_i\in\pi_1} a_i, \qquad c_2 = \operatorname{med}(\pi_2) = \operatorname*{med}_{a_i\in\pi_2} a_i, \tag{3.14}$$

a odgovarajuća kriterijska funkcija cilja je

$$\mathcal{F}_1(\Pi) = \sum_{a_i\in\pi_1} |c_1 - a_i| + \sum_{a_i\in\pi_2} |c_2 - a_i|. \tag{3.15}$$

U ovom slučaju vrijednost kriterijske funkcije $\mathcal{F}_1$ predstavlja sumu „apsolutnih odstupanja" točaka klastera π_1 do centra c_1 i točaka klastera π_2 do centra c_2.

Primjer 3.11. *Za skup $\mathcal{A} = \{1, 3, 4, 8\}$ i sve njegove particije odredit ćemo ℓ_1-centre i vrijednosti ℓ_1-funkcije cilja $\mathcal{F}_1$ (vidi Tablicu 3.3).*

π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$	
$\{1\}$	$\{3,4,8\}$	1	4	0+5=	5
$\{3\}$	$\{1,4,8\}$	3	4	0+7=	7
$\{4\}$	$\{1,3,8\}$	4	3	0+7=	7
$\{8\}$	$\{1,3,4\}$	8	3	0+3=	3
$\{1,3\}$	$\{4,8\}$	1	6	2+4=	6
$\{1,4\}$	$\{3,8\}$	1	3	3+5=	8
$\{1,8\}$	$\{3,4\}$	8	4	7+1=	8

Tablica 3.3: Biranje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{1,3,4,8\}$

Primijetite da se minimalna vrijednost funkcije $\mathcal{F}_1$ postiže na particiji $\Pi^\star = \{\{8\},\{1,3,4\}\}$ koja je također bila i LS-optimalna 2-particija (vidi Tablicu 3.2). U ovom primjeru to se slučajno dogodilo. Općenito, LS-optimalna particija ne mora se podudarati s ℓ_1-optimalnom particijom.

Zadatak 3.4. Konstruirajte primjer skupa $\mathcal{A} \subset \mathbb{R}$ kod kojega će se LS-optimalna i ℓ_1-optimalna particija međusobno razlikovati.

Primjedba 3.1. Broj svih k-particija skupa $\mathcal{A}$ s m elemenata može biti jako velik (vidi Tablicu 3.1). Međutim, u slučaju podataka s jednim obilježjem ($\mathcal{A} \subset \mathbb{R}$), očigledno je da se optimalna particija može očekivati između particija čiji se klasteri međusobno nastavljaju jedan na drugi. To znači da se svi elementi klastera π_2 nalaze desno od klastera π_1, svi elementi klastera π_3 nalaze se desno od klastera π_2, itd. (vidi [53], str. 161). Broj takvih particija znatno je manji (vidi Tablicu 3.4) i iznosi

$$\binom{m-1}{k-1}. \tag{3.16}$$

$\binom{m-1}{k-1}$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=8$	$k=10$
$m=10$	9	36	84	126	126	36	1
$m=30$	29	406	3 654	23 751	118 755	1 560 780	10 015 005
$m=50$	49	1 176	18 424	211 876	1 906 884	85 900 584	2 054 455 634

Tablica 3.4: Broj particija čiji se klasteri međusobno nastavljaju u ovisnosti o broju elemenata i broju klastera

Primjer 3.12. *Za dani skup* $\mathcal{A} = \{1, 2, 4, 5, 8, 9, 11\}$ *pokušat ćemo pronaći* ℓ_1*-optimalnu* 2*-particiju.*

Primijetimo najprije da ovaj skup ima $2^6 - 1 = 63$ različite particije pa bi ispitivanje svih ovih particija bilo zamorno. Podsjetimo se da je u slučaju skupa podataka s jednim obilježjem optimalnu particiju dovoljno tražiti između particija čiji se klasteri međusobno nastavljaju (vidi prethodnu Primjedbu 3.1). Zato se problem traženja optimalne particije u ovom slučaju svodi na ispitivanje samo $\binom{6}{1} = 6$ particija (vidi Tablicu 3.5).

π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$	
$\{1\}$	$\{2,4,5,8,9,11\}$	1	6	(0)+(4+2+1+2+3+5)=	17
$\{1,2\}$	$\{4,5,8,9,11\}$	1	8	(0+1)+(4+3+0+1+3)=	12
$\{1,2,4\}$	$\{5,8,9,11\}$	2	8	(1+0+2)+(3+0+1+3)=	10
$\{1,2,4,5\}$	$\{8,9,11\}$	3	9	(2+1+1+2)+(1+0+2)=	9
$\{1,2,4,5,8\}$	$\{9,11\}$	4	10	(3+2+0+1+4)+(1+1)=	12
$\{1,2,4,5,8,9\}$	$\{11\}$	4	11	(3+2+0+1+4+5)+(0)=	15

Tablica 3.5: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 4, 5, 8, 9, 11\}$

U nekim slučajevima pojavit će se mogućnost izbora više vrijednosti za centar klastera[2]. Kao što se vidi iz Tablice 3.5, optimalna particija je

$$\Pi^\star = \{\{1, 2, 4, 5\}, \{8, 9, 11\}\}, \qquad \mathcal{F}_1(\Pi^\star) = 9.$$

Primjedba 3.2. U slučaju primjene ℓ_1-metričke funkcije nije moguće definirati odgovarajući dualni problem onako kako smo to uradili za LS-kvazimetričku funkciju.

3.2.3 Formulacija problema grupiranja pomoću centara klastera

Korištenjem neke kvazimetričke funkcije $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ za dane realne brojeve $z_1, z_2 \in \mathbb{R}$, $z_1 \neq z_2$, primjenom **principa minimalnih udaljenosti** [31, 51, 53] možemo definirati particiju $\Pi = \{\pi_1, \pi_2\}$ skupa $\mathcal{A}$ na sljedeći

[2]Sjetite se (vidi str. 6) da kod određivanja medijana parnog broja podataka medijan može biti bilo koji broj iz intervala dvaju srednjih podataka.

način:

$$\pi_1 = \{a \in \mathcal{A} \colon d(z_1, a) \leq d(z_2, a)\}, \tag{3.17}$$
$$\pi_2 = \{a \in \mathcal{A} \colon d(z_2, a) < d(z_1, a)\}, \tag{3.18}$$

pri čemu treba voditi računa da svaki element skupa $\mathcal{A}$ pridružimo samo jednom klasteru. Za različite parove brojeva (z_1, z_2) općenito dobivamo različite particije. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može razmatrati kao sljedeći optimizacijski problem

$$\min_{z_1, z_2 \in \mathbb{R}} F(z_1, z_2), \qquad F(z_1, z_2) = \sum_{i=1}^{m} \min\{d(z_1, a_i), d(z_2, a_i)\}, \tag{3.19}$$

gdje je $F \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. Ovaj problem u literaturi se pojavljuje pod nazivom *k-median problem* (u ovom je slučaju $k = 2$) i ekvivalentan je problemu traženja optimalne particije na kojoj kriterijska funkcija cilja $\mathcal{F}$ zadana s (3.2) postiže globalni minimum.

Primijetite da je problem (3.19) optimizacijski problem za običnu realnu funkciju dviju realnih varijabli. Zato bi se ovaj problem mogao rješavati poznatim metodama globalne optimizacije (vidi primjerice [25, 27, 62]), ali je direktno rješavanje ovog optimizacijskog problema spomenutim metodama numerički vrlo zahtjevno. Zbog toga se u recentnoj literaturi mogu pronaći brojne specijalizacije (vidi primjerice [15, 43, 63]).

Pokažimo da je optimizacijski problem (3.19) ekvivalentan optimizacijskom problemu (3.3). Pokažimo najprije da se vrijednost funkcije F zadane u (3.19) na paru (z_1, z_2) podudara s vrijednošću funkcije $\mathcal{F}$ zadane s (3.2) na particiji $\Pi = \{\pi_1, \pi_2\}$, čiji klasteri imaju centre baš z_1 i z_2. Naime,

$$\begin{aligned} F(z_1, z_2) &= \sum_{a \in \pi_1(z_1, z_2)} \min\{d(z_1, a), d(z_2, a)\} + \sum_{a \in \pi_2(z_1, z_2)} \min\{d(z_1, a), d(z_2, a)\} \\ &= \sum_{a \in \pi_1(z_1, z_2)} d(z_1, a) + \sum_{a \in \pi_2(z_1, z_2)} d(z_2, a) = \mathcal{F}(\Pi). \end{aligned} \tag{3.20}$$

Primjer 3.13. *Zadan je sup* $\mathcal{A} = \{1, 3, 4, 8\}$ *iz Primjera 3.9, str. 30. Za izabrane parove brojeva* $(z_1, z_2) \in \mathbb{R}^2$ *primjenom formula* (3.17)–(3.18) *najprije ćemo odrediti pripadnu particiju s klasterima* π_1, π_2*, a nakon toga primjenom* ℓ_1*-metričke funkcije izračunati odgovarajuću vrijednost funkcije* F_1 *zadane s* (3.19) *(vidi Tablicu 3.6).*

z_1	z_2	π_1	π_2	$F_1(z_1, z_2)$	
2	5	$\{1,3\}$	$\{4,8\}$	(1+1)+(1+3)=	6
1	6	$\{1,3\}$	$\{4,8\}$	(0+2)+(2+2)=	6
0	5	$\{1\}$	$\{3,4,8\}$	(1)+(2+1+3)=	7
2	7	$\{1,3,4\}$	$\{8\}$	(1+1+2)+(1)=	5
3	8	$\{1,3,4\}$	$\{8\}$	(2+0+1)+(0)=	3
4	9	$\{1,3,4\}$	$\{8\}$	(3+1+0)+(1)=	5

Tablica 3.6: Izračunavanje vrijednosti funkcije F_1 za skup $\mathcal{A} = \{1,3,4,8\}$

Najniža vrijednost funkcije F_1 postignuta je na paru brojeva $(3,8)$ koji preko (3.17)–(3.18) generiraju optimalnu particiju $\Pi^\star = \{\{1,3,4\},\{8\}\}$. Primijetite da i neki drugi parovi brojeva generiraju tu istu optimalnu particiju, ali je vrijednost funkcije F_1 na paru $(3,8)$ najmanja.

π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$	
$\{1,3\}$	$\{4,8\}$	2	6	2+4=	6
$\{1\}$	$\{3,4,8\}$	1	4	0+5=	5
$\{1,3,4\}$	$\{8\}$	8	3	0+3=	3

Tablica 3.7: Vrijednosti funkcije $\mathcal{F}_1$ na particijama iz Tablice 3.3

Primijetite da je particija $\Pi^\star$ također optimalna i u smislu kriterijske funkcije cilja $\mathcal{F}_1$ zadane s (3.15) i da su centri klastera optimalne particije zadani s (3.14) baš brojevi $(3,8)$, što potvrđuje tvrdnju (3.20). Podsjetimo se kakve su bile vrijednosti funkcije cilja $\mathcal{F}_1$ zadane s (3.15) na ovih nekoliko particija koje su se pojavile u Tablici 3.6. U tu svrhu iz Tablice 3.3 konstruirajmo Tablicu 3.7

Primijetite da je $F_1(1,6) = \mathcal{F}_1(\{\{1,3\},\{4,8\}\})$ jer brojevi 1 i 6 mogu biti centri klastera $\{1,3\}$, odnosno $\{4,8\}$, ali $F_1(0,5) \neq \mathcal{F}_1(\{\{1\},\{3,4,8\}\})$ jer brojevi 0 i 5 nisu centri klastera $\{1\}$, odnosno $\{3,4,8\}$.

Zadatak 3.5. Za skup $\mathcal{A} = \{1,3,4,8\}$ iz prethodnog primjera primjenom ℓ_1-metričke funkcije pokušajte za svaku od 7 particija odrediti područje u ravnini gdje treba birati par brojeva (z_1, z_2) koji preko (3.17)–(3.18) generiraju tu particiju. Rezultate prikažite grafički.

Zadatak 3.6. Za skup $\mathcal{A} = \{1,3,4,8\}$ iz prethodnog primjera primjenom LS-kvazimetričke funkcije pokušajte za svaku od 7 particija odrediti područje u

ravnini gdje treba birati par brojeva (z_1, z_2) koji preko (3.17)–(3.18) generiraju tu particiju. Rezultate prikažite grafički. Za rješavanje ovog zadatka bit će potreno izraditi odgovarajući računalni program.

Zadatak 3.7. Zadan je skup $\mathcal{A} = \{1, 2, 4, 5, 8, 9, 11\}$ iz Primjera 3.12, str. 36. Izračunajte $F_1(1, 6)$, $F_1(9, 11)$, $F_1(3, 9)$, $F_1(3, 10)$ tako da kao u Primjeru 3.13 za svaki par brojeva najprije odredite pripadne particije s klasterima π_1, π_2 a nakon toga primjenom ℓ_1-metričke funkcije izračunate odgovarajuće vrijednosti funkcije F_1 zadane s (3.19). Napravite odgovarajuću analizu rezultata.

Rješenje: $F_1(1, 6) = 14$, $F_1(9, 11) = 25$, $F_1(3, 9) = 9$, $F_1(3, 10) = 10$.

3.3 Grupiranje u k klastera na osnovi jednog obilježja

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup koji na osnovi jednog obilježja treba grupirati u k klastera $\pi_1, \ldots, \pi_k$, koji zadovoljavaju Definiciju 3.1, str. 23. Primjerice, dane u godini možemo grupirati u tri klastera prema prosječnoj dnevnoj temperaturi izraženoj u °C: klaster hladnih dana, klaster dana s umjerenom temperaturom, klaster toplih dana. Svaki element $a \in \mathcal{A}$ temeljem tog obilježja reprezentirat ćemo jednim realnim brojem kojeg ćemo također označavati s a. Zato ćemo nadalje govoriti o multiskupu podataka-realnih brojeva $\mathcal{A} = \{a_1, \ldots, a_m\}$ među kojima može biti i jednakih elemenata. Možemo također koristiti i termine *m-torka realnih brojeva* ili *konačni niz realnih brojeva.*

Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar c_j na sljedeći način

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} d(x, a), \qquad j = 1, \ldots, k. \tag{3.21}$$

Nadalje, ako na skupu svih particija $\mathcal{P}(\mathcal{A}; k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo kriterijsku funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a), \tag{3.22}$$

onda optimalnu k-particiju $\Pi^\star$ tražimo rješavanjem sljedećeg optimizacijskog problema

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A}; k)} \mathcal{F}(\Pi). \tag{3.23}$$

Primijetite da na taj način optimalna particija $\Pi^\star$ ima svojstvo da je suma „rasipanja" (suma odstupanja) elemenata klastera oko njegovog centra minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i međusobnu razdvojenost (separiranost) klastera.

3.3.1 Princip najmanjih kvadrata

Neka je $\mathcal{A} \subset \mathbb{R}$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija. Ako je $d_{LS}\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$, $d_{LS}(x,y) = (x-y)^2$ LS-kvazimetrička funkcija, slično kao u t. 3.2.1, str. 31, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ nazivaju se centroidi i određuju se na sljedeći način

$$c_j = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{a\in\pi_j} (x-a)^2 = \frac{1}{|\pi_j|} \sum_{a\in\pi_j} a, \qquad j = 1, \ldots, k, \tag{3.24}$$

a funkcija cilja (3.22) definirana je s

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} (c_j - a)^2. \tag{3.25}$$

Primjer 3.14. *Zadan je skup $\mathcal{A} = \{2, 4, 8, 10, 16\}$. Treba pronaći sve njegove 3-particije koje zadovoljavaju Definiciju 3.1 i koje se međusobno nastavljaju jedna na drugu. Za njih treba odrediti pripadne centroide i vrijednosti kriterijske funkcije cilja $\mathcal{F}_{LS}$.*

Prema Stirlingovoj formuli (3.1), broj svih 3-particija skupa $\mathcal{A}$ je 25. Međutim, broj 3-particija istog skupa koje se nastavljaju jedna na drugu prema (3.16) iznosi samo $\binom{5-1}{3-1} = \frac{4!}{2!\cdot 2!} = 6$ (vidi Tablicu 3.8).

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
{2}	{4}	{8,10,16}	2	4	11.33	0+0+34.67=34.67	36+16+33.33=85.33
{2}	{4,8}	{10,16}	2	6	13	0+8+18=26	36+8+50=94
{2}	{4,8,10}	{16}	2	7.33	16	0+18.67+0=18.67	36+0+64=100
{2,4}	{8}	{10,16}	3	8	13	2+0+18=20	50+0+50=100
{2,4}	{8,10}	{16}	3	9	16	2+2+0=4	50+2+64=116
{2,4,8}	{10}	{16}	4.67	10	16	18.67+0+0=18.67	33.33+4+64=101.33

Tablica 3.8: 3-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$

Kao što se vidi iz Tablice 3.8, LS-optimalna 3-particija u ovom slučaju je $\{\{2, 4\}, \{8, 10\}, \{16\}\}$ jer na njoj funkcija cilja $\mathcal{F}_{LS}$ zadana s (3.25) postiže najmanju vrijednost (globalni minimum).

Zadatak 3.8. Zadan je skup $\mathcal{A} = \{1, 4, 5, 8, 10, 12, 15\}$. Koliko ovaj skup ima 3-particija, a koliko 3-particija čiji se klasteri međusobno nastavljaju? Ispišite sve 3-particije skupa $\mathcal{A}$ čiji se klasteri međusobno nastavljaju i među njima pronađite LS-optimalnu 3-particiju.

Rješenje: Broj svih particija je 301, a broj svih particija čiji se klasteri međusobno nastavljaju je 15. LS-optimalna 3-particija je $\Pi^\star = \{\{1, 4, 5\}, \{8, 10\}, \{12, 15\}\}$. $\mathcal{F}(\Pi^\star) = \frac{91}{6} = 15.1667$.

Dualni problem

Razmotrimo najprije sljedeći pomoćni rezultat. Sljedeća lema pokazuje da je „rasipanje" skupa $\mathcal{A}$ oko njegovog centra c jednako zbroju „rasipanja" klastera π_j, $j = 1, \ldots, k$ oko njihovih centara c_j, $j = 1, \ldots, k$ i težinskoj sumi kvadrata odstupanja centra c od centara c_j, pri čemu su težine određene veličinom skupova π_j.

Lema 3.1. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup podataka, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija s klasterima $\pi_1, \ldots, \pi_k$. Neka je nadalje*

$$c = \frac{1}{m}\sum_{i=1}^{m} a_i, \qquad c_j = \frac{1}{m_j}\sum_{a\in\pi_j} a, \quad j = 1, \ldots, k, \tag{3.26}$$

gdje je $m_j = |\pi_j|$ broj elemenata klastera π_j. Tada vrijedi

$$\sum_{i=1}^{m}(c - a_i)^2 = \mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi), \tag{3.27}$$

gdje je

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k}\sum_{a\in\pi_j}(c_j - a)^2, \tag{3.28}$$

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j(c_j - c)^2. \tag{3.29}$$

Dokaz. Primijetimo najprije da za svaki $x \in \mathbb{R}$ prema Lemi 5.1, str. 98, vrijedi

$$\sum_{a\in\pi_j}(x - a)^2 = \sum_{a\in\pi_j}(c_j - a)^2 + m_j(c_j - x)^2, \quad j = 1, \ldots, k. \tag{3.30}$$

Ako u (3.30) umjesto x stavimo $c = \frac{1}{m}\sum_{i=1}^{m} a_i$ i zbrojimo sve jednakosti, dobivamo (3.27). □

Zadatak 3.9. Napišite formule za centroid skupa $\mathcal{A}$ i kriterijske funkcije cilja $\mathcal{F}$ i $\mathcal{G}$ za slučaj skupa podataka $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$.

Rješenje: $\mathcal{G}(\Pi) = \sum_{j=1}^{k} \big(\sum_{\pi_j} w_s\big)(c_j - c)^2$.

Slično kao u slučaju dva klastera (str. 32) iz Leme 3.1 neposredno slijedi tvrdnja sljedećeg teorema [15, 68].

Teorem 3.1. *Uz oznake kao u Lemi 3.1 vrijedi:*

$$(i) \quad \operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi) =: \Pi^\star,$$

$$(ii) \quad \min_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \mathcal{F}_{LS}(\Pi^\star) \quad \& \quad \max_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi) = \mathcal{G}(\Pi^\star).$$

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije $\mathcal{F}_{LS}$ zadane s (3.25), možemo maksimizirati funkciju $\mathcal{G}$

$$\operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j (c_j - c)^2. \tag{3.31}$$

Optimizacijski problem (3.31) zovemo dualni problem u odnosu na optimizacijski problem $\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi)$.

Primjer 3.15. *U Tablici 3.8 također možemo vidjeti i odgovarajuće vrijednosti kriterijske funkcije cilja $\mathcal{G}$ za sve particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$.*

Kao što se može vidjeti, kriterijska funkcija cilja $\mathcal{G}$ postiže najveću vrijednost na LS-optimalnoj 3-particiji $\{\{2, 4\}, \{8, 10\}, \{16\}\}$, što je u skladu s prethodnim Teoremom 3.1.

3.3.2 Princip najmanjih apsolutnih odstupanja

Neka je $\mathcal{A} \subset \mathbb{R}$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija. Ako je $d_1 \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, $d_1(x, y) = |x - y|$, ℓ_1-metrička funkcija slično kao u t. 3.2.2, str. 34, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{a\in\pi_j} |x - a| = \operatorname{med}(\pi_j), \quad j = 1, \ldots, k, \tag{3.32}$$

a funkcija cilja (3.22) s

$$\mathcal{F}_1(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} |c_j - a|. \tag{3.33}$$

Ako pri tome iskoristimo (3.34), onda za izračunavanje funkcije cilja (3.33) nije potrebno poznavati centre klastera (3.32), što može značajno ubrzati računski proces.

Primjer 3.16. *Zadan je skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *kao u Primjeru 3.14, str. 40. Treba pronaći sve njegove tročlane particije koje zadovoljavaju Definiciju 3.1 i koje se međusobno nastavljaju jedna na drugu.*

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_1(\Pi)$
{2}	{4}	{8,10,16}	2	4	10	0+0+8=8
{2}	{4,8}	{10,16}	2	6	13	0+4+6=10
{2}	{4,8,10}	{16}	2	8	16	0+6+0=6
{2,4}	{8}	{10,16}	3	8	13	2+0+6=8
{2,4}	{8,10}	{16}	3	9	16	2+2+0=4
{2,4,8}	{10}	{16}	4	10	16	6+0+0=6

Tablica 3.9: Particije skupa $\mathcal{A}$ čiji se klasteri međusobno nastavljaju

Za njih treba odrediti pripadne centre i vrijednosti funkcije cilja $\mathcal{F}_1$ *uz primjenu* ℓ_1*-metričke funkcije, te pronaći globalno* ℓ_1*-optimalnu* 3*-particiju.*

Broj svih tročlanih particija čiji se klasteri međusobno nastavljaju je $\binom{m-1}{k-1} = 6$, a kao što se vidi iz Tablice 3.9, ℓ_1-optimalna 3-particija je $\{\{2, 4\}, \{8, 10\}, \{16\}\}$ jer na njoj funkcija cilja $\mathcal{F}_1$ zadana s (3.33) postiže najmanju vrijednost (globalni minimum).

Zadatak 3.10. Između svih particija skupa $\mathcal{A} = \{1, 4, 5, 8, 10, 12, 15\}$ iz Zadatka 3.8, str. 41, pronađite ℓ_1-optimalnu 3-particiju.

Zadatak 3.11. Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ konačan rastući niz realnih brojeva. Pokažite da vrijedi

$$\sum_{i=1}^{m} |a_i - \mathrm{med}(\mathcal{A})| = \sum_{i=1}^{\lceil \frac{m}{2} \rceil} (a_{m-i+1} - a_i)\,, \tag{3.34}$$

gdje je $\lceil x \rceil$[3] jednak x ako je x cijeli broj, a $\lceil x \rceil$ je najmanji cijeli broj veći od x ako x nije cijeli broj. Primjerice, $\lceil 20 \rceil = 20$, ali $\lceil 20.3 \rceil = 21$.

3.3.3 Grupiranje podataka s težinama

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ na pravcu, pri čemu je svakom podatku a_i pridružena odgovarajuća težina $w_i > 0$. Težine w_i općenito su realni brojevi. Kriterijska funkcija cilja (3.22) tada postaje

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} w_i d(c_j, a_i). \tag{3.35}$$

Specijalno, kod primjene LS-kvazimetričke funkcije centri c_j klastera π_j su težinske aritmetičke sredine podataka iz klastera π_j

$$c_j = \frac{1}{\kappa_j} \sum_{a_i \in \pi_j} w_i a_i, \qquad \kappa_j = \sum_{a_i \in \pi_j} w_i, \tag{3.36}$$

a kod primjene ℓ_1-metričke funkcije centri c_j klastera π_j su težinski medijani podataka koji pripadaju klasteru π_j [50, 76]

$$c_j = \operatorname*{med}_{a_i \in \pi_j}(w_i, a_i). \tag{3.37}$$

Primjer 3.17. *Promatrajmo ponovo skup* $\mathcal{A} = \{1, 4, 5, 8, 10, 12, 15\}$ *iz Zadatka 3.8, str. 41. Svim podacima, osim posljednjeg, pridružimo težinu* 1, *a posljednjem težinu* 3. *Sada LS-optimalna* 3*-particija postaje* $\Pi^\star = \{\{1, 4, 5\}, \{8, 10, 12\}, \{15\}\}$ *s centroidima:* $\frac{10}{3}, 10, 15$ *i vrijednosti funkcije cilja* $\mathcal{F}(\Pi^\star) = \frac{50}{3} = 16.667$.

U slučaju primjene ℓ_1-metričke funkcije za određivanje centara klastera treba znati izračunati težinski medijan podataka. Kao što smo naveli u t. 2.1.3, str. 8, to može biti složen postupak. Ako su težine cijeli brojevi, problem se može svesti na određivanje običnog medijana podataka (vidi Primjer 2.4, str. 9). Ako težine nisu cijeli brojevi, množenjem nekim brojem i zaokruživanjem mogu se svesti na cijele brojeve.

Zadatak 3.12. Pronađite ℓ_1-optimalnu 3-particiju skupa $\mathcal{A}$ iz prethodnog primjera u slučaju kada su sve težine jednake 1 i u slučaju ako su podacima pridružene težine kao u prethodnom primjeru.

[3] U programskom sustavu *Mathematica* veličina $\lceil x \rceil$ dobiva se kao `Ceiling[x]`, a veličina $\lfloor x \rfloor$ kao `Floor[x]`.

3.3.4 Formulacija problema grupiranja pomoću centara klastera

Slično kao u t. 3.2.3, str. 36, za dani skup međusobno različitih brojeva $z_1, \ldots, z_k \in \mathbb{R}$, primjenom **principa minimalnih udaljenosti** [31, 51, 53] možemo definirati particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_j = \{a \in \mathcal{A} : d(z_j, a) \leq d(z_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k, \quad (3.38)$$

pri čemu treba voditi računa o tome da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru. Za različite k-torke brojeva $(z_1, \ldots, z_k)$ općenito dobivamo različite particije. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći optimizacijski problem

$$\min_{z_1,\ldots,z_k \in \mathbb{R}} F(z_1, \ldots, z_k), \qquad F(z_1, \ldots, z_k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} d(z_j, a_i), \quad (3.39)$$

gdje je $F\colon \mathbb{R}^k \to \mathbb{R}_+$. Funkcija F je realna funkcija k realnih varijabli, općenito nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma [6, 72]. U principu, problem (3.39) mogao bi se rješavati poznatim metodama globalne optimizacije [25, 27, 62]), ali zbog izuzetne numeričke zahtjevnosti to se obično ne radi. Umjesto toga, u literaturi se mogu pronaći brojne specijalizacije (vidi primjerice [15, 43, 63]).

Optimizacijski problem (3.39) u literaturi se može pronaći pod nazivom *k-median problem* [26, 31, 72] i ekvivalentan je optimizacijskom problemu (3.23). Naime, vrijednost funkcije F zadane s (3.39) na k-torki $(z_1, \ldots, z_k)$ podudara se s vrijednošću funkcije $\mathcal{F}$ zadane s (3.22) na particiji $\Pi = \{\pi_1, \ldots, \pi_k\}$, čiji klasteri imaju centre baš $z_1, \ldots, z_k$

$$\begin{aligned} F(z_1, \ldots, z_k) &= \sum_{i=1}^{m} \min\{d(z_1, a_i), \ldots, d(z_k, a_i)\} \\ &= \sum_{j=1}^{k} \sum_{a_i \in \pi_j} \min\{d(z_1, a_i), \ldots, d(z_k, a_i)\} \\ &= \sum_{j=1}^{k} \sum_{a_i \in \pi_j} d(z_j, a_i) = \mathcal{F}(\Pi). \end{aligned}$$

3.4 Grupiranje u k klastera na osnovi dva ili više obilježja

Neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup, koji u smislu Definicije 3.1 treba grupirati u $1 \leq k \leq m$ nepraznih disjunktnih klastera.

Primjerice, skup $\mathcal{A} \subset \mathbb{R}^2$ iz Primjera 3.2, str. 25, ima dva obilježja (apscise i ordinate točaka), a možemo ga grupirati u $2, 3, 4, 5, 6$ ili više klastera (vidi Sliku 3.1).

Neka je $\Pi \in \mathcal{P}(\mathcal{A}; k)$ neka particija skupa $\mathcal{A}$. Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar c_j na sljedeći način

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k. \tag{3.40}$$

Nadalje, potpuno analogno kao i ranije, ako na skupu svih particija $\mathcal{P}(\mathcal{A}; k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo kriterijsku funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a), \tag{3.41}$$

onda optimalnu k-particiju $\Pi^\star$ tražimo rješavanjem sljedećeg optimizacijskog problema

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A}; k)} \mathcal{F}(\Pi). \tag{3.42}$$

Primijetite da na taj način optimalna particija $\Pi^\star$ ima svojstvo da je suma „rasipanja" (suma d-udaljenosti elemenata klastera do svog centra) minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i međusobnu razdvojenost (separiranost) klastera.

3.4.1 Princip najmanjih kvadrata

Neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \in \mathbb{R}^n \colon i = 1, \ldots, m\}$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova particija. Ako je $d_{LS}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d_{LS}(a^1, a^2) = \|a^1 - a^2\|_2^2$ LS-kvazimetrička funkcija, slično kao u t. 3.2.1, str. 31, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ nazivaju se centroidi i određuju se na sljedeći način

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} \|x - a\|_2^2 = \tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a = \Big(\tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a_1, \ldots, \tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a_n \Big), \tag{3.43}$$
$$j = 1, \ldots, k,$$

pri čemu $\sum_{a \in \pi_j} a_1$ označava sumu prvih komponenti svih elemenata klastera π_j, a $\sum_{a \in \pi_j} a_n$ sumu n-tih komponenti svih elemenata klastera π_j. Funkcija

cilja (3.41) u ovom slučaju zadana je s

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|c_j - a\|_2^2 \tag{3.44}$$

Primjer 3.18. *Neka je* $\mathcal{A} = \{a^1 = (0,0),\ a^2 = (1,0),\ a^3 = (1,1),\ a^4 = (0,1)\}$ *skup točaka u ravnini. Broj svih njegovih* 2*-particija je* $\mathcal{P}(\mathcal{A}; 2) = 2^{4-1} - 1 = 7$*, a prikazane su u Tablici 3.10. Između svih* 2*-particija skupa* $\mathcal{A}$ *potražimo LS-optimalnu.*

Skup $\mathcal{A}$ sastoji se od elemenata s dva obilježja (koordinate), pa se jednostavnije može zapisati kao $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, 4\}$ i grafički prikazati u ravnini (vidi Sliku 3.7). LS-udaljenost elemenata $a^1 = (x_1, y_1)$ i $a^2 = (x_2, y_2)$ računa se na sljedeći način $d_{LS}(a^1, a^2) = \|a^1 - a^2\|_2^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$.

Prema (3.1), skup $\mathcal{A}$ ima 7 različitih particija. Neka je $\Pi = \{\pi_1, \pi_2\}$ bilo koja od njih. Centroidi njenih klastera zadani su s

$$c_1 = \tfrac{1}{|\pi_1|} \sum_{a \in \pi_1} a, \qquad c_2 = \tfrac{1}{|\pi_2|} \sum_{a \in \pi_2} a,$$

a odgovarajuća LS-funkcija cilja je

$$\mathcal{F}_{LS}(\Pi) = \sum_{a \in \pi_1} \|c_1 - a\|_2^2 + \sum_{a \in \pi_2} \|c_2 - a\|_2^2.$$

U ovom slučaju vrijednost kriterijske funkcije $\mathcal{F}_{LS}$ predstavlja sumu „kvadrata udaljenosti" točaka klastera π_1 do njegovog centroida c_1 i točaka klastera π_2 do njegovog centroida c_2.

U Tablici 3.10 navedene su sve particije s centroidima odgovarajućih klastera i vrijednostima kriterijske funkcije cilja $\mathcal{F}_{LS}$.

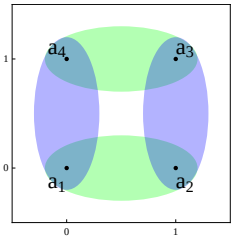
π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$		$\mathcal{G}(\Pi)$
$\{a^1\}$	$\{a^2,a^3,a^4\}$	a^1	$\left(\frac{2}{3},\frac{2}{3}\right)^T$	$0+\frac{4}{3}$	≈ 1.3	$\frac{1}{2}+\frac{1}{6}\approx 0.67$
$\{a^2\}$	$\{a^1,a^3,a^4\}$	a^2	$\left(\frac{1}{3},\frac{2}{3}\right)^T$	$0+\frac{4}{3}$	≈ 1.3	$\frac{1}{2}+\frac{1}{6}\approx 0.67$
$\{a^3\}$	$\{a^1,a^2,a^4\}$	a^3	$\left(\frac{1}{3},\frac{1}{3}\right)^T$	$0+\frac{4}{3}$	≈ 1.3	$\frac{1}{2}+\frac{1}{6}\approx 0.67$
$\{a^4\}$	$\{a^1,a^2,a^3\}$	a^4	$\left(\frac{2}{3},\frac{1}{3}\right)^T$	$0+\frac{4}{3}$	≈ 1.3	$\frac{1}{2}+\frac{1}{6}\approx 0.67$
$\{a^1,a^2\}$	$\{a^3,a^4\}$	$\left(\frac{1}{2},0\right)^T$	$\left(\frac{1}{2},1\right)^T$	$\frac{1}{2}+\frac{1}{2}$	$=1$	$\frac{1}{2}+\frac{1}{2}=1$
$\{a^1,a^4\}$	$\{a^2,a^3\}$	$\left(0,\frac{1}{2}\right)^T$	$\left(1,\frac{1}{2}\right)^T$	$\frac{1}{2}+\frac{1}{2}$	$=1$	$\frac{1}{2}+\frac{1}{2}=1$
$\{a^1,a^3\}$	$\{a^2,a^4\}$	$\left(\frac{1}{2},\frac{1}{2}\right)^T$	$\left(\frac{1}{2},\frac{1}{2}\right)^T$	$1+1$	$=2$	$0+0=0$

Tablica 3.10: Particije, centri i funkcije cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$ iz Primjera 3.18

Kao što se vidi iz Tablice 3.10, dvije su particije:

$$\{\{a^1,a^2\},\{a^3,a^4\}\} \quad \text{i} \quad \{\{a^1,a^4\},\{a^2,a^3\}\}$$

LS-optimalne jer na njima kriterijska funkcija cilja $\mathcal{F}_{LS}$ postiže globalni minimum (vidi također Sliku 3.7).



Slika 3.7: LS-optimalne particije

Dualni problem za podatke s dva ili više obilježja

Analogno, kao u t. 3.2.1, str. 31, korištenjem Leme 5.1, str. 98, može se pokazati da vrijedi

$$\sum_{i=1}^{m}\|a^i-c\|_2^2=\sum_{j=1}^{k}\sum_{a\in\pi_j}\|c_j-a\|_2^2+\sum_{j=1}^{k}m_j\|c_j-c\|_2^2, \tag{3.45}$$

gdje je $c=\frac{1}{m}\sum_{i=1}^{m}a^i$ centroid cijelog skupa $\mathcal{A}$, a $m_j=|\pi_j|$ broj elemenata klastera π_j. Jednakost (3.45) dozvoljava nam da umjesto minimizacije funk-

cije $\mathcal{F}_{LS}$ zadane s (3.44) LS-optimalnu particiju također možemo potražiti rješavanjem *dualnog optimizacijskog problema*

$$\operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j \|c - c_j\|_2^2. \tag{3.46}$$

Određenim prilagođavanjem [15] problem se svodi na poznate probleme i metode linearne algebre.

Primjer 3.19. *Kod Primjera 3.18, str. 47, može se razmatrati i odgovarajući dualni problem.*

Specijalno, u ovom slučaju jednakost (3.45) glasi

$$\sum_{i=1}^{m} \|c - a^i\|_2^2 = \Big(\sum_{a\in\pi_1} \|c_1 - a\|_2^2 + \sum_{a\in\pi_2} \|c_2 - a\|_2^2 \Big) + \big(m_1\|c_1 - c\|_2^2 + m_2\|c_2 - c\|_2^2 \big),$$

a dualni optimizacijski problem (3.46) postaje

$$\operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = m_1\|c_1 - c\|_2^2 + m_2\|c_2 - c\|_2^2.$$

Za svaku 2-particiju u Tablici 3.10, str. 48, plavom bojom prikazane su vrijednosti dualne kriterijske funkcije cilja $\mathcal{G}$. Kao što se vidi, funkcija $\mathcal{G}$ prima maksimalnu vrijednost na LS-optimalnim 2-particijama $\{\{a^1, a^2\}, \{a^3, a^4\}\}$ i $\{\{a^1, a^4\}, \{a^2, a^3\}\}$, na kojima je kriterijska funkcija $\mathcal{F}_{LS}$ zadana s (3.44) primila minimalnu vrijednost.

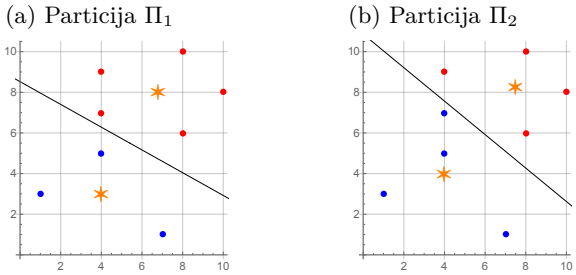
Primjer 3.20. *Skup* $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 8\} \subset \mathbb{R}^2$ *zadan je s*

i	1	2	3	4	5	6	7	8
x_i	1	4	4	4	7	8	8	10
y_i	3	5	7	9	1	6	10	8

Uz primjenu LS-kvazimetričke funkcije za 2-particije:

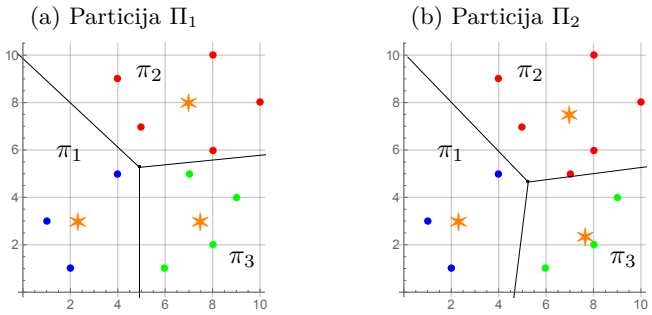
$$\Pi_1 = \{\{a^1, a^2, a^5\}, \{a^3, a^4, a^6, a^7, a^8\}\},$$
$$\Pi_2 = \{\{a^1, a^2, a^3, a^5\}, \{a^4, a^6, a^7, a^8\}\},$$

prikazane na Slici 3.8, čiji su klasteri označeni plavom odnosno crvenom bojom, treba odrediti centroide i odgovarajuće vrijednosti funkcija cilja $\mathcal{F}_{LS}$ *i* $\mathcal{G}$ *te na osnovi toga ustanoviti koja je particija bliža optimalnoj.*



Slika 3.8: Dvije particije skupa $\mathcal{A}$ iz Primjera 3.20

Za particiju Π_1 dobivamo: $c_1 = (4, 3)$, $c_2 = (6.8, 8)$, $\mathcal{F}_{LS} = 26 + 38.8 = 64.8$ i $\mathcal{G} = 61.575$, a za particiju Π_2: $c_1 = (4, 4)$, $c_2 = (7.5, 8.25)$, $\mathcal{F}_{LS} = 38+27.75 = 65.75$ i $\mathcal{G} = 60.625$. Dakle, 2-particija Π_1 bliža je LS-optimalnoj. Primjenom *Mathematica*-modula `WKMeans[]`, str. 175, provjerite je li to ujedno i globalno LS-optimalna 2-particija. Primijetite (formula (3.1)) da u ovom slučaju ukupno postoji $2^7 - 1 = 127$ različitih 2-particija.



Slika 3.9: Usporedba dviju particija iz Zadatka 3.13

Zadatak 3.13. Zadan je skup $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, m\}$ prikazan na Slici 3.9, gdje je

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	2	4	4	5	6	7	8	8	8	9	10
y_i	3	1	5	9	7	1	5	2	6	10	4	8

Treba ustanoviti na kojoj od dviju niže navedenih 3-particija LS-funkcija cilja $\mathcal{F}_{LS}$ zadana s (3.44) prima manju vrijednost.

Rješenje:

$\Pi_1 = \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^9, a^{10}, a^{12}\}, \{a^6, a^7, a^8, a^{11}\}\}$ ... Slika 3.9a

$\Pi_2 = \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^7, a^9, a^{10}, a^{12}\}, \{a^6, a^8, a^{11}\}\}$ ... Slika 3.9b

$\Pi_1 : c_1 = (2.33, 3)$, $c_2 = (7, 8)$, $c_3 = (7.5, 3)$; $\mathcal{F}_{LS} = 12.67 + 34 + 15 = 61.67$;
$\mathcal{G} = 127.25$,
$\Pi_2 : c_1 = (2.33, 3)$, $c_2 = (7, 7.5)$, $c_3 = (7.67, 2.33)$; $\mathcal{F}_{LS} = 12.67+41.5+9.33=63.5$;
$\mathcal{G} = 125.42$.

Dakle, manja vrijednost LS-funkcije cilja $\mathcal{F}_{LS}$ (i veća vrijednost dualne funkcije $\mathcal{G}$) postiže se na 3-particiji Π_1 pa nju smatramo LS-optimalnijom. Primjenom *Mathematica*-modula `WKMeans[]`, str. 175, provjerite je li to ujedno i globalno LS-optimalna 3-particija

3.4.2 Princip najmanjih apsolutnih odstupanja

Neka je $\mathcal{A} \subset \mathbb{R}^n$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija. Ako je $d_1 \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d_1(x, y) = \|x - y\|_1$, ℓ_1-metrička funkcija, slično kao u t. 3.2.2, str. 34, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} \|x - a\|_1 = \left(\operatorname*{med}_{a \in \pi_j} a_1, \ldots, \operatorname*{med}_{a \in \pi_j} a_n \right) =: \operatorname{med}(\pi_j), \quad j = 1, \ldots, k, \tag{3.47}$$

pri čemu $\operatorname*{med}_{a \in \pi_j} a_1$ označava medijan prvih komponenti svih elemenata klastera π_j, a $\operatorname*{med}_{a \in \pi_j} a_n$ medijan n-tih komponenti svih elemenata klastera π_j. ℓ_1-funkcija cilja (3.41) u ovom je slučaju zadana s

$$\mathcal{F}_1(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|c_j - a\|_1 \tag{3.48}$$

Primjer 3.21. *Za skup $\mathcal{A}$ iz Primjera 3.20, str. 49 treba odrediti ℓ_1-optimalnu 2-particiju.*

Elementi ovog skupa imaju dva obilježja (koordinate točaka), a ℓ_1 udaljenost elemenata $a^1 = (x_1, y_1)$, $a^2 = (x_2, y_2)$ određuje se formulom $d_1(a^1, a^2) = \|a^1 - a^2\|_1 = |x_1 - x_2| + |y_1 - y_2|$. Centri klastera π_1, π_2 particije $\Pi = \{\pi_1, \pi_2\}$ zadani su s

$$c_1 \in \operatorname{med}(\pi_1), \qquad c_2 \in \operatorname{med}(\pi_2),$$

a kriterijska funkcija cilja $\mathcal{F}_1$ zadana je s

$$\mathcal{F}_1(\Pi) = \sum_{a \in \pi_1} \|c_1 - a\|_1 + \sum_{a \in \pi_2} \|c_2 - a\|_1,$$

i predstavlja sumu „rasipanja" (sumu ℓ_1 udaljenosti) točaka klastera π_1 do centra c_1 i točaka klastera π_2 do centra c_2.

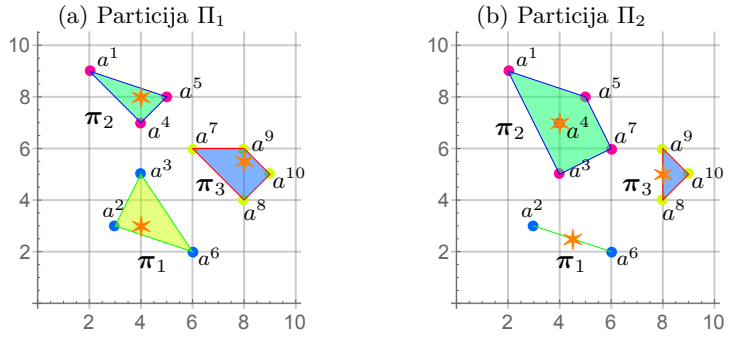
Na osnovi Slike 3.8 očekujemo da bi se ℓ_1-optimalna 2-particija mogla pojaviti između particija navedenih u Tablici 3.11. Za ove particije izračunati su centri odgovarajućih klastera i vrijednosti funkcije cilja $\mathcal{F}_1$ zadane s (3.48). Na osnovi ovih izračuna zaključujemo da bi ℓ_1-optimalna 2-particija mogla biti particija Π_1 (vidi t. 9.2.2, str. 167, i odgovarajući *Mathematica*-program `LS+LAD-grupiranje.nb` raspoloživ na `http://www.mathos.unios.hr/images/homepages/scitowsk/Programi.zip`).

	π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$
Π_1	$\{a^1, a^2, a^3, a^4\}$	$\{a^5, a^6, a^7, a^8\}$	(4,6)	(8,7)	(11)+(14)=25
Π_2	$\{a^1, a^2, a^5\}$	$\{a^3, a^4, a^6, a^7, a^8\}$	(4,3)	(8,8)	(10)+(16)=26
Π_3	$\{a^1, a^2, a^3, a^4, a^6, a^7, a^8\}$	$\{a^5\}$	(4,7)	(7,1)	(30)+(0)=30

Tablica 3.11: Nekoliko particija skupa $\mathcal{A}$ iz Primjera 3.20

Naravno da ovakvo zaključivanje nije utemeljeno na obranjivim argumentima. Korektan zaključak bit će moguće donijeti tek primjenom neke metode za traženje optimalne particije (t. 4, str. 57; t. 5, str. 91).

Zadatak 3.14. Na particije iz Zadatka 3.13, str. 50, primijenite princip najmanjih apsolutnih odstupanja.



Slika 3.10: Usporedba dviju particija

Primjer 3.22. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 10\}$, *gdje je*

i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	4	4	5	6	6	8	8	9
y_i	9	3	5	7	8	2	6	4	6	5

Treba ustanoviti na kojoj od dvije niže navedene 3*-particije* ℓ_1*-funkcija cilja* (3.48) *prima manju vrijednost.*

$$\Pi_1 = \{\{a^2, a^3, a^6\}, \{a^1, a^4, a^5\}, \{a^7, a^8, a^9, a^{10}\}\} \quad \ldots \quad \textit{Slika 3.10a,}$$
$$\Pi_2 = \{\{a^2, a^6\}, \{a^1, a^3, a^4, a^5, a^7\}, \{a^8, a^9, a^{10}\}\} \quad \ldots \quad \textit{Slika 3.10b.}$$

Niže su izračunati ℓ_1-centri pojedinih klastera u obje particije i vrijednost funkcije cilja na obje particije. Vidi se da je Π_1 „bolja" particija jer se na njoj postiže niža vrijednost funkcije cilja.

	c_1	c_2	c_3	$\mathcal{F}_1$
Π_1	$(4, 3)$	$(4, 8)$	$(8, 5.5)$	$(1+2+3)+(3+1+1)+(2.5+1.5+.5+1.5)=17$
Π_2	$(4.5, 2.5)$	$(4, 7)$	$(8, 5)$	$(2+2)+(4+2+0+2+3)+(1+1+1)=18$

Primjedba 3.3. Kao što smo u t. 3.3.3, str. 44, razmatrali problem grupiranja jednodimenzionalnih težinskih podataka, slično bismo mogli postupiti i u slučaju grupiranja težinskih dvodimenzionalnih i višedimenzionalnih podataka.

3.4.3 Formulacija problema grupiranja pomoću centara klastera

Slično kao u t. 3.2.3, za dani skup međusobno različitih točaka $z_1, \ldots, z_k \in \mathbb{R}^n$, primjenom principa minimalnih udaljenosti [31, 51, 53] možemo definirati particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ na sljedeći način

$$\pi_j = \{a \in \mathcal{A} : d(z_j, a) \leq d(z_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k, \quad (3.49)$$

pri čemu treba voditi računa o tome da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru. Za različite k-torke točaka $(z_1, \ldots, z_k)$ općenito dobivamo različite particije. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći optimizacijski problem

$$\min_{z_1,\ldots,z_k \in \mathbb{R}^n} F(z_1, \ldots, z_k), \qquad F(z_1, \ldots, z_k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} d(z_j, a^i), \quad (3.50)$$

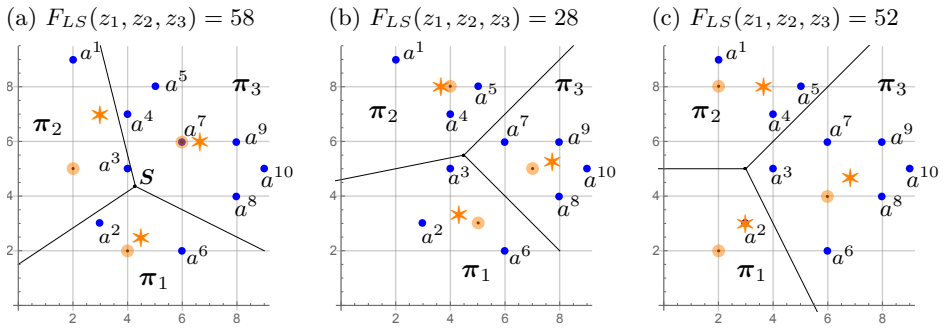
gdje je $F\colon \mathbb{R}^{k\times n} \to \mathbb{R}_+$. Funkcija F je realna funkcija $k \times n$ realnih varijabli: svaki centar ima po n komponenti. To znači da već u slučaju particije s $k = 2$ klastera podataka s $n = 2$ obilježja minimizirajuća funkcija F ima 4 nezavisne varijable. Općenito, funkcija F nije konveksna ni diferencijabilna,

a može imati puno lokalnih i globalnih minimuma [23, 72]. U principu, problem (3.50) mogao bi se rješavati poznatim metodama globalne optimizacije [25, 27, 62]), ali zbog izuzetno velikog broja nezavisnih varijabili i zbog izuzetne numeričke zahtjevnosti to se obično ne radi. Umjesto toga, u literaturi se mogu pronaći brojne specijalizacije (vidi primjerice [15, 43, 63]).

Primjer 3.23. *Promatramo skup podataka $\mathcal{A}$ iz Primjera 3.22 uz primjenu LS-kvazimetričke funkcije $d_{LS}(x, y) = \|x - y\|_2^2$.*

Izaberimo tri točke $z_1 = (2, 5)$, $z_2 = (4, 2)$, $z_1 = (6, 6)$ u ravnini (narančaste točke na Slici 3.11a) i za njih odredimo odgovarajuću vrijednost funkcije $F(z_1, z_2, z_3)$. Po definiciji funkcije F, to znači da za svaku točku $a \in \mathcal{A}$ treba pronaći najbližu od točaka z_1, z_2, z_3.

Umjesto toga, pokušajmo geometrijski odrediti „područje utjecaja" pojedine točke z_1, z_2, z_3, tj. područje svih točaka u ravnini koje su najbliže točki z_1, područje svih točaka u ravnini koje su najbliže točki z_2 i područje svih točaka u ravnini koje su najbliže točki z_3. Nakon toga, lako ćemo odrediti vrijednost funkcije $F(z_1, z_2, z_3)$.



Slika 3.11: Princip minimalnih udaljenosti za različite trojke točaka (narančaste točke)

Najprije pronađimo simetrale spojnica točaka $\{z_1, z_2\}$, $\{z_1, z_3\}$ i $\{z_2, z_3\}$ (vidi Sliku 3.11a). Iz geometrije znamo da se te simetrale sijeku u jednoj točki S (centar trokutu $\triangle z_1, z_2, z_3$ opisane kružnice!). Na taj način cijelu ravninu podijelili smo na tri područja polupravcima sa zajedničkim početkom u točki S. Ova tri polupravca čine tzv. Voronoijev dijagram [6, 53, 61].

Točke skupa $\mathcal{A}$ koje su najbliže točki z_1 jasno su naznačene Voronoijevim dijagramom na Slici 3.11a i čine prvi klaster $\pi_1 = \{(3, 3), (6, 2)\}$. Slično je

$$\pi_2 = \{(2, 9), (4, 5)\}, \quad \pi_3 = \{(4, 7), (5, 8), (6, 6), (8, 4), (8, 6), (9, 5)\}.$$

Vrijednost funkcije $F_{LS}(z_1, z_2, z_3)$ sada se lako računa

$$\begin{aligned} F_{LS}(z_1, z_2, z_3) &= (\|a^2 - z_1\|_2^2 + \|a^6 - z_1\|_2^2) + (\|a^1 - z_2\|_2^2 + \|a^3 - z_2\|_2^2) \\ &\quad + (\|a^4 - z_3\|_2^2 + \|a^5 - z_3\|_2^2 + \|a^7 - z_3\|_2^2 + \|a^8 - z_3\|_2^2 + \|a^9 - z_3\|_2^2 + \|a^{10} - z_3\|_2^2) \\ &= (2 + 4) + (16 + 4) + (0 + 5 + 5 + 8 + 4 + 10) = 58. \end{aligned}$$

Ako još odredimo i centroide klastera (narančaste zvjezdice na Slici 3.11a):

$$(4.5, 2.5), \quad c_2 = (3, 7), \quad c_3 = (6.67, 6),$$

možemo izračunati i vrijednost kriterijske funkcije $\mathcal{F}_{LS}$ na particiji $\Pi = \{\pi_1, \pi_2, \pi_3\}$: $\mathcal{F}_{LS}(\Pi) = 44.33$. Primijetite da je u ovom slučaju $\mathcal{F}_{LS}(\Pi) \neq F_{LS}(z_1, z_2, z_3)$.

	(z_1, z_2, z_3)	$\{\pi_1, \pi_2, \pi_3\}$	F_{LS}	$\mathcal{F}_{LS}$
(a)	$(4,2), (2,5), (6,6)$	$\{a^2, a^6\}$, $\{a^1, a^3\}$ $\{a^4, a^5, a^7, a^8, a^9, a^{10}\}$	58	44.33
(b)	$(5,3), (4,8), (7,5)$	$\{a^2, a^3, a^6\}$, $\{a^1, a^4, a^5\}$ $\{a^7, a^8, a^9, a^{10}\}$	28	23.5
(c)	$(2,2), (2,8), (6,4)$	$\{a^2\}$, $\{a^1, a^4, a^5\}$ $\{a^6, a^7, a^8, a^9, a^{10}\}$	52	34.83

Tablica 3.12: Nekoliko particija skupa $\mathcal{A}$ iz Primjera 3.22

Na sličan su način za nekoliko različitih trojki točaka (z_1, z_2, z_3) u Tablici 3.12 navedeni odgovarajući klasteri π_1, π_2, π_3, vrijednosti kriterijske funkcije F_{LS} i kriterijske funkcije $\mathcal{F}_{LS}$. Primijetite da što su centroidi klastera bliži točkama (z_1, z_2, z_3), vrijednosti kriterijskih funkcija F_{LS} i $\mathcal{F}_{LS}$ također su međusobno bliže.

Zadatak 3.15. Slično kao u Zadatku 3.7, str. 39, i analogno prethodnom primjeru, za skup $\mathcal{A}$ iz Primjera 3.22 izračunajte vrijednosti funkcije F_1 u nekoliko izabranih trojki točaka tako da kao u Primjeru 3.13, str. 37, za svaku trojku točaka najprije odredite pripadne particije s klasterima π_1, π_2, π_3 a nakon toga primjenom ℓ_1-metričke funkcije izračunate odgovarajuće vrijednosti funkcije F_1 zadane s (3.48). Napravite odgovarajuću analizu rezultata. Može li se u ovom slučaju koristiti Voronoijev dijagram?

Poglavlje 4

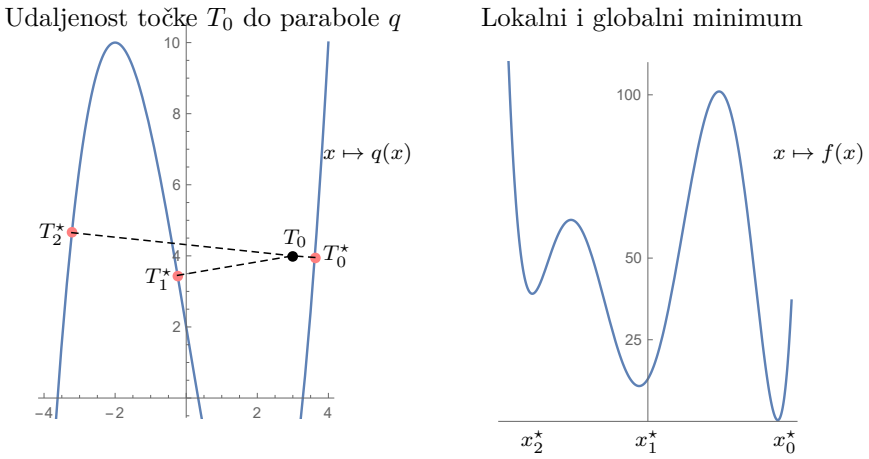
Traženje optimalne particije

Particiju $\Pi^\star \in \mathcal{P}(\mathcal{A}; k)$ na kojoj funkcija cilja postiže svoju najnižu vrijednost zovemo *globalno optimalna particija* (GOP). Problem traženja GOP složeni je optimizacijski problem nediferencijabilne i nekonveksne optimizacije. Zbog velikog broja nezavisnih varijabli isključena je direktna primjena klasičnih metoda minimizacije. U literaturi se mogu pronaći razne heurističke metode [1, 23, 43, 46, 54, 63] koje nemaju ambiciju pronalaženja GOP, ali mogu dati *lokalno optimalnu particiju* (LOP) koja se može činiti dosta blisa GOP[1]. Zbog toga ćemo malo više pažnje posvetiti najpoznatijoj metodi za traženje LOP: k-means algoritmu. To je iterativni proces koji na osnovi početne aproksimacije (početnih centara ili početne particije) daje LOP. Algoritam u konačno koraka daje rješenje i u svakom koraku snižava vrijednost funkcije cilja. Takvo rješenje ne mora biti globalno optimalno, što znači da se može dogoditi da ima boljih rješenja od dobivenog. Još jedan nedostatak ove metode je mogućnost da se tijekom iterativnog procesa može dogoditi da neki od klastera postane prazan skup, tj. može se dogoditi smanjenje broja klastera. Tek u t. 4.5, str. 87, pokazat ćemo jednu mogućnost za pronalaženje optimalne particije koja može biti prihvatljivo blizu GOP.

Pojam „globalno optimalno" i „lokalno optimalno" ilustrirat ćemo na sljedećem jednostavnom primjeru.

Primjer 4.1. *Treba odrediti točku na kubnoj paraboli zadanoj s funkcijom* $q\colon \mathbb{R} \to \mathbb{R}$, $q(x) = .5x^3 - 6x + 2$, *a koja je najbliža točki* $T_0 = (3, 4)$ *u smislu obične Euklidske* ℓ_2 *metrike.*

[1] Bez ambicije davanja matematički korektne definicije pojmova *globalno optimalne particije* (GOP) i *lokalno optimalne particije* (LOP) može se reći da je GOP najbolja particija u smislu (3.42), str. 46, a LOP je najbolja u nekom užem području.



Slika 4.1: Traženje točke na kubnoj paraboli koja je najbliža danoj točki T_0

Primijetite da je ℓ_2 udaljenost točke $T_0 = (x_0, y_0)$ do neke točke $T = (x, q(x))$ na grafu funkcije q zadana s

$$d_2(T_0, T) = \sqrt{(x - x_0)^2 + (q(x) - y_0)^2}.$$

Određivanje točke $T^\star = (x^\star, q(x^\star))$ na kojoj se postiže minimalna udaljenost točke T_0 do grafa funkcije q svodi se na optimizacijski problem koji ima više od jednog lokalnog minimuma. Kako je δ_2 neprekidno derivabilna funkcija na $\mathbb{R}$, a $x \mapsto \sqrt{x}$ monotono rastuća funkcija, naš se problem može svesti na rješavanje jednog nelinearnog problema najmanjih kvadrata (Least Squares Problem), odnosno na problem minimizacije polinoma 6-tog stupnja (vidi Sliku 4.1b)

$$f(x) = (x - x_0)^2 + (q(x) - y_0)^2 = .25x^6 - 6x^4 - 2x^3 + 37x^2 + 18x + 13.$$

Graf funkcije f ima tri lokalna minimuma koji se postižu u točkama $x_0^\star \approx 3.62$, $x_1^\star \approx -0.24$, $x_2^\star \approx -3.22$ (vidi Sliku 4.1b). Na Slici 4.1a označene su točke $T_0^\star$, $T_1^\star$ i $T_2^\star$ s apscisama $x_0^\star$, $x_1^\star$ i $x_2^\star$.

Globalni minimum funkcije f postiže se u točki $x_0^\star$ gdje funkcija f prima najnižu vrijednost. Kao što se može vidjeti na Slici 4.1a točka $T_0^\star$ je točka koja leži na kubnoj paraboli q, a najbliža je točki T_0.

4.1 Motivacija: Traženje lokalno optimalne 2-particije skupa podataka s jednim obilježjem

Spomenuti k-means algoritam za traženje LOP najprije ćemo pokazati u najjednostavnijem slučaju za traženje lokalno optimalne 2-particije[2] skupa podataka s jednim obilježjem ($n = 1$).

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ skup (konačni niz) podataka, a $\mathcal{P}(\mathcal{A}; 2)$ skup svih njegovih 2-particija (vidi Definiciju 3.1, str. 23). Prema (3.1) skup $\mathcal{P}(\mathcal{A}; 2)$ ima $2^{m-1} - 1$ različitih particija, što za veliki m može biti izuzetno veliki broj (vidi Tablicu 3.1). Iako u ovom najjednostavnijem slučaju ($n = 1$) optimalne particije tražimo između particija čiji se klasteri međusobno nastavljaju, a kojih prema (3.16) ima samo $\binom{m-1}{2-1} = m - 1$, i to može biti vrlo zahtjevan postupak.

Neka je $\Pi = \{\pi_1, \pi_2\} \in \mathcal{P}(\mathcal{A}; 2)$ neka 2-particija skupa $\mathcal{A}$. Ako uvedemo neku kvazimetričku funkciju $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ (vidi t. 3, str. 23), onda svakom klasteru možemo pridružiti njegov centar

$$c_1 \in \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_1} d(x, a), \qquad c_2 \in \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_2} d(x, a),$$

i prema (3.2) uvesti kriterijsku funkciju cilja

$$\mathcal{F}(\Pi) = \sum_{a \in \pi_1} d(c_1, a) + \sum_{a \in \pi_2} d(c_2, a). \tag{4.1}$$

Funkcija $\mathcal{F}$ pokazuje ukupno „rasipanje" elemenata klastera π_1 do centra c_1 i elemenata klastera π_2 do centra c_2. Što je vrijednost funkcije $\mathcal{F}$ manja, time je „rasipanje" manje, što znači da su klasteri kompaktniji i međusobno bolje razdvojeni. Problem traženja globalno optimalne 2-particije možemo zapisati na sljedeći način

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};2)} \mathcal{F}(\Pi).$$

Zbog općenito velikog broja particija u $\mathcal{P}(\mathcal{A}; 2)$ traženje GOP putem pretraživanja cijelog skupa $\mathcal{P}(\mathcal{A}; 2)$ nije moguće u prihvatljivom vremenu. To je razlog što se u primjenama često zadovoljavamo pronalaženjem neke LOP. Najpoznatiji algoritam za traženje LOP je k-means algoritam, koji se može opisati sa sljedeća dva koraka [31, 34, 39, 51, 53, 60].

[2] Zbog pojednostavljivanja rečeničnih konstrukcija i jasnoće zapisa, umjesto „particija s 2 klastera" ili „dvočlana particija" jednostavno ćemo pisati 2-particija.

Algoritam 4.1. *[k-means algoritam]*

Korak A: Pridruživanje (assignment step). Poznavanjem različitih brojeva $z_1 \neq z_2$, skup $\mathcal{A}$ treba grupirati u dva disjunktna klastera π_1, π_2 korištenjem principa minimalnih udaljenosti

$$\pi_1 = \{a \in \mathcal{A} \colon d(z_1, a) \leq d(z_2, a)\},$$
$$\pi_2 = \{a \in \mathcal{A} \colon d(z_1, a) > d(z_2, a)\}.$$

Korak B: Korekcija (update step). Za poznatu particiju $\Pi = \{\pi_1, \pi_2\}$ skupa $\mathcal{A}$ treba definirati centre klastera

$$c_1 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_1} d(x, a), \qquad c_2 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_2} d(x, a).$$

(Brojevi c_1, c_2 iz Koraka B nakon jedne iteracije postaju brojevi z_1, z_2 u sljedećoj iteraciji.)

U Koraku A, principom minimalnih udaljenosti cijeli skup $\mathcal{A}$ razdjeljuje se u dvije skupine: na elemente koji su bliži broju z_1 i na elemente koji su bliži broju z_2. Ako je neki element skupa $\mathcal{A}$ jednako udaljen od brojeva z_1 i z_2, pridružit će se prvom (lijevom) klasteru π_1. Geometrijski, klasteri se mogu odrediti tako da odredimo polovište točaka koje reprezentiraju brojeve z_1 i z_2. Tada svi elementi lijevo od polovišta pripadaju klasteru π_1, a svi elementi desno od polovišta pripadaju klasteru π_2. Ako se neki element pojavi baš na polovištu, svrstat ćemo ga u lijevi klaster π_1.

U Koraku B, svakom klasteru particije Π pridružujemo njima odgovarajuće centre. Ako je d, LS-kvazimetrička funkcija, centri su aritmetičke sredine klastera, a ako je d, ℓ_1-metrička funkcija, centri su medijani klastera. Poznavanjem centara možemo odrediti i vrijednost kriterijske funkcije cilja (4.1).

k-means algoritam sukcesivno ponavlja navedene korake, a može se pokrenuti zadavanjem početne particije $\Pi = \{\pi_1, \pi_2\}$ ili zadavanjem početnih centara (međusobno različitih brojeva z_1, z_2).

Primjedba 4.1. U ovom udžbeniku koristit ćemo *k-means algoritam uz primjenu LS-kvazimetričke funkcije* i *k-means algoritam uz primjenu ℓ_1-metričke funkcije*, iako se u znanstvenoj literaturi može pronaći i primjena brojnih drugih kvazimetričkih funkcija (vidi primjerice [31, 53, 54]). Spomenimo još da se k-means algoritam uz primjenu ℓ_1-metričke funkcije u literaturi pojavljuje i pod nazivom k-medijan algoritam.

Primjedba 4.2. Primijetite da se u slučaju podataka s jednim obilježjem ($\mathcal{A} \subset \mathbb{R}$) spomenuti princip minimalnih udaljenosti ne razlikuje u slučaju primjene LS-kvazimetričke funkcije ili u slučaju primjene ℓ_1-metričke funkcije.

Primjedba 4.3. Prilikom korištenja k-means algoritma uz primjenu ℓ_1-metričke funkcije u cilju traženja ℓ_1-optimalne 2-particije u Koraku B Algoritma 4.1 može se dogoditi da neki centar c_j može primiti proizvoljnu vrijednost iz nekog intervala $[\alpha, \beta] \subset \mathbb{R}$ (svojstvo medijana!). U tom slučaju, najbolje je uzeti polovište intervala $c_j = \frac{\alpha+\beta}{2}$.

4.1.1 Inicijalizacija k-means algoritma početnom particijom

Ako algoritam pokrenemo zadavanjem početne particije, najprije će se izvoditi Korak B. Nakon određivanja centara c_1 i c_2 klastera π_1 i π_2, u mogućnosti smo odrediti vrijednost funkcije cilja $\mathcal{F}$ (također možemo odrediti i prosječno „rasipanje" podataka po klasterima). Nakon toga pokreće se Korak A za brojeve c_1 i c_2, na osnovi kojih principom minimalnih udaljenosti određujemo novu particiju s novim klasterima. Postupak se dalje ponavlja toliko dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera tada također postanu jednaki).

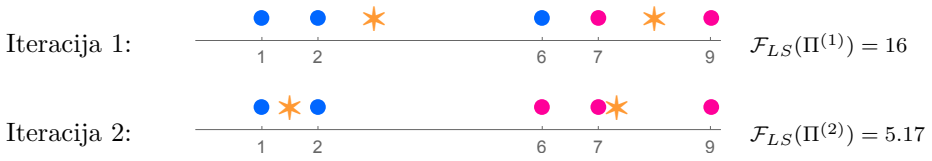
U svakom koraku k-means algoritma vrijednost funkcije cilja $\mathcal{F}$ snižava se i asimptotski približava lokalno najmanjoj mogućoj vrijednosti [31, 68].

Kako bi k-means algoritam dao particiju što bližu globalno optimalnoj, početna particija s kojom započinjemo iterativni postupak mora biti što bolje izabrana.

Primjer 4.2. *Treba pronaći LS-optimalnu* 2*-particiju skupa* $\mathcal{A} = \{1, 2, 6, 7, 9\}$ *primjenom* k*-means Algoritma 4.1 uz početnu particiju* $\Pi^{(0)} = \{\{1, 2, 6\}, \{7, 9\}\}$. *Postupak koristi LS-kvazimetričku funkciju, a vidljiv je u Tablici 4.1 i na Slici 4.2.*

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
1	{1,2,6}	{7,9}	3	8	16	30
2	{1,2}	{6,7,9}	3/2	22/3	$31/6 \approx 5.167$	40.83
3	{1,2}	{6,7,9}	3/2	22/3	$31/6 \approx 5.167$	40.83

Tablica 4.1: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$



Slika 4.2: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$

Nakon što smo prema Korak B odredili centre klastera početne particije ($c_1 = 3, c_2 = 8$), prelazimo na Korak A i određujemo njihovo polovište 5.5. Lijevo od polovišta su elementi $1, 2$, a desno elementi $6, 7, 9$. Tako smo odredili klastere nove particije i u mogućnosti smo ponovo pokrenuti Korak B, itd. Na kraju dobivamo LOP, $\Pi^\star = \{\{1, 2\}, \{6, 7, 9\}\}$. Je li ona i globalno optimalna? Odgovor na ovo pitanje možemo dobiti tako da dobivenu particiju usporedimo sa svim particijama čiji klasteri se nastavljaju jedan na drugi. Prema (3.16), str. 35, takvih particija ima $\binom{5-1}{2-1} = 4$. Provjerite je li dobivena particija LS-optimalna.

Zadatak 4.1. Odredite ℓ_1-optimalnu 2-particiju skupa $\mathcal{A}$ iz Primjera 4.2.

Rješenje: $\Pi^\star = \{\{1, 2\}, \{6, 7, 9\}\}$; $c_1 = 3/2$, $c_2 = 7$; $\mathcal{F}_1(\Pi^\star) = 4$.

Sljedeći primjer pokazuje da standardni k-means algoritam daje lokalno optimalno rješenje, koje nije ujedno i globalno optimalno rješenje.

Primjer 4.3. *Treba pronaći LS-optimalnu* 2*-particiju skupa* $\mathcal{A} = \{0, 2, 3\}$*, primjenom* k*-means algoritma uz početnu particiju* $\Pi^{(0)} = \{\{0, 2\}, \{3\}\}$.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$
1	{0,2}	{3}	1	3	2
2	{0,2}	{3}	1	3	2

Tablica 4.2: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{0, 2, 3\}$

Kao što se vidi iz Tablici 4.2, k-means algoritam uz primjenu LS-kvazimetričke funkcije ne može pronaći bolju particiju od početne. Međutim, bolja particija u ovom je slučaju particija $\Pi^\star = \{\{0\}, \{2, 3\}\}$ jer je $\mathcal{F}(\Pi^\star) = 0.5$. Na ovom jednostavnom primjeru pokazano je da k-means algoritam daje LOP. Izborom neke druge početne particije možda bismo dobili GOP. Pokušajte!

4.1.2 Inicijalizacija k-means algoritma početnim centrima

k-means algoritam također se može pokrenuti i izborom početnih centara z_1 i z_2. U tom slučaju najprije treba primijeniti princip minimalnih udaljenosti iz Koraka A i tako odrediti početnu particiju s odgovarajućim klasterima π_1 i π_2. Nakon toga pokrećemo Koraka B, određujemo centre c_1 i c_2 klastera π_1 i π_2 i izračunavamo vrijednost funkcije cilja $\mathcal{F}$. Nakon toga, pokreće se Korak A s brojevima c_1 i c_2, na osnovi kojih principom minimalnih udaljenosti određujemo novu particiju s novim klasterima. Postupak se dalje ponavlja tako dugo dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera također postanu jednaki).

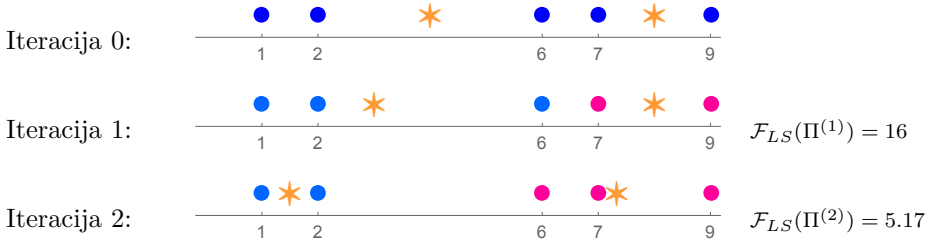
Kako bi u ovom slučaju k-means algoritam dao particiju što bliže globalno optimalnoj, početni centri z_1 i z_2 trebaju biti što bolje izabrani: trebaju biti što bliži centrima klastera unaprijed nepoznate GOP. U literaturi se mogu pronaći metode za procjenu dobrih početnih centara (vidi primjerice [69, 81]). Nažalost, to obično nisu jednostavni postupci.

Primjer 4.4. *Treba pronaći LS-optimalnu 2-particiju skupa* $\mathcal{A} = \{1, 2, 6, 7, 9\}$ *primjenom k-means algoritma uz početne centre* $z_1 = 4$, $z_2 = 8$.

Uz primjenu LS-kvazimetričke funkcije korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 4$, $z_2 = 8$ određujemo početnu particiju $\Pi^{(0)} = \{\{1, 2, 6\}, \{7, 9\}\}$. Daljnji tijek iterativnog procesa prikazan je u Tablici 4.3 i na Slici 4.3.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$
1	{1,2,6}	{7,9}	3	8	16
2	{1,2}	{6,7,9}	1.5	7.33	5.17
3	{1,2}	{6,7,9}	1.5	7.33	5.17

Tablica 4.3: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$



Slika 4.3: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$

Je li dobivena 2-particija $\Pi^\star = \{\{1,2\},\{6,7,9\}\}$ globalno LS-optimalna? Provjerite to tako da particiju $\Pi^\star$ usporedite sa svim drugim 2-particijama čiji se klasteri međusobno nastavljaju. Koliko takvih particija ima u ovom primjeru?

Zadatak 4.2. Pronađite ℓ_1-optimalnu 2-particiju skupa $\mathcal{A}$ iz Primjera 4.4 uz iste početne centre.

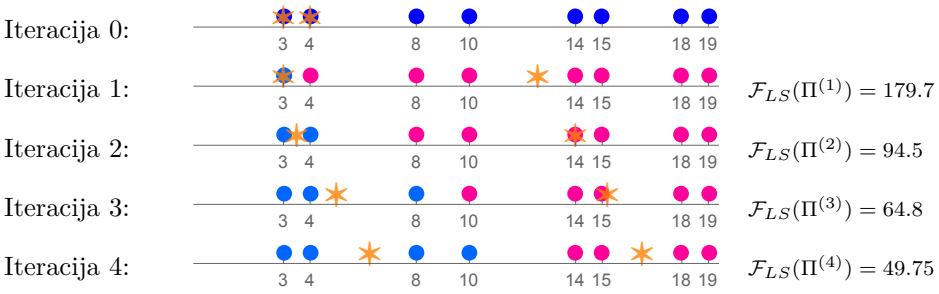
Rješenje: $\Pi^\star = \{\{1,2\},\{6,7,9\}\}$; $c_1 = 3/2$, $c_2 = 7$; $\mathcal{F}_1(\Pi^\star) = 4$.

Primjer 4.5. *Primjenom k-means algoritma treba pronaći LS-optimalnu 2-particiju skupa* $\mathcal{A} = \{3,4,8,10,14,15,18,19\}$ *uz početne centre* $z_1 = 3$, $z_2 = 4$.

Uz primjenu LS-kvazimetričke funkcije, korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 3$, $z_2 = 4$ dobivamo početnu particiju $\Pi^{(0)} = \{\{3\},\{4,8,10,14,15,18,19\}\}$. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.4 ili na Slici 4.4.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$
1	{3}	{4, 8, 10, 14, 15, 18, 19}	3	12.57	179.7
2	{3, 4}	{8,10, 14, 15, 18, 19}	3.5	14	94.5
3	{3, 4, 8}	{10,14, 15, 18, 19}	5	15.2	64.8
4	{3, 4, 8, 10}	{14, 15, 18, 19}	6.25	16.5	49.75
5	{3, 4, 8, 10}	{14, 15, 18, 19}	6.25	16.5	49.75

Tablica 4.4: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{3,4,8,10,14,15,18,19\}$



Slika 4.4: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{3,4,8,10,14,15,18,19\}$

Zadatak 4.3. Budući da optimalnu particiju skupa s jednim obilježjem ima smisla tražiti samo između particija čiji se klasteri međusobno nastavljaju (vidi Primjedbu 3.1, str. 35), i da skup iz prethodnog primjera ima samo 7

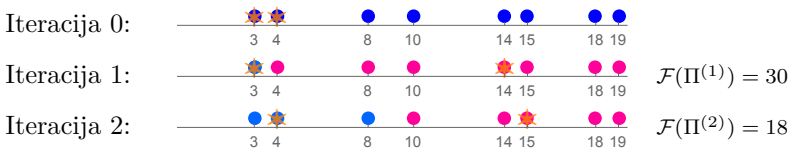
(formula (3.16)) takvih particija, provjerite je li k-means algoritam pronašao GOP.

Primjer 4.6. *Za skup $\mathcal{A}$ iz prethodnog primjera potražit ćemo ℓ_1-optimalnu 2-particiju korištenjem k-means algoritma uz primjenu ℓ_1-metričke funkcije s istim početnim centrima $z_1 = 3$, $z_2 = 4$.*

Primjenom principa minimalnih udaljenosti za početne centre $z_1 = 3$ i $z_2 = 4$ dobivamo istu početnu particiju $\Pi^{(0)} = \{\{3\}, \{4, 8, 10, 14, 15, 18, 19\}\}$ kao u prethodnom primjeru. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.5 i na Slici 4.5.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$
1	{3}	{4, 8, 10, 14, 15, 18, 19}	3	14	30
2	{3, 4, 8}	{10, 14, 15, 18, 19}	4	15	18
3	{3, 4, 8}	{10, 14, 15, 18, 19}	4	15	18

Tablica 4.5: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$



Slika 4.5: k-means algoritam uz primjenu ℓ_1-metričke funkcije

Postupak je bio kraći, a dobivena ℓ_1-optimalna 2-particija (Tablica 4.5) razlikuje se od odgovarajuće LS-optimalne 2-particije. Ako isti postupak provedemo s nešto drugačijim početnim centrima $z_1 = 5$ i $z_2 = 16$, dobivamo početnu particiju $\Pi^{(0)} = \{\{3, 4, 8, 10\}, \{14, 15, 18, 19\}\}$, koja je ujedno i lokalno optimalna u ovom slučaju. Na ovom jednostavnom primjeru vidjeli smo kako k-means algoritam uz izbor različitih početnih centara pronalazi različite LOP.

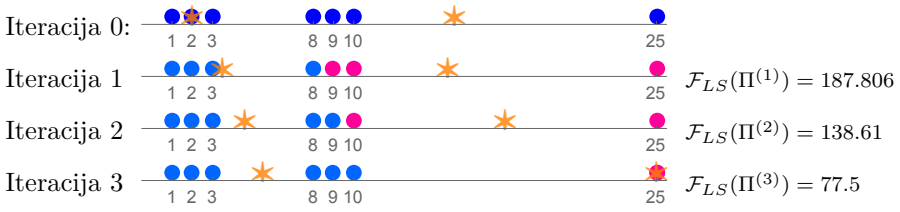
Zadatak 4.4. Među sedam 2-particija skupa $\mathcal{A}$ iz Primjera 4.6, odnosno Primjera 4.5, čiji se klasteri međusobno nastavljaju (vidi Primjedbu 3.1, str. 35) pronađite globalno ℓ_1-optimalnu 2-particiju. Što možete reći o optimalnosti particija:
$\Pi^{(1)} = \{\{3, 4, 8\}, \{10, 14, 15, 18, 19\}\}$ i $\Pi^{(2)} = \{\{3, 4, 8, 10\}, \{14, 15, 18, 19\}\}$
u odnosu na globalno ℓ_1-optimalnu?

Primjer 4.7. *Primjenom k-means algoritma treba pronaći LS-optimalnu 2-particiju skupa* $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$ *uz početne centre* $z_1 = 2$, $z_2 = 15$ *(vidi Primjer 5 u [53]).*

Najprije primjenom LS-kvazimetričke funkcije korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 2$, $z_2 = 15$ odredimo početnu particiju $\Pi^{(0)} = \{\{1, 2, 3, 8\}, \{9, 10, 25\}\}$. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.6 ili na Slici 4.6.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$
1	{1, 2, 3, 8}	{9,10, 25}	3.5	14.667	189.667
2	{1, 2, 3, 8,9}	{10, 25}	4.6	17.5	138.61
3	{1, 2, 3, 8, 9,10}	{25}	5.5	25	77.5
4	{1, 2, 3, 8, 9,10}	{25}	5.5	25	77.5

Tablica 4.6: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$



Slika 4.6: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$

Broj svih dvočlanih particija ovog skupa je $2^{7-1} - 1 = 63$, ali postoji samo 6 particija čiji se klasteri međusobno nastavljaju. Odmah uočavamo da skup $\mathcal{A}$ sadrži dvije značajno različite skupine realnih brojeva $\mathcal{A}_1 = \{1, 2, 3\}$ te $\mathcal{A}_2 = \{8, 9, 10\}$. Također, skup $\mathcal{A}$ sadrži i element 25, kojeg možemo shvatiti kao jako stršeći podatak („outlier") nastao zbog određene pogreške, a prirodno dolazi iz skupine $\mathcal{A}_2$. Primjenom k-means algoritma uz početne centre $z_1 = 2$ i $z_2 = 15$ dobivamo početnu particiju $\Pi^{(0)} = \{\pi_1^{(0)}, \pi_2^{(0)}\}$, $\pi_1^{(0)} = \{1, 2, 3, 8\}$, $\pi_2^{(0)} = \{9, 10, 25\}$. Direktnom provjerom svih particija može se pokazati da je k-means algoritam pronašao upravo globalno optimalnu particiju. Iz ovog primjera vidljivo je da k-means algoritam uz primjenu LS-kvazimetričke funkcije daje particiju koja značajno ovisi o stršećem podatku, tako da upravo stršeći podatak čini zaseban klaster (vidi Sliku 4.6).

Zadatak 4.5. Dopunite Tablicu 4.6 odgovarajućim vrijednostima dualne funkcije $\mathcal{G}$.

Zadatak 4.6. Za skup $\mathcal{A}$ iz Primjera 4.7 primjenom k-means algoritma odredite ℓ_1-optimalnu 2-particiju uz iste početne centre $z_1 = 2$ i $z_2 = 15$ (vidi Primjer 6 u [53]). Usporedite i komentirajte dobivenu ℓ_1-optimalnu 2-particiju s odgovarajućom LS-optimalnom 2-particijom dobivenom u Primjeru 4.7.

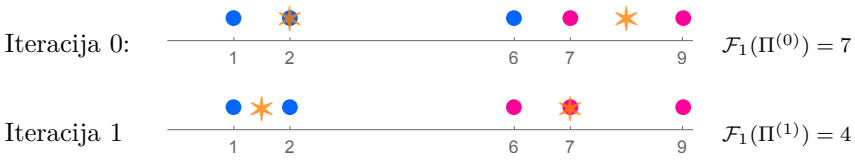
Rješenje: $\Pi^\star = \{\{1,2,3\},\{8,9,10,25\}\}$; $c_1 = 2$, $c_2 = 9.5$; $\mathcal{F}_1(\Pi^\star) = 20$.

Primjer 4.8. *Treba pronaći ℓ_1-optimalnu 2-particiju skupa $\mathcal{A} = \{1,2,6,7,9\}$ primjenom k-means algoritma uz početnu particiju $\Pi^{(0)} = \{\{1,2,6\},\{7,9\}\}$.*

Cijeli iterativni postupak vidljiv je u Tablici 4.7 i na Slici 4.7.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$
0	{1,2,6}	{7,9}	2	8	5+2=7
1	{1,2}	{6,7,9}	1.5	7	1+3=4
2	{1,2}	{6,7,9}	1.5	7	1+3=4

Tablica 4.7: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{1,2,6,7,9\}$



Slika 4.7: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{1,2,6,7,9\}$

Zadatak 4.7. Za skup $\mathcal{A}$ iz pethodnog primjera primjenom k-means algoritma odredite LS-optimalnu 2-particiju. Razlikuje li se dobivena LS-optimalna 2-particija od ℓ_1-optimalne 2-particije dobivene u Primjeru 4.8.

Rješenje: Ne!

4.2 Traženje lokalno optimalne k-particije podataka s jednim obilježjem

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ skup (konačni niz) podataka, a $\mathcal{P}(\mathcal{A}; k)$ skup svih njegovih k-particija (vidi Definiciju 3.1, str. 23). Broj elemenata skupa $\mathcal{P}(\mathcal{A}; k)$ određuje se prema (3.1) (vidi također Tablicu 3.1).

Neka je $\Pi = \{\pi_1, \ldots, \pi_k\} \in \mathcal{P}(\mathcal{A}; k)$ neka k-particija. Ako uvedemo neku kvazimetričku funkciju $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ (vidi t. 3, str. 23), onda svakom klasteru možemo pridružiti njegov centar

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k,$$

a prema (3.2) uvodimo kriterijsku funkciju cilja

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a). \tag{4.2}$$

Funkcija $\mathcal{F}$ pokazuje ukupno „rasipanje" elemenata svih klastera do njihovih centara. Što je vrijednost funkcije $\mathcal{F}$ manja, time je „rasipanje" manje, što znači da su klasteri kompaktniji i međusobno bolje razdvojeni. Problem traženja optimalne k-particije zapisuje se na sljedeći način

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi).$$

Već smo naglasili da traženje optimalne particije putem pretraživanja cijelog skupa $\mathcal{P}(\mathcal{A}; k)$ nije moguće u prihvatljivom vremenu. Pretraživanje skupa svih particija čiji se klasteri međusobno nastavljaju također može biti vremenski vrlo zahtjevno.

Za traženje lokalno optimalne k-particije koristit ćemo dobro poznati k-means algoritam. Ovaj algoritam nema ambiciju pronalaženja generalno najbolje GOP, o čemu ćemo nešto više reći u t. 4.5, str. 87. k-means algoritam može se opisati sa sljedeća dva koraka [31, 34, 39, 51, 53, 60].

Algoritam 4.2. *[k-means algoritam]*

Korak A: Priduživanje (assignment step). Poznavanjem međusobno različitih brojeva $z_1, \ldots, z_k$, skup $\mathcal{A}$ treba grupirati u k disjunktnih

klastera $\pi_1, \ldots, \pi_k$ *korištenjem principa minimalnih udaljenosti*

$$\pi_j = \{a \in \mathcal{A} \colon d(z_j, a) \leq d(z_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k.$$

Korak B: Korekcija (update step). Za poznatu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ *skupa* $\mathcal{A}$*, treba definirati centre klastera*

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k.$$

(Brojevi $c_1, \ldots, c_k$ *iz Koraka B nakon jedne iteracije postaju brojevi* $z_1, \ldots, z_k$ *u sljedećoj iteraciji.)*

U Koraku A principom minimalnih udaljenosti cijeli skup $\mathcal{A}$ razdjeljuje se u k skupina prema njihovoj d-bliskosti pojedinim točkama z_j. Geometrijski, klasteri se mogu odrediti tako da odredimo polovište susjednih točaka. Ta polovišta razdjeljuju nove klastere. Ako se neki element pojavi baš na nekom polovištu, svrstat ćemo ga u lijevi klaster.

U Koraku B svakom klasteru particije Π pridružujemo njegove centre. Ako je d, LS-kvazimetrička funkcija, ti centri su aritmetičke sredine elemenata klastera, a ako je d, ℓ_1-metrička funkcija, ti centri su medijani elemenata klastera. Poznavanjem centara možemo odrediti i vrijednost kriterijske funkcije cilja (4.2).

k-means algoritam sukcesivno ponavlja navedene korake toliko dugo dok se particije ne počnu ponavljati. U tom se slučaju centri klastera trenutne i prethodne particije se podudaraju, a vrijednost funkcije cilja prestane opadati. Algoritam se može pokrenuti zadavanjem početne particije $\Pi^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_k^{(0)}\}$ ili zadavanjem početnih centara (međusobno različitih brojeva $z_1^{(0)}, \ldots, z_k^{(0)}$).

Primjedba 4.4. Kao što smo već spomenuli u Primjedbi 4.1, str. 60, u ovom udžbeniku koristit ćemo *k-means algoritam uz primjenu LS-kvazimetričke funkcije* i *k-means algoritam uz primjenu* ℓ_1*-metričke funkcije*, iako se u znanstvenoj literaturi može pronaći i primjena brojnih drugih kvazimetričkih funkcija (vidi primjerice [31, 53, 54]). Spomenimo još da se k-means algoritam uz primjenu ℓ_1-metričke funkcije u literaturi pojavljuje i pod nazivom *k-medijan algoritam*.

Primjedba 4.5. Kao što smo već spomenuli u Primjedbi 4.2, str. 61, u slučaju podataka s jednim obilježjem ($\mathcal{A} \subset \mathbb{R}$), spomenuti princip minimalnih udaljenosti ne razlikuje se u slučaju primjene LS-kvazimetričke funkcije ili u slučaju primjene ℓ_1-metričke funkcije.

4.2.1 Inicijalizacija k-means algoritma početnom particijom

Ako algoritam pokrenemo zadavanjem početne particije, najprije će se izvoditi Korak B. Nakon određivanja centara $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$, u mogućnosti smo odrediti vrijednost funkcije cilja $\mathcal{F}$ (također možemo odrediti i prosječno „rasipanje" podataka po klasterima). Nakon toga pokreće se Korak A za brojeve $c_1, \ldots, c_k$, na osnovi kojih principom minimalnih udaljenosti određujemo novu particiju s novim klasterima. Postupak se dalje ponavlja toliko dugo dok trenutnu i prethodna particija ne postanu jednake (centri njihovih klastera također postanu jednaki).

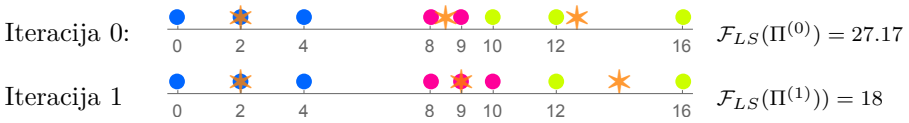
Vrijednost funkcije cilja $\mathcal{F}$ snižava se u svakom koraku k-means algoritma i asimptotski približava lokalno najmanjoj mogućoj vrijednosti [31, 68].

Kako bi k-means algoritam dao particiju što bližu GOP, particija s kojom započinjemo iterativni postupak mora biti što bolje izabrana. Tu particiju nazivamo početna k-particija.

Primjer 4.9. *Primjenom k-means algoritma (Algoritam 4.2) treba pronaći LS-optimalnu 3-particiju skupa $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$ krenuvši od početne particije $\Pi^{(0)} = \{\{0, 2, 4\}, \{8, 9\}, \{10, 12, 16\}\}$.*

Iteracija	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$
0	{0,2,4}	{8,9}	{10, 12, 16}	2	8.5	12.67	27.17
1	{0,2,4}	{8,9,10}	{12, 16}	2	9	14	18
2	{0,2,4}	{8,9,10}	{12, 16}	2	9	14	18

Tablica 4.8: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$



Slika 4.8: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$

LS-optimalnu 3-particiju skupa $\mathcal{A}$ potražit ćemo korištenjem k-means algoritma uz primjenu LS-kvazimetričke funkcije. Nakon što smo prema Koraku B odredili centre klastera $c_1 = 2$, $c_2 = 8.5$ i $c_3 = 12.67$ naše početne particije (iteracija 0 na Slici 4.8), prelazimo na Korak A. Najprije odredimo polovišta susjednih centara. Lijevo od polovišta centara c_1 i c_2 nalaze se elementi $0, 2, 4$ prvog početnog klastera π_1, između polovišta centara c_1 i c_2 i polovišta centara c_2 i c_3 nalaze se elementi $8, 9$ drugog početnog klastera π_2 i element 10 trećeg početnog klastera, dok se desno od polovišta centara c_2, c_3 nalaze se elementi $12, 16$ trećeg početnog klastera π_3. Tako smo odredili klastere

nove particije (iteracija 1 na Slici 4.8) i u mogućnosti smo ponovo pokrenuti Korak B itd. Na kraju dobivamo LOP, $\Pi^\star = \{\{0,2,4\},\{8,9,10\},\{12,16\}\}$. Je li ona i GOP? Odgovor na ovo pitanje možemo dobiti tako da ovu particiju usporedimo sa svim particijama čiji se klasteri međusobno nastavljaju jedan na drugi. Prema (3.16), str. 35, takvih particija ima $\binom{8-1}{3-1} = 21$. Provjerite je li dobivena particija LS-optimalna.

Zadatak 4.8. Odredite ℓ_1-optimalnu 3-particiju skupa $\mathcal{A}$ iz prethodnog primjera uz istu početnu particiju $\Pi^{(0)} = \{\{0,2,4\},\{8,9\},\{10,12,16\}\}$. Razlikuje li se dobivena ℓ_1-optimalna 3-particija od ranije dobivene LS-optimalne 3-particije istog skupa?

Rješenje: Ne razlikuje!

Primjer 4.10. *Zadan je skup* $\mathcal{A} = \{1,2,4,8,9,10,12,16,18,20\}$. *Primjenom k-means algoritma treba pronaći LS-optimalnu* 3*-particiju skupa* $\mathcal{A}$ *polazeći od početne particije* $\Pi^{(0)} = \{\{1,2,4\},\{8,9\},\{10,12,16,18,20\}\}$.

Tijek iterativnog postupka k-means algoritma uz primjenu LS-kvazimetričke funkcije može se vidjeti u Tablici 4.9 i na Slici 4.9.

Iter.	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$
0	{1,2,4}	{8,9}	{10, 12, 16, 18, 20}	2.33	8.5	15.2	73.97
1	{1,2,4}	{8,9,10}	{12, 16, 18, 20}	2.33	9	16.5	41.67
2	{1,2,4}	{8,9,10,12}	{16, 18, 20}	2.33	9.75	18	21.46
3	{1,2,4}	{8,9,10,12}	{16, 18, 20}	2.33	9.75	18	21.46

Tablica 4.9: Traženje LS-optimalne 3-particije skupa $\{1,2,4,8,9,10,12,15,18,20\}$

Slika 4.9: Traženje LS-optimalne 3-particije skupa $\{1,2,4,8,9,10,12,15,18,20\}$

Zadatak 4.9. Odredite ℓ_1-optimalnu 3-particiju skupa $\mathcal{A}$ iz prethodnog primjera uz istu početnu particiju $\Pi^{(0)} = \{\{1,2,4\},\{8,9\},\{10,12,15,18,20\}\}$. Razlikuje li se dobivena ℓ_1-optimalna 3-particija od ranije dobivene LS-optimalne 3-particije istog skupa?

Rješenje: Ne razlikuje!

Zadatak 4.10. Dopunite Tablicu 4.8 i Tablicu 4.9 odgovarajućim vrijednostima dualne funkcije cilja $\mathcal{G}$.

4.2.2 Inicijalizacija k-means algoritma početnim centrima

k-means algoritam možemo pokrenuti i izborom početnih centara $z_1, \ldots, z_k$. U tom slučaju najprije treba primijeniti princip minimalnih udaljenosti iz Koraka A i tako odrediti početnu particiju s klasterima $\pi_1, \ldots, \pi_k$.

Nakon toga pokrećemo Koraka B, određujemo centre $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ i izračunavamo vrijednost funkcije cilja $\mathcal{F}$. Nakon toga, kao u t. 4.2.1 pokreće se Korak A s brojevima $c_1, \ldots, c_k$, na osnovi kojih principom minimalnih udaljenosti određujemo novu particiju s novim klasterima. Postupak se dalje ponavlja toliko dugo dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera također postanu jednaki).

Kako bi u ovom slučaju k-means algoritam dao particiju što bližu globalno optimalnoj, početni centri $z_1, \ldots, z_k$ trebaju biti što bolje izabrani: trebaju biti što bliži centrima klastera unaprijed nepoznate optimalne particije. U literaturi se mogu pronaći metode za procjenu dobrih početnih centara (vidi primjerice [69, 81]).

Primjer 4.11. *Zadan je skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *iz* Primjera 3.14, *str. 40. Za početne centre* $z_1 = 3$, $z_2 = 8$, $z_3 = 10$ *primjenom k-means algoritma treba pronaći LS-optimalnu* 3*-particiju. Također treba odrediti i vrijednosti odgovarajuće dualne funkcije cilja* $\mathcal{G}$.

Najprije primjenom LS-kvazimetričke funkcije korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 3$, $z_2 = 8$, $z_3 = 10$ određujemo početnu particiju $\Pi^{(0)} = \{\{2, 4\}, \{8\}, \{10, 16\}\}$. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.10 i na Slici 4.10.

Iter.	Klasteri			Centri			Funkcija cilja	Funkcija cilja
	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}$	$\mathcal{G}$
1	{2,4}	{8}	{10,16}	3	8	13	20	6+9+128=143
2	{2,4}	{8,10}	{16}	3	9	16	4	8+32+121=161
3	{2,4}	{8,10}	{16}	3	9	16	4	8+32+121=161

Tablica 4.10: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$

Slika 4.10: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$

Primjer 4.12. *Zadan je skup* $\mathcal{A} = \{2, 3, 5, 7, 9, 12, 13, 15\}$. *Primjenom k-means algoritma uz početne centre* $z_1 = 4$, $z_2 = 10$, $z_3 = 14$ *treba pronaći* ℓ_1*-optimalnu* 3*-particiju.*

Uz primjenu ℓ_1-metričke funkcije korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 4$, $z_2 = 10$, $z_3 = 14$ određujemo početnu particiju $\Pi^{(0)} = \{\{2, 3, 5, 7\}, \{9, 12\}, \{13, 15\}\}$, koja je ujedno i ℓ_1-optimalna 3-particija (vidi Tablicu 4.11).

Iter.	Klasteri			Centri			Funkcija cilja
	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_1$
1	{2,3,5,7}	{9,12}	{13,15}	4	10.5	14	7+3+2=12
2	{2,3,5,7}	{9,12}	{13,15}	4	10.5	14	7+3+2=12

Tablica 4.11: Traženje ℓ_1-optimalne 3-particije skupa $\mathcal{A} = \{2, 3, 5, 7, 9, 12, 13, 15\}$

4.3 Traženje lokalno optimalne k-particije podataka s više obilježja

Neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \colon i = 1, \ldots, m\} \subset \mathbb{R}^n$ skup podataka, a $\mathcal{P}(\mathcal{A}; k)$ skup svih njegovih k-particija (vidi Definiciju 3.1, str. 23). Broj elemenata skupa $\mathcal{P}(\mathcal{A}; k)$ određuje se prema (3.1) (vidi također Tablicu 3.1).

Neka je $\Pi = \{\pi_1, \ldots, \pi_k\} \in \mathcal{P}(\mathcal{A}; k)$ neka k-particija. Ako uvedemo neku kvazimetričku funkciju $d \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ (vidi t. 3, str. 23), onda svakom klasteru možemo pridružiti njegov centar

$$c_j \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k,$$

a prema (3.2) uvodimo kriterijsku funkciju cilja

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a). \tag{4.3}$$

Funkcija $\mathcal{F}$ pokazuje ukupno „rasipanje" elemenata svih klastera do njihovih centara. Što je vrijednost funkcije $\mathcal{F}$ manja, time je „rasipanje" manje, što znači da su klasteri kompaktniji i međusobno bolje razdvojeni. Problem traženja optimalne k-particije zapisujemo na sljedeći način

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi).$$

Traženje optimalne particije putem pretraživanja cijelog skupa $\mathcal{P}(\mathcal{A}; k)$ općenito nije moguće u prihvatljivom vremenu.

Za traženje LOP koristit ćemo već korišteni k-means algoritam. Ovaj algoritam nema ambiciju pronaći najbolju (globalno optimalnu) particiju. O tome ćemo nešto više reći u t. 4.5, str. 87. k-means algoritam u slučaju podataka s dva ili više obilježja može se opisati sa sljedeća dva koraka [31, 34, 39, 51, 53, 60].

Algoritam 4.3. *[k-means algoritam]*

Korak A: *Pridruživanje (assignment step). Poznavanjem međusobno različitih točaka $z_1, \ldots, z_k$, skup $\mathcal{A}$ treba grupirati u k disjunktnih klastera $\pi_1, \ldots, \pi_k$ korištenjem principa minimalnih udaljenosti*

$$\pi_j = \{a \in \mathcal{A} \colon d(z_j, a) \leq d(z_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k.$$

Korak B: *Korekcija (update step). Za poznatu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ treba definirati centre klastera*

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k.$$

(Centri $c_1, \ldots, c_k$ iz Koraka B nakon jedne iteracije postaju točke $z_1, \ldots, z_k$ u sljedećoj iteraciji.)

U Koraku A principom minimalnih udaljenosti cijeli skup $\mathcal{A}$ razdjeljuje se u k skupina prema njihovoj d-bliskosti pojedinim točkama z_j.

U Koraku B svakom klasteru particije Π pridružujemo njegove centre. Ako je d, LS-kvazimetrička funkcija, ti su centri centroidi klastera (vidi t. 3.4.1, str. 46), a ako je d, ℓ_1-metrička funkcija, ti su centri medijani klastera (vidi t. 3.4.2, str. 51). Poznavanjem centara možemo odrediti i vrijednost kriterijske funkcije cilja (4.3).

k-means algoritam sukcesivno ponavlja navedene korake toliko dugo dok se particije ne počnu ponavljati (u tom slučaju vrijednost funkcije cilja prestane opadati, a centri klastera trenutne i prethodne particije međusobno se podudaraju). Algoritam se može pokrenuti zadavanjem početne particije $\Pi^{(0)} = \{\pi_1, \ldots, \pi_k\}$ ili zadavanjem početnih centara (k međusobno različitih točaka $z_1, \ldots, z_k$).

Primjedba 4.6. Kao što smo već spomenuli u Primjedbi 4.4, str. 69, u ovom udžbeniku koristit ćemo *k-means algoritam uz primjenu LS-kvazimetričke funkcije* i *k-means algoritam uz primjenu ℓ_1-metričke funkcije*, iako se u znanstvenoj literaturi može pronaći i primjena brojnih drugih kvazimetričkih funkcija (vidi primjerice [31, 53, 54]). Spomenimo još da se k-means algoritam uz primjenu ℓ_1-metričke funkcije u literaturi pojavljuje i pod nazivom k-medijan algoritam.

Primjedba 4.7. U slučaju podataka s više obilježja ($\mathcal{A} \subset \mathbb{R}^n$, $n \geq 2$), princip minimalnih udaljenosti bitno se razlikuje u slučaju primjene LS-kvazimetričke funkcije, odnosno u slučaju primjene ℓ_1-metričke funkcije. Ispišite formule iz Koraku A u jednom i drugom slučaju i pokušajte uočiti spomenutu razliku na konkretnim podacima.

4.3.1 Inicijalizacija k-means algoritma početnom particijom

Ako algoritam pokrenemo početnom particijom $\Pi^{(0)} = \{\pi_1, \ldots, \pi_k\}$, najprije će se izvoditi Korak B. Nakon određivanja centara $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$, možemo odrediti vrijednost funkcije cilja $\mathcal{F}$ (također možemo odrediti i prosječno „rasipanje" podataka po klasterima). Nakon toga pokreće se Korak A za točke $c_1, \ldots, c_k$, na osnovi kojih principom minimalnih udaljenosti određujemo novu particiju s novim klasterima. Postupak se dalje ponavlja toliko dugo dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera također postanu jednaki).

U svakom koraku k-means algoritma vrijednost funkcije cilja $\mathcal{F}$ snižava se i asimptotski približava lokalno najmanjoj mogućoj vrijednosti [31, 68].

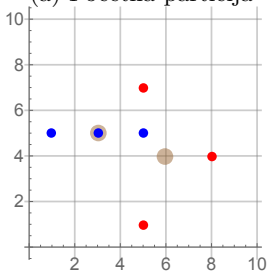
Kako bi k-means algoritam dao particiju što bliže globalno optimalnoj,

particija s kojom započinjemo iterativni postupak mora biti što bolje izabrana.

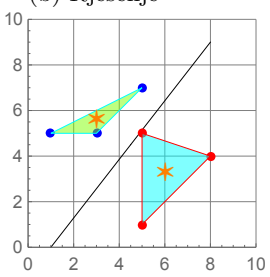
Primjer 4.13. *Skup* $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 6\} \subset \mathbb{R}^2$ *zadan je niže navedenom tablicom i prikazan na Slici 4.11a. Uz poznavanje početne particije* $\Pi^{(0)} = \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^6\}\}$ *primjenom k-means algoritma (Algoritam 4.3) treba odrediti LS-optimalnu 2-particiju skupa* $\mathcal{A}$.

i	1	2	3	4	5	6
x_i	1	3	5	5	5	8
y_i	5	5	5	1	7	4

(a) Početna particija



(b) Rješenje



Slika 4.11: k-means iterativni proces

U cilju traženja LS-optimalne 2-particije skupa $\mathcal{A}$ koristit ćemo k-means algoritam uz primjenu LS-kvazimetričke funkcije. Tijek iterativnog procesa možemo pratiti u Tablici 4.12.

Iter.	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
1	$\{a^1, a^2, a^3\}$	$\{a^4, a^5, a^6\}$	(3,5)	(6,4)	32	15
2	$\{a^1, a^2, a^5\}$	$\{a^3, a^4, a^6\}$	$(3, \frac{17}{3})$	$(6, \frac{10}{3})$	$76/3 \approx 25.3$	$65/3 \approx 21.67$
3	$\{a^1, a^2, a^5\}$	$\{a^3, a^4, a^6\}$	$(3, \frac{17}{3})$	$(6, \frac{10}{3})$	$76/3 \approx 25.3$	$65/3 \approx 21.67$

Tablica 4.12: Traženje LS-optimalne 2-particije skupa $\mathcal{A} \subset \mathbb{R}^2$

Prilikom određivanja novih klastera primjenom principa minimalnih udaljenosti (Korak A) u slučaju traženja LS-optimalne 2-particije skupa točaka iz ravnine, jednostavno treba povući simetralu spojnice centroida. Ova simetrala razdjeljuje dva klastera (vidi Sliku 4.11b). Postupak postaje znatno

složeniji za $k > 2$, a pogotovo za $n > 2$, i vodi na konstrukciju poznatih Voronoijevih dijagrama [48, 60, 73].

Primijetimo još da je za izračunavanje odgovarajuće dualne funkcije $\mathcal{G}$ prethodno bilo potrebno odrediti centroid čitavog skupa $c = (\frac{9}{2}, \frac{9}{2})$. Tada je

$$\mathcal{G}(\Pi) = |\pi_1|(c - c_1)^2 + |\pi_2|(c - c_2)^2.$$

Primjer 4.14. *Skup $\mathcal{A}$ iz Primjera 4.13 grupirat ćemo primjenom ℓ_1-metričke funkcije počevši od iste početne particije. Tijek iterativnog postupka može se pratiti u Tablici 4.13* [3].

It.	π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$
1	$\{a^1, a^2, a^3\}$	$\{a^4, a^5, a^6\}$	(3,5)	(5,4)	4+9=13
2	$\{a^1, a^2\}$	$\{a^3, a^4, a^5, a^6\}$	$(2, 5)$	$(5, 4.5)$	$2 + 10 = 12$
3	$\{a^1, a^2\}$	$\{a^3, a^4, a^5, a^6\}$	$(2, 5)$	$(5, 4.5)$	$2 + 10 = 12$

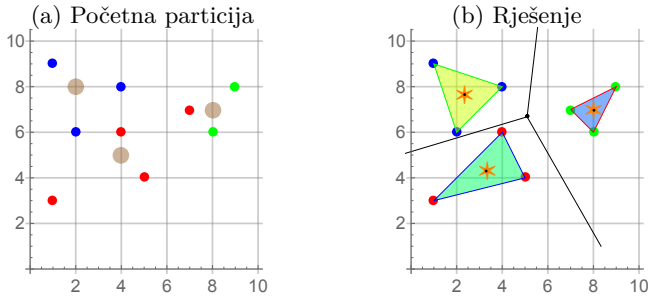
Tablica 4.13: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} \subset \mathbb{R}^2$

Primjer 4.15. *Skup $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, 9\} \subset \mathbb{R}^2$ zadan je niže navedenom tablicom i prikazan na Slici 4.12a. Uz poznavanje početne particije $\Pi^{(0)}$ s klasterima $\pi_1^{(0)} = \{a^1, a^2, a^3\}$, $\pi_2^{(0)} = \{a^4, a^5, a^6, a^7\}$, $\pi_3^{(0)} = \{a^8, a^9\}$ primjenom k-means algoritma treba odrediti LS-optimalnu 3-particiju skupa $\mathcal{A}$.*

	$\pi_1^{(0)}$			$\pi_2^{(0)}$				$\pi_3^{(0)}$	
i	1	2	3	4	5	6	7	8	9
x_i	1	2	4	1	4	7	5	8	9
y_i	9	6	8	3	6	7	4	6	8

Korištenjem k-means algoritma uz primjenu LS-kvazimetričke funkcije već u prvoj iteraciji dobivamo rješenje prikazano na Slici 4.12b. Na LS-lokalno optimalnoj particiji $\Pi^\star = \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^7\}, \{a^6, a^8, a^9\}\}$ funkcija cilja postiže vrijednost $\mathcal{F}(\Pi^\star) = \frac{80}{3}$. Na Slici 4.12b klastere optimalne particije $\Pi^\star$ razdvaja Voronoijev dijagram.

[3] Primijetite da se u slučaju primjene ℓ_1-metričke funkcije kod implementacije principa minimalnih udaljenosti (Korak A) ne može koristiti simetrala spojnice centara.



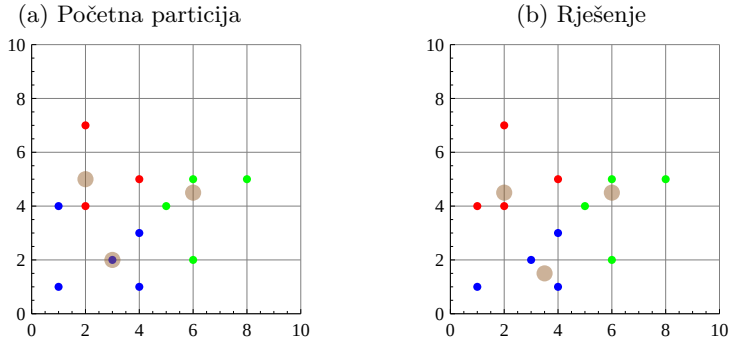
Slika 4.12: k-means iterativni proces

Zadatak 4.11. Pronađite ℓ_1-optimalnu 3-particiju skupa $\mathcal{A}$ iz Primjera 4.15.

Primjer 4.16. *Skup* $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 12\} \subset \mathbb{R}^2$ *zadan je niže navedenom tablicom i prikazan na Slici 4.13a. Uz poznavanje početne particije* $\Pi^{(0)}$ *s klasterima* $\pi_1^{(0)} = \{a^1, a^2, a^3, a^4, a^5\}$, $\pi_2^{(0)} = \{a^6, a^7, a^8\}$, $\pi_3^{(0)} = \{a^9, a^{10}, a^{11}, a^{12}\}$ *primjenom k-means algoritma treba odrediti* ℓ_1*-optimalnu* 3*-particiju.*

	$\pi_1^{(0)}$					$\pi_2^{(0)}$			$\pi_3^{(0)}$			
i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	1	3	4	4	2	2	4	5	6	6	8
y_i	1	4	2	1	3	4	7	5	4	2	5	5

Centri početne particije su: $c_1 = (3, 2)$, $c_2 = (2, 5)$, $c_3 = (6, 4.5)$ (Slika 4.13a), a funkcija cilja $\mathcal{F}_1(\Pi^{(0)}) = 23$.



Slika 4.13: k-means iterativni proces

Primjenom principa minimalnih udaljenosti uz ℓ_1-metričku funkciju dobivamo novu particiju $\Pi^{(1)}$ s klasterima $\pi_1 = \{a^1, a^3, a^4, a^5\}$, $\pi_2 = \{a^2, a^6, a^7, a^8\}$, $\pi_3 = \{a^9, a^{10}, a^{11}, a^{12}\}$. Možemo li u ovom slučaju koristiti Voronoijev dijagram? Nakon toga izračunamo nove centre $c_1 = (3.5, 1.5)$, $c_2 = (2, 4.5)$, $c_3 = (6, 4.5)$ (vidi Sliku 4.13b) i novu vrijednost funkcije cilja $\mathcal{F}_1(\Pi^{(1)}) = 21$. Je li to ℓ_1-optimalna 3-particija skupa $\mathcal{A}$?

Zadatak 4.12. Pronađite LS-optimalnu 3-particiju skupa $\mathcal{A}$ iz Primjera 4.16 uz istu početnu particiju.

Rješenje:
Optimalna particija $\Pi^\star$ s klasterima $\pi_1^\star = \{\{a^1, a^3, a^4, a^5\}$, $\pi_2^\star = \{a^2, a^6, a^7, a^8\}$, $\pi_3^\star = \{a^9, a^{10}, a^{11}, a^{12}\}$ podudara se s ℓ_1-optimalnom 3-particijom. Centroidi klastera su $c_1 = (3, 7/4)$, $c_2 = (9/4, 5)$, $c_3 = (25/4, 4)$, a vrijednosti kriterijskih funkcija cilja $\mathcal{F}_{LS}(\Pi^\star) = 30.25$; $\mathcal{G}(\Pi^\star) = 58.33$.

4.3.2 Inicijalizacija k-means algoritma početnim centrima

k-means algoritam možemo pokrenuti i izborom početnih centara $z_1, \ldots, z_k$. U tom slučaju najprije treba primijeniti princip minimalnih udaljenosti iz Koraka A i tako odrediti početnu particiju $\Pi^{(0)}$ s klasterima $\pi_1^{(0)}, \ldots, \pi_k^{(0)}$. Nakon toga pokrećemo Koraka B, određujemo centre $c_1, \ldots, c_k$ klastera $\pi_1^{(0)}$, $\ldots, \pi_k^{(0)}$ i izračunavamo vrijednost funkcije cilja $\mathcal{F}$. Nakon toga pokreće se Korak A s točkama $c_1, \ldots, c_k$, na osnovi kojih principom minimalnih udaljenosti određujemo novu particiju s novim klasterima. Postupak se dalje ponavlja toliko dugo dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera također postanu jednaki).

Kako bi u ovom slučaju k-means algoritam dao particiju što bliže globalno optimalnoj, početni centri $z_1, \ldots, z_k$ trebaju biti što bolje izabrani: trebaju biti što bliži centrima klastera unaprijed nepoznate optimalne particije. U literaturi se mogu pronaći metode za procjenu dobrih početnih centara (vidi primjerice [69, 81]).

Primjer 4.17. *Skup $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, 8\} \subset \mathbb{R}^2$ zadan je niže navedenom tablicom i prikazan na Slici 4.14a. Primjenom k-means algoritma uz poznavanje početnih centara $z_1 = (4, 4)$, $z_2 = (8, 4)$ treba pronaći LS-optimalnu 2-particiju.*

i	1	2	3	4	5	6	7	8
x_i	2	3	5	6	7	8	9	10
y_i	3	6	8	5	7	1	5	3

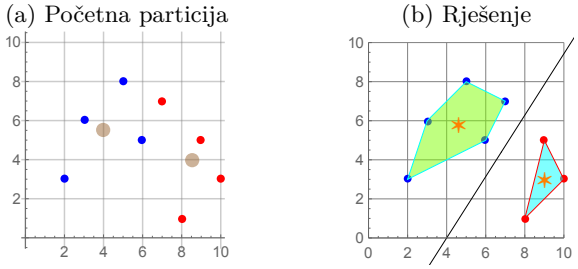
Najprije uz primjenu LS-kvazimetričke funkcije korištenjem principa minimalnih udaljenosti treba odrediti početnu particiju

$$\Pi^{(0)} = \{\pi_1^{(0)}, \pi_2^{(0)}\}, \qquad \pi_1^{(0)} = \{a^1, a^2, a^3, a^4\}, \quad \pi_2^{(0)} = \{a^5, a^6, a^7, a^8\}.$$

Geometrijski se to postiže tako da odredimo simetralu spojnice točaka z_1, z_2 (Voronoijev dijagram!) Daljnji tijek iterativnog postupka vidljiv je u Tablici 4.14, a rješenje se može vidjeti i na Slici 4.14b.

Iter.	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
1	$\{a^1, a^2, a^3, a^4\}$	$\{a^5, a^6, a^7, a^8\}$	(4,5.5)	(8.5,4)	48	45
2	$\{a^1, a^2, a^3, a^4, a^5\}$	$\{a^6, a^7, a^8\}$	(4.6,5.8)	(9,3)	42	51
3	$\{a^1, a^2, a^3, a^4, a^5\}$	$\{a^6, a^7, a^8\}$	(4.6,5.8)	(9,3)	42	51

Tablica 4.14: Traženje LS-optimalne 2-particije skupa $\mathcal{A} \subset \mathbb{R}^2$



Slika 4.14: k-means iterativni proces

Zadatak 4.13. Odredite ℓ_1-optimalnu 2-particiju skupa $\mathcal{A}$ iz Primjera 4.17 uz primijenu k-means algoritma i iste početne centre $z_1 = (4, 4)$, $z_2 = (8, 4)$. Razlikuje li se dobiveno rješenje?

Rješenje: $\Pi^\star = \{\{a^1, a^2, a^3, a^4, a^5\}, \{a^6, a^7, a^8\}\}$; $c_1 = (5, 6)$, $c_2 = (9, 3)$; $\mathcal{F}_1(\Pi^\star) = 21$.

Zadatak 4.14. Odredite LS-optimalnu i ℓ_1-optimalnu 2-particiju skupa $\mathcal{A}$ iz Primjera 4.17 uz primijenu k-means algoritma i početne centre: $z_1 = (1, 1)$ i $z_2 = (5, 5)$. Razlikuje li se dobivene particije?

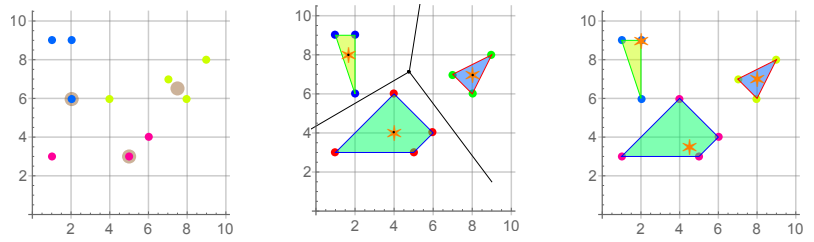
Primjer 4.18. *Skup* $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 10\} \subset \mathbb{R}^2$ *zadan je niže navedenom tablicom i prikazan na Slici 4.15a. Primjenom* k*-means algoritma i poznavanjem početnih centara* $z_1 = (2, 8)$, $z_2 = (5, 4)$, $z_3 = (6, 6)$ *treba odrediti LS-optimalnu* 3*-particiju skupa* $\mathcal{A}$.

i	1	2	3	4	5	6	7	8	9	10
x_i	1	2	2	1	5	6	4	7	8	9
y_i	9	9	6	3	3	4	6	7	6	8

Na osnovi početnih centara, uz primjenu LS-kvazimetričke funkcije, korištenjem principa minimalnih udaljenosti najprije treba odrediti početnu particiju $\Pi^{(0)} = \{\pi_1^{(0)}, \pi_2^{(0)}, \pi_3^{(0)}\}$. Geometrijski, početna particija Π dobiva se tako da najprije nacrtamo sve tri simetrale spojnica točaka z_1, z_2, z_3 i nakon toga odredimo tzv. Voronoijev dijagram (vidi primjerice [31, 51, 60]). Dobivamo (vidi Sliku 4.15a)

$$\pi_1^{(0)} = \{a^1, a^2, a^3\}, \quad \pi_2^{(0)} = \{a^4, a^5, a^6\}, \quad \pi_3^{(0)} = \{a^7, a^8, a^9, a^{10}\}.$$

(a) Početna particija (b) LS-optimalna particija (c) ℓ_1-optimalna particija



Slika 4.15: k-means iterativni proces

Centroidi ovih klastera su $c_1 = (\frac{5}{3}, 8)$, $c_2 = (4, \frac{10}{3})$, $c_3 = (7, \frac{27}{4})$. Nakon toga izračunamo vrijednost funkcije cilja $\mathcal{F}_{LS} = \frac{457}{12} = 38.08$ i vrijednost odgovarajuće dualne funkcije $\mathcal{G} = \frac{5119}{60} = 85.32$.

Na kraju k-means algoritma dobivamo particiju $\Pi^\star = \{\pi_1^\star, \pi_2^\star, \pi_3^\star\}$, gdje je

$$\pi_1^\star = \{a^1, a^2, a^3\}, \quad \pi_2^\star = \{a^4, a^5, a^6, a^7\}, \quad \pi_3^\star = \{a^8, a^9, a^{10}\},$$

uz $\mathcal{F}_{LS}^\star = \frac{92}{3} = 30.67$ i $\mathcal{G}^\star = \frac{1391}{15} = 92.73$.

Primjer 4.19. *Za skup $\mathcal{A}$ iz prethodnog primjera potražit ćemo ℓ_1-optimalnu 3-particiju uz iste početne centre.*

Korištenjem k-means algoritma, uz primjenu ℓ_1-metričke funkcije dobivamo istu optimalnu particiju s centrima $c_1 = (2, 9)$, $c_2 = (\frac{9}{2}, \frac{7}{2})$, $c_3 = (8, 7)$, (vidi Sliku 4.15c) i vrijednosti funkcije cilja $\mathcal{F}_1 = 18$.

4.4 Traženje lokalno optimalne k-particije podataka s težinama

Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n \colon i = 1, \ldots, m\}$ skup podataka kojima su pridružene odgovarajuće težine $w_i > 0$, a $\mathcal{P}(\mathcal{A}; k)$ skup svih njegovih k-particija. Neka je nadalje $\Pi = \{\pi_1, \ldots, \pi_k\} \in \mathcal{P}(\mathcal{A}; k)$ neka k-particija. Ako uvedemo neku kvazimetričku funkciju $d \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, onda analogno (3.40), str. 46, svakom klasteru možemo pridružiti njegov težinski centar

$$c_j \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a^s \in \pi_j} w_s d(x, a^s), \quad j = 1, \ldots, k,$$

a analogno (3.41), str. 46, odnosno (3.50), str. 53, možemo uvesti odgovarajuće težinske kriterijske funkcije cilja

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^s \in \pi_j} w_s d(c_j, a^s),$$
$$F(z_1, \ldots, z_k) = \sum_{i=1}^{m} w_i \min_{j=1,\ldots,k} d(z_j, a^i).$$

Pokazat ćemo nekoliko primjera na podacima s jednim obilježjem i nekoliko primjera na podacima s dva obilježja.

4.4.1 Princip najmanjih kvadrata

Specijalno, ako izaberemo LS-kvazimetričku funkciju, centri klastera postaju težinski centroidi

$$c_j = \tfrac{1}{\kappa_j} \sum_{a^s \in \pi_j} w_s a^s, \qquad \kappa_j = \sum_{a^s \in \pi_j} w_s, \quad j = 1, \ldots, k,$$

a kriterijske funkcije cilja postaju

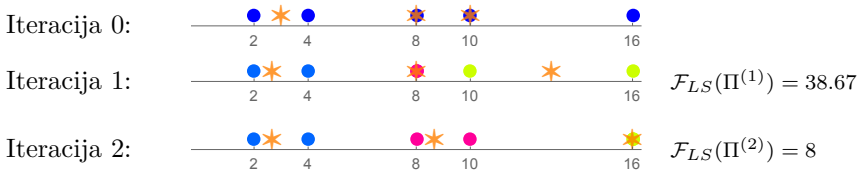
$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^s \in \pi_j} w_s \|c_j - a^s\|_2^2,$$
$$\mathcal{G} = \sum_{j=1}^{k} \Big(\sum_{a^s \in \pi_j} w_s \Big) \|c - c_j\|_2^2, \quad c = \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i, \quad W = \sum_{i=1}^{m} w_i,$$
$$F(z_1, \ldots, z_k) = \sum_{i=1}^{m} w_i \min_{j=1,\ldots,k} \|z_j - a^i\|_2^2.$$

Primjer 4.20. *Zadan je skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *iz Primjera 3.14, str. 40, pri čemu smo svakom elementu skupa* $\mathcal{A}$ *pridružili odgovarajuću težinu* $w_i \in \{2, 1, 4, 2, 2\}$. *Primjenom k-means algoritma uz početne centre* $z_1 = 3$, $z_2 = 8$, $z_3 = 10$ *treba pronaći LS-optimalnu* 3*-particiju.*

Uz primjenu LS-kvazimetričke funkcije, korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 3$, $z_2 = 8$, $z_3 = 10$ dobivamo početnu particiju $\Pi^{(0)} = \{\{2, 4\}, \{8\}, \{10, 16\}\}$. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.15 i na Slici 4.16.

Iteracija	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}$
1	$\{2, 4\}$	$\{8\}$	$\{10, 16\}$	8/3	8	13	38.67
2	$\{2, 4\}$	$\{8, 10\}$	$\{16\}$	8/3	26/3	16	8
3	$\{2, 4\}$	$\{8, 10\}$	$\{16\}$	8/3	26/3	16	8

Tablica 4.15: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$ s težinama



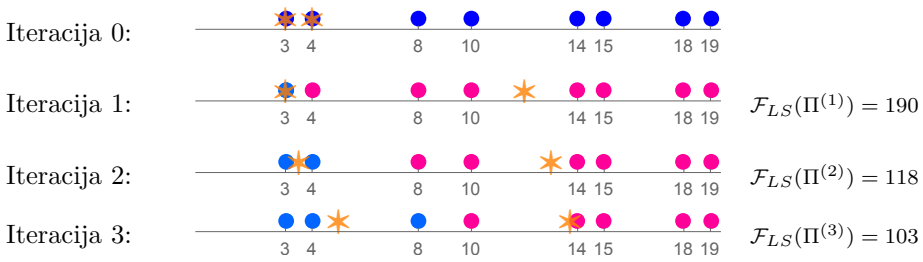
Slika 4.16: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$ s težinama

Primjer 4.21. *Primjenom k-means algoritma treba pronaći LS-optimalnu* 2*-particiju skupa* $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$ *iz Primjera 4.5, str. 64, pri čemu ćemo svakom elementu skupa* $\mathcal{A}$ *pridružiti odgovarajuću težinu* $w_i \in \{1, 1, 1, 3, 1, 1, 1, 1\}$. *Lokalno optimalnu* 2*-particiju tražit ćemo krenuvši od početnih centara* $z_1 = 3$ *i* $z_2 = 4$.

Primijetite da se ovaj primjer razlikuje od Primjera 4.5 samo po tome što četvrti podatak ima težinu $w_4 = 3$, dok se u Primjeru 4.5 pretpostavljalo da su sve težine međusobno jednake (primjerice, $w_i = 1$). Uz primjenu LS-kvazimetričke funkcije, korištenjem principa minimalnih udaljenosti na osnovi početnih centara $z_1 = 3$ i $z_2 = 4$ dobivamo početnu particiju $\Pi^{(0)} = \{\{3\}, \{4, 8, 10, 14, 15, 18, 19\}\}$. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.16 ili na Slici 4.17.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
1	{3}	{4, 8, 10, 14, 15, 18, 19}	3	12.57	190.0	72.9
2	{3, 4}	{8,10, 14, 15, 18, 19}	3.5	13	113.75	144.4
3	{3, 4, 8}	{10,14, 15, 18, 19}	5	13.71	103.43	159.47
4	{3, 4, 8}	{10,14, 15, 18, 19}	5	13.71	103.43	159.47

Tablica 4.16: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$ s težinama



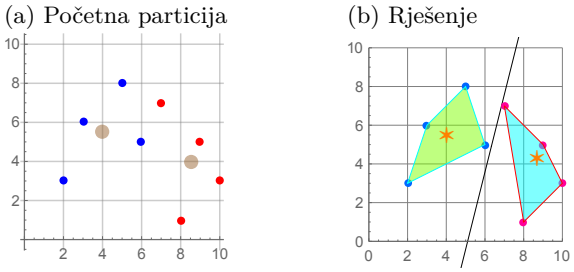
Slika 4.17: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$

Kao što možemo primijetiti, vrlo mala promjena u težinama podataka dovela je do lokalno optimalne particije koja se razlikuje od one dobivene u Primjeru 4.5.

Zadatak 4.15. Budući da optimalnu particiju skupa s jednim obilježjem ima smisla tražiti samo između particija čiji se klasteri međusobno nastavljaju (vidi Primjedbu 3.1, str. 35) i da skup iz prethodnog primjera ima samo 7 (formula (3.16)) takvih particija, provjerite je li k-means algoritam pronašao GOP.

Primjer 4.22. *U Primjeru 4.17, str. 79, tražili smo LS-optimalnu* 2*-particiju skupa* $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, 8\} \subset \mathbb{R}^2$ *poznavanjem početnih centara* $z_1 = (4, 4)$ *i* $z_2 = (8, 4)$ *i uz pretpostavku da su težine svih podataka međusobno jednake (primjerice,* $w_i = 1$*). Sada ćemo pretpostaviti da smo podatku* $(9, 5)$ *pridružili težinu* 3*, dok su težine svih ostalih podataka ostale* 1 *te potražiti lokalno optimalnu* 2*-particiju krenuvši od istih početnih centara.*

Kao što se vidi na Slici 4.18, dobivena particija bitno se razlikuje od one dobivene u Primjeru 4.17



Slika 4.18: k-means iterativni proces

4.4.2 Princip najmanjih apsolutnih odstupanja

Ako izaberemo ℓ_1-metričku funkciju, centri klastera postaju težinski medijani podataka tog klastera

$$c_j \in \underset{\pi_j}{\operatorname{med}}(w_s, a^s), \quad j = 1, \ldots, k,$$

a kriterijske funkcije cilja postaju

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^s \in \pi_j} w_s \|c_j - a^s\|_1,$$

$$F(z_1, \ldots, z_k) = \sum_{i=1}^{m} w_i \min_{j=1,\ldots,k} \|z_j - a^i\|_1.$$

Kao što smo već spomenuli, problem određivanja težinskog medijana složeni je postupak [24, 50], ali u slučaju cjelobrojnih težina, kao što smo pokazali u t. 2.1.3, str. 8, problem se može svesti na određivanje običnog medijana modificiranog skupa podataka. Nova verzija programskog sustava *Mathematica* omogućava izračunavanje težinskog medijana podataka iz $\mathbb{R}^n$ naredbom: `Median[WeightedData[A,W]`, gdje je `A` skup iz $\mathbb{R}^n$, a `W` lista težina.

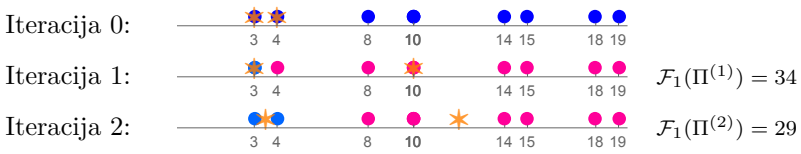
Primjer 4.23. *Ponovo promatramo skup* $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$ *iz Primjera 4.6, str. 65. Ovim elementima pridružit ćemo odgovarajuće težine* $w_i \in \{1, 1, 1, 3, 1, 1, 1, 1\}$ *i potražiti* ℓ_1*-optimalnu 2-particiju primjenom k-medijan algoritma s istim početnim centrima* $z_1 = 3$, $z_2 = 4$.

Uz primjenu ℓ_1-metričke funkcije korištenjem principa minimalnih udaljenosti za početne centre $z_1 = 3$, $z_2 = 4$ dobivamo istu početnu particiju kao u Primjeru 4.6

$\Pi^{(0)} = \{\{3\}, \{4, 8, 10, 14, 15, 18, 19\}\}$. Daljnji tijek iterativnog procesa može se pratiti u Tablici 4.17 i na Slici 4.19.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_1(\Pi)$
1	{3}	{4, 8, 10, 14, 15, 18, 19}	3	10	34
2	{3, 4, 8}	{10, 14, 15, 18, 19}	3.5	12	29
3	{3, 4, 8}	{10, 14, 15, 18, 19}	3.5	12	29

Tablica 4.17: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$



Slika 4.19: Traženje ℓ_1-optimalne 2-particije skupa $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$

Kao što možemo primijetiti, vrlo mala promjena u težinama podataka dovela je do lokalno optimalne particije koja se razlikuje od one u Primjeru 4.6.

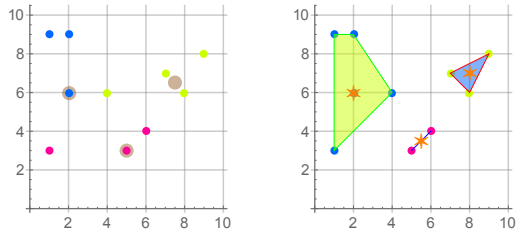
Zadatak 4.16. Kako optimalnu particiju skupa s jednim obilježjem ima smisla tražiti samo među particijama čiji se klasteri međusobno nastavljaju (vidi Primjedbu 3.1, str. 35) i kako skup iz prethodnog primjera ima samo 7 (formula (3.16)) takvih particija, provjerite je li k-means algoritam pronašao GOP.

Zadatak 4.17. Odredite ℓ_1-optimalnu 3-particiju skup $\mathcal{A}$ iz prethodnog primjera uz iste početne centre $z_1 = 3$, $z_2 = 8$, $z_3 = 10$. Razlikuje li se dobivena ℓ_1-optimalna 3-particija od ranije dobivene LS-optimalne 3-particije istog skupa?

Rješenje: Ne razlikuje!

Primjer 4.24. *U Primjeru 4.19, str. 81, tražili smo ℓ_1-optimalnu 3-particiju skupa $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, 10\} \subset \mathbb{R}^2$ (iz Primjera 4.18, str. 80) poznavanjem početnih centara $z_1 = (2, 8)$, $z_2 = (5, 4)$, $z_3 = (6, 6)$ i uz pretpostavku da su težine svih podataka međusobno jednake (primjerice, $w_i = 1$).*

(a) Početna particija (b) ℓ_1-optimalna particija



Slika 4.20: k-medijan iterativni proces

Sada ćemo pretpostaviti da smo podatku $(2, 6)$ *pridružili težinu* 3*, dok su težine svih ostalih podataka ostale* 1 *te potražiti* ℓ_1*-lokalno optimalnu* 3*-particiju krenuvši od istih početnih centara.*

Kao što možemo vidjeti, uspoređivanjem Slike 4.15 i Slike 4.20 dobivena particija bitno se razlikuje od one dobivene u Primjeru 4.19.

Zadatak 4.18. Odredite ℓ_1-optimalnu 2-particiju skupa $\mathcal{A}$ iz prethodnog primjera uz primijenu k-medijan algoritma i početne centre $z_1 = (2, 8)$, $z_2 = (6, 6)$.

4.5 Traženje globalno optimalne particije

Kao što smo već ranije spomenuli, traženje GOP složen je optimizacijski problem. Direktno pretraživanje skupa svih particija $\mathcal{P}(\mathcal{A}; k)$ nije prihvatljivo zbog veličine tog skupa (vidi Tablicu 3.1, str. 24). Zato se u stručnoj literaturi ovi problemi nazivaju NP-teški problemi [40]. Traženje GOP putem minimizacije funkcije cilja $\mathcal{F}$ zadane s (3.41), str. 46, ili funkcije F zadane s (3.50), str. 53, također nije prihvatljivo jer je F nekonveksna i općenito nediferencijabilna funkcija s izuzetno velikim brojem varijabli [23, 46, 54], koja najčešće posjeduje i veliki broj stacionarnih točaka. Postoje neke specijalne situacije kod kojih se GOP može dobiti klasičnim metodama minimizacije ili pretraživanjem u prihvatljivom vremenu. Primjerice,

- u slučaju malog broja podataka;

- ako je broj obilježja $n = 1$, onda se GOP traži među particijama čiji se klasteri međusobno nastavljaju (Primjedba 3.1, str. 35). U tom slučaju skup $\mathcal{P}(\mathcal{A}; k)$ znatno je manji (vidi Tablicu 3.4, str. 35), ali i

u tom slučaju realni problemi iz primjena najčešće se ne mogu riješiti na taj način.

Zbog svega navedenog, u stručnoj literaturi mogu se pronaći brojni prijedlozi za rješavanje problema traženja GOP [1, 23, 43, 46, 54, 63]. Ipak, nijedna od poznatih metoda ne daje GOP već LOP, koja „na oko" može biti dobra, ali ne postoji mogućnost formalne provjere radi li se baš o GOP.

U ovom udžbeniku opisat ćemo jednu jednostavnu metodu za traženje GOP, koja zahtijeva samo poznavanje k-means algoritma. Metodu je 2006. godine predložio Friedrich Leisch[4] [34].

Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n \colon i = 1, \ldots, m\} \subset [\alpha, \beta]$, gdje je $[\alpha, \beta]$ hiperpravokutnik u kome se nalaze svi podaci. Ako je $n = 1$, $[\alpha, \beta]$ je obični segment realnih brojeva, ako je $n = 2$, $[\alpha, \beta]$ je pravokutnik u ravnini itd.

Pretpostavimo da skup $\mathcal{A}$ treba grupirati u $1 \leq k \leq m$ klastera $\pi_1, \ldots, \pi_k$ (sukladno Definiciji 3.1, str. 23), koji čine particiju Π. Pretpostavimo također da se kriterijska funkcija cilja $F \colon \mathbb{R}^{n \times k} \to \mathbb{R}$,

$$F(c_1, \ldots, c_k) = \sum_{i=1}^{m} \min_{j=1,\ldots k} d(c_j, a^i) \tag{4.4}$$

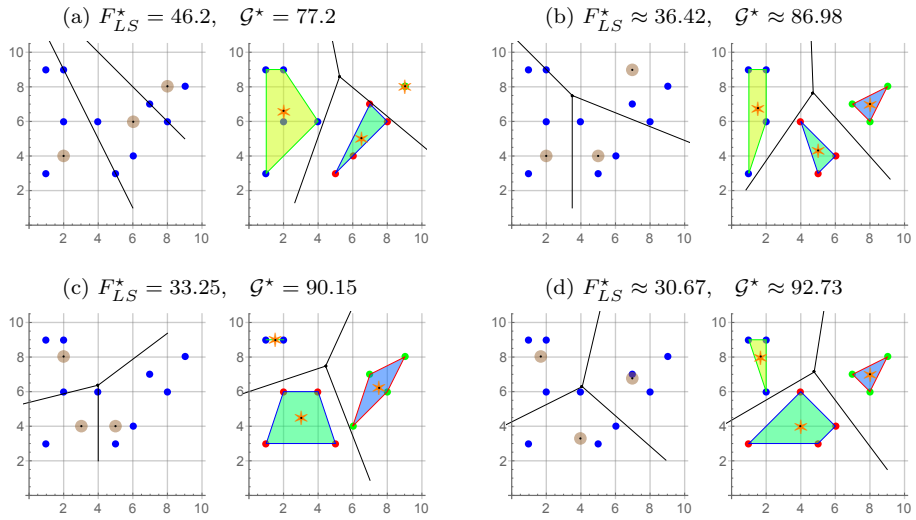
definira pomoću izabrane kvazimetričke funkcije $d \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$.

U hiperpravokutniku slučajno izaberemo k međusobno različitih točaka $z_1, \ldots, z_k$ i provedemo k-means algoritam. Na taj način dobivamo particiju $\Pi^{(1)} = \{\pi_1, \ldots, \pi_k\}$ s centrima $c_1, \ldots, c_k \in [\alpha, \beta]$ i vrijednost funkcije cilja $F(c_1, \ldots, c_k)$. Ako postupak ponovimo više puta i pamtimo samo particiju s najnižom vrijednosti funkcije cilja F, možemo se nadati da smo pronašli particiju koja je bliska GOP. U praktičnim primjenama (vidi primjerice [34, 54]) metoda pokazuje dobre rezultate, ali postupak može trajati prilično dugo. Metoda se može popraviti dodatnim postupcima za izbor početnih centara.

Primjer 4.25. *Razmotrimo ponovo problem traženja globalne LS-optimalne particije skupa $\mathcal{A}$ iz Primjera 4.18, str. 80.*

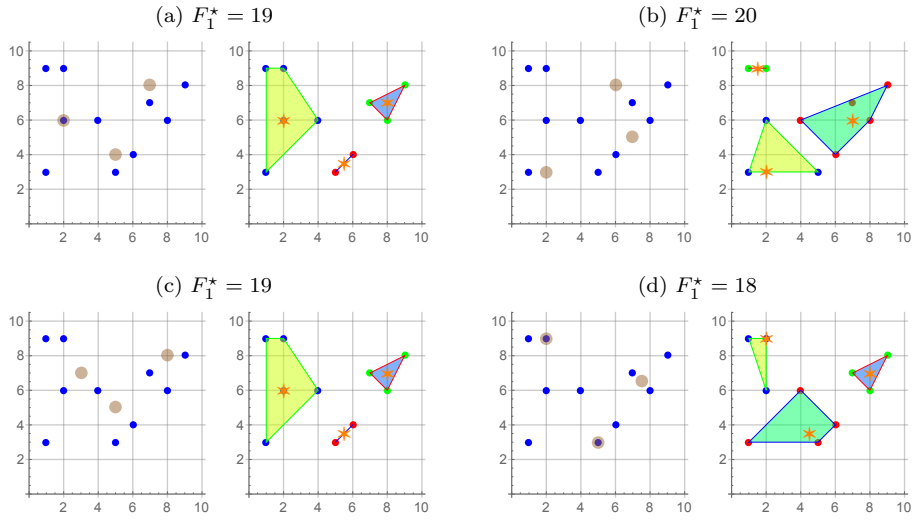
Na Slici 4.21 vidimo različite LOP dobivene različitim izborom početnih centara. Primijetite da LOP s najnižom vrijednosti funkcije cilja F nije dobivena s istim početnim centrima kao u Primjeru 4.18. Primijetite također da nije slučajno (vidi t. 3.3.4, str. 45), da se vrijednosti kriterijskih funkcija $\mathcal{F}$ i F podudaraju na LOP.

[4] Friedrich Leisch, Department of Statistics and Probability Theory, Vienna University of Technology, 1040 Vienna, Austria



Slika 4.21: Traženje LS-GOP metodom izbora slučajnih početnih centara

Primjer 4.26. *Razmotrimo sada problem traženja globalne ℓ_1-optimalne particije istog skupa $\mathcal{A}$ iz Primjera 4.18, str. 80, a koji smo također razmatrali u prethodnom primjeru kao primjer traženja globalne LS-optimalne particije.*



Slika 4.22: Traženje ℓ_1-GOP metodom izbora slučajnih početnih centara

Na Slici 4.22 vidimo različite LOP dobivene različitim izborom početnih centara. Primijetite da se najbolja LS-LOP iz prethodnog primjera podudara s najboljom ℓ_1-LOP. To je ujedno GOP u smislu LS-kvazimetrike i u smislu ℓ_1-metrike.

Poglavlje 5

Aglomerativni hijerarhijski algoritmi

Jedna mogućnost traženja optimalne particije su tzv. aglomerativni hijerarhijski algoritmi. Ovi algoritmi najviše se primjenjuju u području društvenih znanosti, biologije, medicine, arheologije, ali i u računarskim znanostima [30, 42, 71, 73].

5.1 Uvod i motivacija

Osnovna ideja aglomerativnih hijerarhijskih algoritama sastoji se u tome da polazeći od poznate particije $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ sastavljene od $1 < k \leq m$ klastera konstruiramo particiju $\Pi^{(r)}$ s $r < k$ klastera tako da barem dva klastera particije $\Pi^{(k)}$ spojimo u jedan. Jedna mogućnost u slučaju manjeg broja podataka je pokrenuti algoritam od samog skupa $\mathcal{A}$. U tom smislu uvodimo sljedeću definiciju.

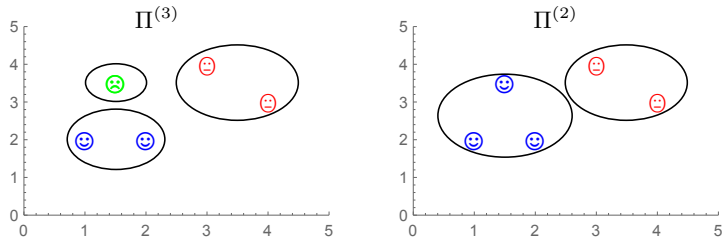
Definicija 5.1. Kažemo da je particija $\Pi^{(k)}$ ugnježđena (*nested*) u particiju $\Pi^{(r)}$ i pišemo $\Pi^{(k)} \sqsubset \Pi^{(r)}$ ako

(i) $r < k$;

(ii) svaki klaster iz $\Pi^{(k)}$ podskup je nekog klastera iz $\Pi^{(r)}$.

Na Slici 5.1 prikazana je particija $\Pi^{(3)}$ koja je ugnježđena u particiju $\Pi^{(2)}$.

U ovom udžbeniku razmatrat ćemo samo aglomerativne hijerarhijske algoritme koji u svakom koraku povezuju najviše po dva klastera promatrane

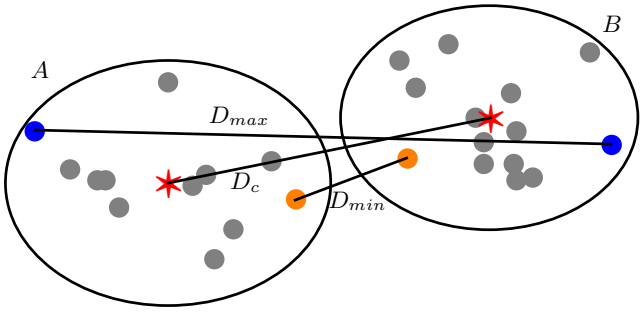


Slika 5.1: Particija $\Pi^{(3)}$ je ugnježđena u particiju $\Pi^{(2)}$ ($\Pi^{(3)} \sqsubset \Pi^{(2)}$)

k-particije $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$. Ta dva klastera odabrat ćemo tako da razmotrimo sve moguće parove klastera. Ukupan broj ovih parova jednak je broju svih kombinacija bez ponavljanja od k elemenata drugog razreda

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}.$$

Ako uvedemo neku kvazimetričku funkciju $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ (kao u t. 2, str. 3), onda kao mjeru sličnosti/različitosti dva klastera A, B možemo promatrati različite mjere njihove međusobne udaljenosti [31, 72, 73]:



Slika 5.2: Različite mjere udaljenosti klastera A i B

$$D_c(A,B) = d(c_A, c_B), \quad \text{[udaljenost centara } c_A, c_B \text{ klastera]} \tag{5.1}$$

$$D_{min}(A,B) = \min_{a\in A, b\in B} d(a,b) \quad \text{[minimalna udaljenost]} \tag{5.2}$$

$$D_{max}(A,B) = \max_{a\in A, b\in B} d(a,b) \quad \text{[maksimalna udaljenost]} \tag{5.3}$$

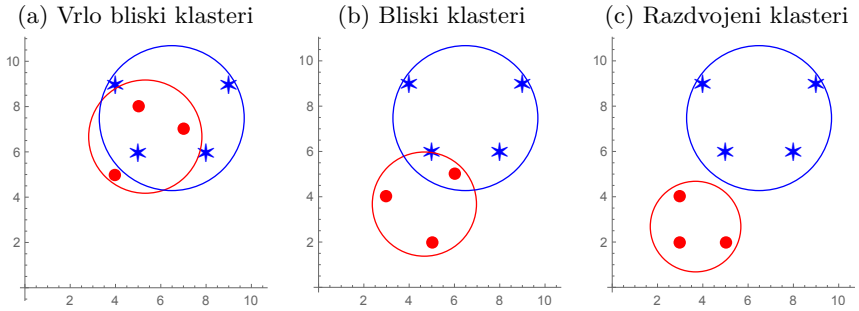
$$D_{avg}(A,B) = \frac{1}{|A||B|} \sum_{a\in A} \sum_{b\in B} d(a,b) \quad \text{[prosječna udaljenost]} \tag{5.4}$$

$$\texttt{HD}(A,B) = \max\{\max_{a\in A}\min_{b\in B} d(a,b), \max_{b\in B}\min_{a\in A} d(a,b)\}$$

[Hausdorffova udaljenost] (5.5)

Zadatak 5.1. Kolika je Hausdorffova udaljenost klastera A i B na Slici 5.2?

Primjer 5.1. *Na Slici 5.3 prikazan je klaster A (3 crvene točkice) i klaster B (4 plave zvjezdice) u različitim međusobnim položajima, a u Tablici 5.1 navedene su odgovarajuće D_c, D_{min} i Hausdorffove* HD *udaljenosti za LS-kvazimetričku i ℓ_1-metričku funkciju za navedena tri slučaja.*



Slika 5.3: Klaster A (crvene točke) i klaster B (plave točke)

	Vrlo bliski	Bliski	Razdvojeni
D_c uz ℓ_1-metričku funkciju	2	5	9
D_c uz LS-kvazimetričku funkciju	2.05	18	31.4
D_{min} uz ℓ_1-metričku funkciju	2	2	4
D_{min} uz LS-kvazimetričku funkciju	2	2	8
HD uz ℓ_1-metričku funkciju	4	7	11
HD uz LS-kvazimetričku funkciju	8	25	61

Tablica 5.1: Različite mjere udaljenosti dva klastera

U nastavku teksta detaljnije ćemo razmotriti primjenu mjere sličnosti (udaljenosti) klastera definirane s (5.1) koristeći ranije spomenute kvazimetričke funkcije.

Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1,\ldots,m\}$ zadani skup. Aglomeracijski hijerarhijski algoritam može započeti od neke particije $\Pi^{(\mu)}$ koja sadržava μ klastera ($1 < \mu \leq m$) i završiti s particijom $\Pi^{(k)}$ koja sadržava k klastera

$(1 \leq k < \mu \leq m)$. Mi ćemo algoritam najčešće pokretati od particije s m klastera

$$\Pi^{(m)} = \{\{a^1\}, \{a^2\}, \ldots, \{a^m\}\},$$

tj. od particije u kojoj je svaki element skupa $\mathcal{A}$ za sebe poseban klaster. Udaljenost $D(A, B)$ između klastera A i B definirat ćemo kao udaljenost njihovih centara

$$c_A = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in A} d(x, a), \qquad c_B = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{b \in B} d(x, b),$$

gdje će d biti LS-kvazimetrička funkcija ili ℓ_1-metrička funkcija.

Primijetite da je vrijednost kriterijske funkcije cilja $\mathcal{F}$ na particiji $\Pi^{(m)}$ jednaka nuli. U prvom koraku biramo dva najsličnija klastera, tj. dva najbliža elementa skupa $\mathcal{A}$. Njih ćemo spojiti u jedan klaster. Primijetite da se na taj način povećava vrijednost kriterijske funkcije cilja $\mathcal{F}$.

Algoritam se može zaustaviti na particiji s unaprijed zadanim brojem klastera, a moguće je razmotriti i problem određivanja particije s najprikladnijim brojem klastera [30, 78], što ćemo detaljnije razmatrati u Poglavlju 6.

Algoritam 1 (Agglomerative Nesting (AGNES))

Input: m, $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$, $1 < k < m$, $\mu = 0$;

1: Definirati početnu particiju $\Pi^{(m)} = \{\pi_1, \ldots, \pi_m\}$, $\pi_j = \{a^j\}$;
2: Za particiju $\Pi^{(m-\mu)}$ konstruirati matricu sličnosti $R_{m-\mu} \in \mathbb{R}^{(m-\mu)\times(m-\mu)}$, $r_{ij} = D(\pi_i, \pi_j)$;
3: Riješiti optimizacijski problem $\{i_0, j_0\} \subseteq \operatorname*{argmin}_{1<i<j\leq m} r_{i,j}$;
4: Konstruirati novu particiju $\Pi^{(m-\mu-1)} = (\Pi^{(m-\mu)} \setminus \{\pi_{i_0}, \pi_{j_0}\}) \bigcup \{\pi_{i_0} \bigcup \pi_{j_0}\}$;
5: **if** $\mu < m - k$, **then**
6: $\quad \mu := \mu + 1$ i prijeći na Korak 2;
7: **else**
8: $\quad$ STOP;
9: **end if**

Output: $\{\Pi^{(k)}\}$.

U `Koraku 3` traži se pozicija $\{i_0, j_0\}$ najmanjeg elementa matrice razlika $R_{m-\mu}$. U `Koraku 4` konstruira se nova particija tako da se klasteri π_{i_0}, π_{j_0} spoje u jedan klaster. U `Koraku 5` provjerava se kriterij zaustavljanja algoritma.

Koristan ilustrativni prikaz Algoritma 1 može se napraviti pomoću tzv. dendrograma, koji ilustrira svaki korak algoritma i daje razinu sličnosti. Algogitam ćemo ilustrirati na sljedećem jednostavnom primjeru.

Primjer 5.2. *Za skup* $\mathcal{A} = \{(0,0),(1,5),(3,5),(5,5),(5,3),(5,1)\}$*, polazeći od particije* $\Pi^{(6)} = \{\{(0,0)\},\{(1,5)\},\{(3,5)\},\{(5,5)\},\{(5,3)\},\{(5,1)\}\}$ *pokrenut ćemo* ***Algoritam AGNES*** *s mjerom sličnosti* (5.1) *generiranom* ℓ_1*-metričkom funkcijom (vidi Sliku 5.4a)*

$$D_1(A,B) = \|c_A - c_B\|_1, \quad c_A = \operatorname*{med}_{a\in A} a \quad c_B = \operatorname*{med}_{b\in B} b$$

Odredimo najprije matricu sličnosti $R_6 \in \mathbb{R}^{6\times 6}$ s elementima $r_{ij} = D_1(\{a^i\},\{a^j\})$ zadanu s

$$R_6 = \begin{bmatrix} 0 & 6 & 8 & 10 & 8 & 6 \\ 6 & 0 & \boxed{2} & 4 & 6 & 8 \\ 8 & 2 & 0 & 2 & 4 & 6 \\ 10 & 4 & 2 & 0 & 2 & 4 \\ 8 & 6 & 4 & 2 & 0 & 2 \\ 6 & 8 & 6 & 4 & 2 & 0 \end{bmatrix}$$

Minimalni element gornjeg trokuta matrice R_6 postiže se na četiri mjesta. Izaberimo prvo od njih: to je element $r_{2,3} = 2$. To znači da ćemo spojiti klastere $\{a^2\}$ i $\{a^3\}$ (tj. elemente a^2 i a^3) u novi klaster $\{a^2, a^3\}$ s centrom u točki $(2,5)$ (vidi Sliku 5.4b). Tako dobivamo novu particiju $\Pi^{(5)}$ s novim centrima

$\Pi^{(5)}$	$\{(0,0)\}$	$\{(1,5),(3,5)\}$	$\{(5,5)\}$	$\{(5,3)\}$	$\{(5,1)\}$
Centri	$(0,0)$	$(2,5)$	$(5,5)$	$(5,3)$	$(5,1)$

.

Vrijednost funkcije cilja postaje $\mathcal{F}_1(\Pi^{(5)}) = 2$.

U sljedećem koraku algoritma određujemo matricu sličnosti $R_5 \in \mathbb{R}^{5\times 5}$ s elementima $r_{ij} = D(\pi_i^{(5)}, \pi_j^{(5)})$ zadanu s

$$R_5 = \begin{bmatrix} 0 & 7 & 10 & 8 & 6 \\ 7 & 0 & 3 & 5 & 7 \\ 10 & 3 & 0 & \boxed{2} & 4 \\ 8 & 5 & 2 & 0 & 2 \\ 6 & 7 & 4 & 2 & 0 \end{bmatrix}$$

Minimalni element gornjeg trokuta matrice R_5 postiže se na dva mjesta. Izaberimo prvo od njih: to je element $r_{3,4} = 2$. To znači da ćemo spojiti treći i četvrti klaster particije $\Pi^{(5)}$, tj. spojit ćemo klastere $\{(5,5)\}$ i $\{(5,3)\}$ u novi klaster $\{(5,5),(5,3)\}$ s centrom u točki $(5,4)$ (vidi Sliku 5.4c). Tako dobivamo novu particiju $\Pi^{(4)}$ s novim centrima

$\Pi^{(4)}$	$\{(0,0)\}$	$\{(1,5),(3,5)\}$	$\{(5,5),(5,3)\}$	$\{(5,1)\}$
Centri	$(0,0)$	$(2,5)$	$(5,4)$	$(5,1)$

Vrijednost funkcije cilja postaje $\mathcal{F}_1(\Pi^{(4)}) = 4$.

U sljedećem koraku algoritma određujemo matricu sličnosti $R_4 \in \mathbb{R}^{4\times 4}$ s elementima $r_{ij} = D(\pi_i^{(4)}, \pi_j^{(4)})$ zadanu s

$$R_4 = \begin{bmatrix} 0 & 7 & 9 & 6 \\ 7 & 0 & 4 & 7 \\ 9 & 4 & 0 & \boxed{3} \\ 6 & 7 & 3 & 0 \end{bmatrix}$$

Minimalni element gornjeg trokuta matrice R_4 postiže se na elementu $r_{3,4} = 3$. To znači da ćemo spojiti treći i četvrti klaster particije $\Pi^{(4)}$, tj. spojit ćemo klastere $\{(5,5),(5,3)\}$ i $\{(5,1)\}$ u novi klaster $\{(5,5),(5,3),(5,1)\}$ s centrom u točki $(5,3)$ (vidi Sliku 5.4d). Tako dobivamo novu particiju $\Pi^{(3)}$ s novim centrima

$\Pi^{(3)}$	$\{(0,0)\}$	$\{(1,5),(3,5)\}$	$\{(5,5),(5,3),(5,1)\}$
Centri	$(0,0)$	$(2,5)$	$(5,3)$

Vrijednost funkcije cilja postaje $\mathcal{F}_1(\Pi^{(3)}) = 6$.

U sljedećem koraku algoritma određujemo matricu sličnosti $R_3 \in \mathbb{R}^{3\times 3}$ s elementima $r_{ij} = D(\pi_i^{(3)}, \pi_j^{(3)})$ zadanu s

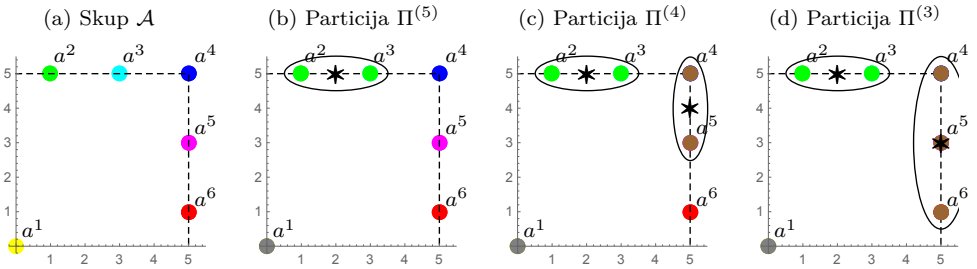
$$R_3 = \begin{bmatrix} 0 & 7 & 8 \\ 7 & 0 & \boxed{5} \\ 8 & 5 & 0 \end{bmatrix}$$

Minimalni element gornjeg trokuta matrice R_3 postiže se na elementu $r_{2,3} = 5$. To znači da ćemo spojiti drugi i treći klaster particije $\Pi^{(3)}$, tj. spojit ćemo klastere $\{(1,5),(3,5)\}$ i $\{(5,5),(5,3),(5,1)\}$ u novi klaster $\{(1,5),(3,5),(5,5),(5,3),(5,1)\}$ s centrom u točki $(5,5)$. Tako dobivamo novu particiju $\Pi^{(2)}$ s novim centrima (vidi Sliku 5.5)

$\Pi^{(2)}$	$\{(0,0)\}$	$\{(1,5),(3,5),(5,5),(5,3),(5,1)\}$
Centri	$(0,0)$	$(5,5)$

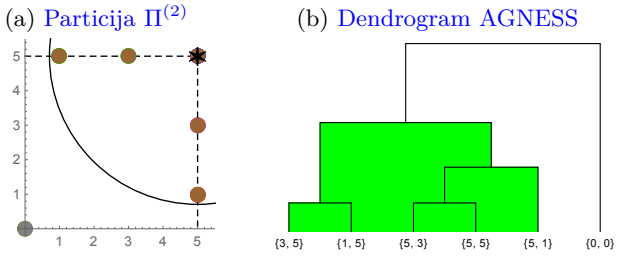
.

Vrijednost funkcije cilja postaje $\mathcal{F}_1(\Pi^{(2)}) = 10$.



Slika 5.4: Aglomeracijski hijerarhijski algoritam za Primjer 5.2

Daljnje spajanje rezultiralo bi dobivanjem particije s jednim klasterom – čitavim skupom $\mathcal{A}$. Na Slici 5.5b prikazan je dendrogram koji grafički ilustrira Algoritam 1 za dobivanje particije $\Pi^{(2)}$ s dva kompaktna dobro razdvojena klastera. Denrogram pokazuje redoslijed povezivanja klastera.



Slika 5.5: Particija $\Pi^{(2)}$ i dendrogram za Primjer 5.2

Primjedba 5.1. U svakom koraku Algoritma 1 moramo međusobno usporediti svih $m - \mu$ klastera, dakle $\binom{m-\mu}{2}$ parova klastera. Dakle, ako bismo proveli cijeli Algoritam 1 do osnovnog skupa $\mathcal{A}$, bilo bi potrebno

$$\sum_{\mu=0}^{m-1} \binom{m-\mu}{2} = \sum_{\mu=1}^{m} \binom{\mu}{2} = \frac{(m-1)m(m+1)}{6}$$

usporedbi. Budući da je u ovom izrazu dominantna potencija m^3, uobičajeno je pisati da je složenost ovog algoritma $\mathcal{O}(m^3)$ (vidi primjerice [40]).

Primjedba 5.2. Druga klasa hijerarhijskih algoritama su tzv. Algoritmi dijeljenja (*Divisive Algorithms*). Postupak se sastoji u tome da se osnovni skup $\mathcal{A}$ najprije podijeli u dva klastera tako da se pri tome postigne maksimalno sniženje vrijednosti funkcije cilja. Nadalje, svaki od klastera nastavlja se dalje dijeliti istim postupkom.

5.2 Primjena principa najmanjih kvadrata

Kao što smo ranije spomenuli, mjeru sličnosti dva klastera možemo definirati pomoću poznatih kvazimetričkih funkcija. U ovom poglavlju analizirat ćemo primjenu LS-kvazimetričke funkcije, a sličnost (udaljenost) dva klastera definirat ćemo pomoću udaljenosti njihovih centroida (5.1) ili pomoću minimalnih udaljenosti njihovih elemenata (5.2).

5.2.1 Sličnost definirana pomoću udaljenosti centroida

Sukladno (5.1), udaljenost dva skupa A i B možemo definirati pomoću udaljenosti njihovih centroida c_A, c_B

$$D_{LS}(A,B) = d_{LS}(c_A, c_B) = \|c_A - c_B\|_2^2. \tag{5.6}$$

Sljedeća lema daje važan identitet koji povezuje elemente skupa $\mathcal{A}$ s njegovim centroidom.

Lema 5.1. *Ako je* $A = \{a^i \in \mathbb{R}^n : i = 1, \ldots, p\}$ *skup, a* $c_A = \frac{1}{p}\sum_{i=1}^{p} a^i$ *njegov centroid, tada vrijedi*

$$\sum_{i=1}^{p}(a^i - c_A) = 0, \tag{5.7}$$

$$\sum_{i=1}^{p}\|a^i - x\|_2^2 = \sum_{i=1}^{p}\|a^i - c_A\|_2^2 + p\|x - c_A\|_2^2, \quad \forall x \in \mathbb{R}^n. \tag{5.8}$$

Dokaz. Jednakost (5.7) neposredno slijedi iz

$$\sum_{i=1}^{p}(a^i - c_A) = \frac{p}{p}\sum_{i=1}^{p} a^i - pc_A = pc_A - pc_A = 0.$$

U svrhu dokaza jednakosti (5.8) najprije primijetimo da zbog (5.7) vrijedi

$$\sum_{i=1}^{p}\langle a^i - c_A, c_A - x\rangle = \Big\langle \sum_{i=1}^{p}(a^i - c_A), c_A - x\Big\rangle = 0,$$

gdje je $\langle\cdot\rangle$ uobičajeni sklarni produkt. Zato vrijedi:

$$\begin{aligned}\sum_{i=1}^{p}\|a^i - x\|_2^2 &= \sum_{i=1}^{p}\|(a^i - c_A) + (c_A - x)\|_2^2 \\ &= \sum_{i=1}^{p}\|a^i - c_A\|_2^2 + 2\sum_{i=1}^{p}\langle a^i - c_A, c_A - x\rangle + \sum_{i=1}^{p}\|c_A - x\|_2^2 \\ &= \sum_{i=1}^{p}\|a^i - c_A\|_2^2 + p\|c_A - x\|_2^2. \end{aligned}$$

□

Sljedeći teorem pokazuje vezu između vrijednosti funkcije cilja $\mathcal{F}_{LS}$ primijenjenu na dva klastera s njenom vrijednosti na uniji ta dva klastera.

Teorem 5.1. *Ako je skup* $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ *sastavljen od dva disjunktna klastera* $\mathcal{A} = A \cup B$,

$$A = \{a^1, \ldots, a^p\}, \quad |A| = p, \quad c_A = \frac{1}{p}\sum_{i=1}^{p} a^i;$$

$$B = \{b^1, \ldots, b^q\}, \quad |B| = q, \quad c_B = \frac{1}{q}\sum_{j=1}^{q} b^j,$$

tada je centroid skupa $\mathcal{A} = A \cup B$ *zadan s*

$$c = \frac{p}{p+q}c_A + \frac{q}{p+q}c_B, \tag{5.9}$$

i vrijedi

$$\mathcal{F}_{LS}(A \cup B) = \mathcal{F}_{LS}(A) + \mathcal{F}_{LS}(B) + p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2, \tag{5.10}$$

gdje je $\mathcal{F}_{LS}$, *LS-funkcija cilja (vidi t. 2, str. 3).*

Dokaz. Jednakost (5.9) neposredno slijedi iz

$$c = c(A \cup B) = \frac{1}{p+q}\Big(\sum_{i=1}^{p} a^i + \sum_{j=1}^{q} b^j\Big) = \frac{p}{p+q}\frac{1}{p}\sum_{i=1}^{p} a^i + \frac{q}{p+q}\frac{1}{q}\sum_{j=1}^{q} b^j.$$

Korištenjem Leme 5.1 dobivamo

$$\begin{aligned}\mathcal{F}_{LS}(A \cup B) =& \sum_{i=1}^{p}\|a^i - c\|_2^2 + \sum_{j=1}^{q}\|b^j - c\|_2^2 \\ =& \sum_{i=1}^{p}\|a^i - c_A\|_2^2 + p\|c - c_A\|_2^2 + \sum_{j=1}^{q}\|b^j - c_B\|_2^2 + q\|c - c_B\|_2^2 \\ =& \mathcal{F}_{LS}(A) + \mathcal{F}_{LS}(B) + p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2.\end{aligned}$$

□

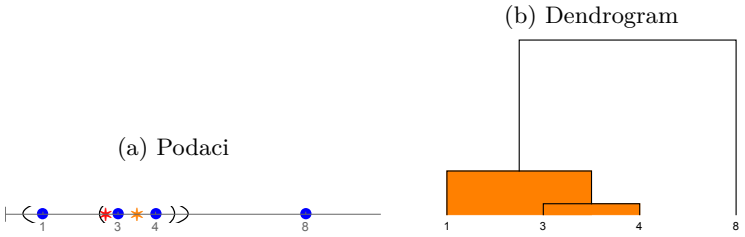
Primjer 5.3. *Zadan je skup* $\mathcal{A} = \{1, 3, 4, 8\}$ *prikazan na Slici 5.6a. Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.1) *i LS-kvazimetričke funkcije* $d_{LS}(x, y) = (x - y)^2$. *Primijenit ćemo tvrdnje dokazane u Teoremu 5.1.*

Krenimo od početne particije $\Pi^{(4)} = \{\{1\},\{3\},\{4\},\{8\}\}$ za koju je $\mathcal{F}_{LS}(\Pi^{(4)}) = 0$. Minimalni element matrice sličnosti $R_4 \in \mathbb{R}^{4\times 4}$ je $r_{2,3} = 1$ i on pokazuje da najprije treba spojiti drugi i treći klaster: $\{3\}$ i $\{4\}$. Tako dobivamo particiju $\Pi^{(3)} = \{\{1\},\{3,4\},\{8\}\}$ s centroidima $c_i \in \{1, 3.5, 8\}$ i vrijednosti funkcije cilja $\mathcal{F}_{LS}(\Pi^{(3)}) = 0.5$.

$$R_4 = \begin{bmatrix} 0 & 4 & 9 & 49 \\ 4 & 0 & \boxed{1} & 25 \\ 9 & 1 & 0 & 16 \\ 49 & 25 & 16 & 0 \end{bmatrix}, \qquad R_3 = \begin{bmatrix} 0 & \boxed{6.25} & 49 \\ 6.25 & 0 & 20.25 \\ 49 & 20.25 & 0 \end{bmatrix}.$$

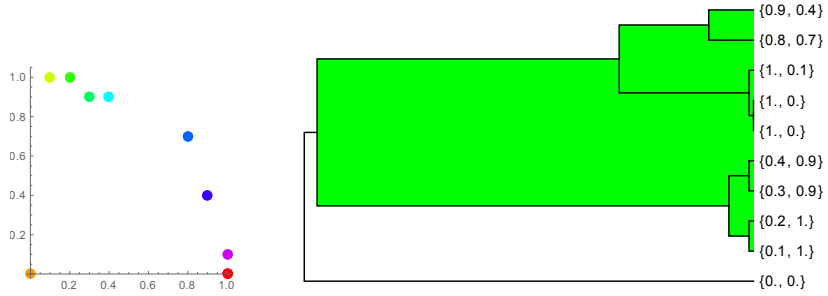
Minimalni element matrice sličnosti $R_3 \in \mathbb{R}^{3\times 3}$ je $r_{1,2} = 6.25$ i on pokazuje da treba spojiti prvi i drugi klaster: $\{1\}$ i $\{3,4\}$ particije $\Pi^{(3)}$. Tako dobivamo particiju $\Pi^{(2)} = \{\{1,3,4\},\{8\}\}$ s centroidima $c_i \in \{8/3, 8\}$ i vrijednosti funkcije cilja $\mathcal{F}_{LS}(\Pi^{(2)}) = \frac{14}{3} = 4.667$.
Rezultate možemo provjeriti primjenom programskog sustava *Mathematica*. Odgovarajući dendrogram prikazan je na Slici 5.6b.

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A = {1, 3, 4, 8};
        Agglomerate[A, Linkage -> "Centroid"]
        DendrogramPlot[A, Linkage -> "Centroid", HighlightLevel -> 2]
```



Slika 5.6: Primjena Algoritma 1 na skup $\mathcal{A} = \{1,3,4,8\}$ korištenjem sličnosti definirane s (5.1) i LS-kvazimetričke funkcije

Primjer 5.4. *Skup $\mathcal{A}$ sastavljen je od ishodišta i još 9 slučajnih točaka u blizini trigonometrijske kružnice u prvom kvadrantu (vidi Sliku 5.7a). Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.1) *i LS-kvazimetričke funkcije $d_{LS}(x,y) = \|x-y\|_2^2$. Rezultat povezivanja vidljiv je na dendrogramu prikazanom na Slici 5.7b.*



Slika 5.7: Podaci oko trigonometrijske kružnice i odgovarajući dendrogram Algoritma 1 dobiven na bazi sličnosti (5.1) i LS-kvazimetričke funkcije

Izraz $\Delta := p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2$ iz (5.10) može se pojednostaviti kao u sljedećem korolaru.

Korolar 5.1. *Ako je skup* $\mathcal{A} = \{a^i \in \mathbb{R}^n \colon i = 1, \ldots, m\}$ *s centroidom* $c = \frac{1}{m}\sum_{i=1}^{m} a^i$ *unija od dva klastera* $\mathcal{A} = A \cup B$,

$$A = \{a^1, \ldots, a^p\}, \quad |A| = p, \quad c_A = \frac{1}{p}\sum_{i=1}^{p} a^i;$$
$$B = \{b^1, \ldots, b^q\}, \quad |B| = q, \quad c_B = \frac{1}{q}\sum_{j=1}^{q} b^j,$$

tada je

$$\Delta := p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2 = \frac{pq}{p+q}\|c_A - c_B\|_2^2. \tag{5.11}$$

Dokaz. Kako je prema (5.9)

$$p\|c_A - c\|_2^2 = p\|\tfrac{p}{p+q}c_A + \tfrac{q}{p+q}c_B - c_A\|_2^2 = \tfrac{pq^2}{(p+q)^2}\|c_A - c_B\|_2^2,$$
$$q\|c_B - c\|_2^2 = q\|\tfrac{p}{p+q}c_A + \tfrac{q}{p+q}c_B - c_B\|_2^2 = \tfrac{p^2q}{(p+q)^2}\|c_A - c_B\|_2^2,$$

vrijedi

$$\Delta = \tfrac{pq^2}{(p+q)^2}\|c_A - c_B\|_2^2 + \tfrac{p^2q}{(p+q)^2}\|c_A - c_B\|_2^2,$$

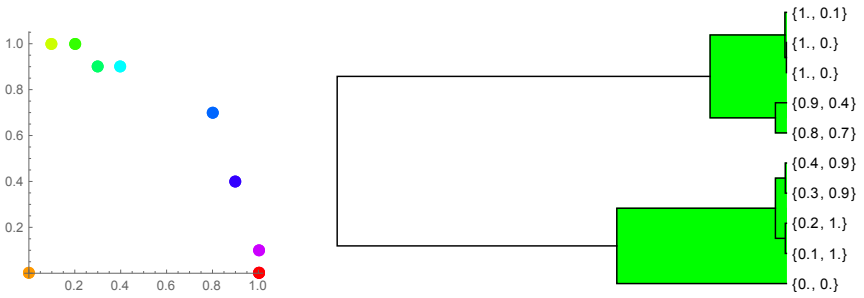
iz čega neposredno slijedi (5.11). □

Zbog toga se kao mjera sličnosti dva klastera A i B umjesto (5.6) može koristiti tzv. Wardova udaljenost[1]

$$D_W(A,B) = \frac{|A||B|}{|A|+|B|}\|c_A - c_B\|_2^2, \tag{5.12}$$

gdje je $|A|$ broj elemenata klastera A, a $|B|$ broj elemenata klastera B. Kao što se može vidjeti iz Teorema 5.1 i Korolara 5.1, ako kao mjeru sličnosti dva klastera koristimo Wardovu udaljenost (5.12), onda će spajanjem klastera A i B vrijednost funkcije cilja porasti upravo za $D_W(A,B)$. Može se pokazati [73] da ovaj izbor osigurava najmanji mogući porast funkcije cilja $\mathcal{F}_{LS}$.

Primjer 5.5. *Primjer 5.4 izradit ćemo primjenom Wardove udaljenosti. Rezultat povezivanja vidljiv je na dendrogramu prikazanom na Slici 5.8.*



Slika 5.8: Podaci oko trigonometrijske kružnice i odgovarajući dendrogram Algoritma 1 dobiven primjenom Wardove udaljenosti

Zadatak 5.2. Primjer 5.3 izradite primjenom Wardove udaljenosti. Rezultate možemo provjeriti primjenom programskog sustava *Mathematica*.

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A = {1, 3, 4, 8};
        Agglomerate[A, Linkage -> "Ward"]
```

5.2.2 Sličnost definirana pomoću minimalnih udaljenosti

Već na ovako jednostavnim primjerima vidi se da primjena formule (5.1) za određivanje sličnosti dva klastera ima značajnu računsku složenost. Sličan rezultat, ali znatno jednostavnije, možemo dobiti primjenom formule

[1] J.H.Ward, Jr., *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, **58**(1963), 236–244.

sličnosti (5.2)

$$D_{min}(A,B) = \min_{a\in A, b\in B} d(a,b), \tag{5.13}$$

gdje je d također LS-kvazimetrička ili Wardova funkcija udaljenosti.

Primjer 5.6. *Zadan je skup* $\mathcal{A} = \{1,3,4,8\}$ *iz Primjera 5.3. Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.13) *i LS-kvazimetričke funkcije* $d_{LS}(x,y) = (x-y)^2$.

Krenimo od početne particije particije $\Pi^{(4)} = \{\{1\},\{3\},\{4\},\{8\}\}$. Minimalni element matrice sličnosti $R_4 \in \mathbb{R}^{4\times 4}$ je $r_{2,3} = 1$ i on pokazuje da najprije treba spojiti drugi i treći klaster: $\{3\}$ i $\{4\}$ particije $\Pi^{(4)}$. Tako dobivamo particiju $\Pi^{(3)} = \{\{1\},\{3,4\},\{8\}\}$. Primijetite da ovdje više ne treba računati centroide.

Minimalni element matrice sličnosti $R_3 \in \mathbb{R}^{3\times 3}$ particije $\Pi^{(3)}$ je $r_{1,2} = 4$ i on pokazuje da treba spojiti prvi i drugi klaster: $\{1\}$ i $\{3,4\}$ particije $\Pi^{(3)}$. Tako dobivamo particiju $\Pi^{(2)} = \{\{1,3,4\},\{8\}\}$ koja se podudara s onom dobivenom u Primjeru 5.3, gdje smo koristili sličnost definiranu s (5.1).

$$R_4 = \begin{bmatrix} 0 & 2^2 & 3^2 & 7^2 \\ - & 0 & \boxed{1^2} & 5^2 \\ - & - & 0 & 4^2 \\ - & - & - & 0 \end{bmatrix}, \qquad R_3 = \begin{bmatrix} 0 & \boxed{2^2} & 7^2 \\ - & 0 & 4^2 \\ - & - & 0 \end{bmatrix}.$$

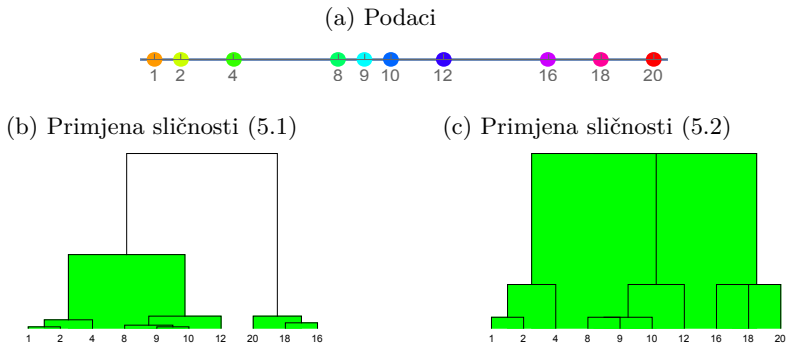
Rezultate možemo provjeriti primjenom programskog sustava *Mathematica*

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A = {1, 3, 4, 8};
        Agglomerate[A, Linkage -> "Single"]
```

Primjer 5.7. *Zadan je skup* $\mathcal{A} = \{1,2,4,8,9,10,12,15,18,20\}$ *(vidi Sliku 5.9a). Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.1) *i* (5.2) *i LS-kvazimetričke funkcije* $d_{LS}(x,y) = (x-y)^2$.

Algoritam ćemo provesti primjenom niže navedenog *Mathematica*-programa.

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A = {1, 2, 4, 8, 9, 10, 12, 16, 18, 20};
        DendrogramPlot[A, Linkage -> "Centroid", HighlightLevel -> 2,
                          LeafLabels -> (# &)]
        DendrogramPlot[A, Linkage -> "Single", HighlightLevel -> 2,
                        LeafLabels -> (# &)]
```

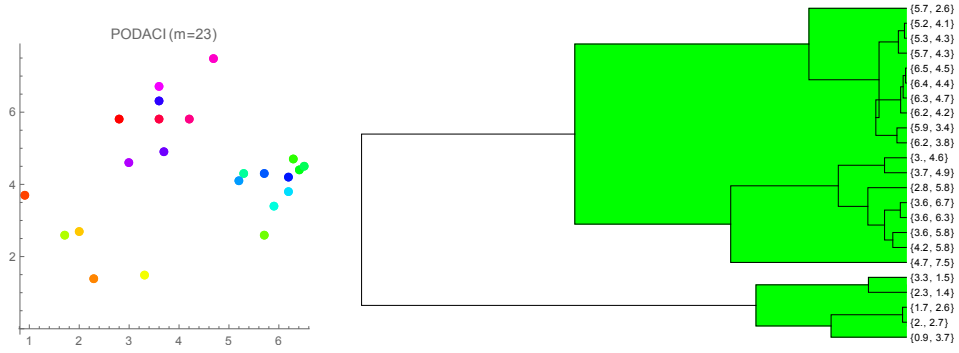


Slika 5.9: Djelovanje Algoritma 1 primjenom LS-kvazimetričke funkcije

Primjer 5.8. *Zadan je skup* $\mathcal{A} \subset \mathbb{R}^2$ *prikazan na Slici 5.10. Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.1)*, pri čemu koristimo LS-kvazimetričku funkciju* $d_{LS}(x, y) = (x - y)^2$.

Na dendrogramu dobivenom niže navedenim *Mathematica*- programom jasno se uočavaju tri kompaktna dobro razdvojena klastera.

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= DendrogramPlot[A, Linkage -> "Centroid", HighlightLevel -> 2,
                       LeafLabels -> (# &), Orientation -> Left]
```



Slika 5.10: Djelovanje Algoritma 1 primjenom LS-kvazimetričke funkcije i sličnosti definirane s (5.1)

5.3 Primjena principa najmanjih apsolutnih odstupanja

U ovom poglavlju analizirat ćemo primjenu ℓ_1-metričke funkcije. Sličnost dva klastera definirat ćemo ponovo pomoću udaljenosti njihovih centara (5.1) ili pomoću minimalne udaljenosti njihovih elemenata (5.2).

5.3.1 Sličnost definirana pomoću udaljenosti centara

Neka su A i B dva klastera neke particije Π skupa $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$. Njihovi centri zadani su s

$$c_A = \text{med}(A), \qquad c_B = \text{med}(B),$$

a njihova međusobna udaljenost zadana je pomoću ℓ_1-udaljenosti njihovih centara (5.1) i ℓ_1-metričke funkcije

$$D_1(A, B) = d_1(c_A, c_B) = \|c_A - c_B\|_1. \tag{5.14}$$

Već ranije u Primjeru 5.2 na str. 95, primijenili smo ℓ_1-metričku funkciju i princip povezivanja klastera zadan s (5.1), tj. (5.14). U sljedećem primjeru princip ćemo ilustrirati na jednostavnom primjeru skupa podataka s jednim obilježjem ($n = 1$).

Primjer 5.9. *Ponovo promatramo skup $\mathcal{A} = \{1, 3, 4, 8\}$ iz Primjera 5.3, str. 99. Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.14) *i ℓ_1-metričke funkcije.*

Krenimo od početne particije $\Pi^{(4)} = \{\{1\}, \{3\}, \{4\}, \{8\}\}$ za koju je $\mathcal{F}_1(\Pi^{(4)}) = 0$. Minimalni element matrice sličnosti $R_4 \in \mathbb{R}^{4\times 4}$ je $r_{2,3} = 1$ i on pokazuje da najprije treba spojiti drugi i treći klaster: $\{3\}$ i $\{4\}$. Tako dobivamo particiju $\Pi^{(3)} = \{\{1\}, \{3, 4\}, \{8\}\}$ s centrima $c_i \in \{1, 3.5, 8\}$ i vrijednosti funkcije cilja $\mathcal{F}_1(\Pi^{(3)}) = 1$.

$$R_4 = \begin{bmatrix} 0 & 2 & 3 & 7 \\ 2 & 0 & \boxed{1} & 5 \\ 3 & 1 & 0 & 4 \\ 7 & 5 & 4 & 0 \end{bmatrix}, \qquad R_3 = \begin{bmatrix} 0 & \boxed{2.5} & 7 \\ 2.5 & 0 & 4.5 \\ 7 & 4.5 & 0 \end{bmatrix}.$$

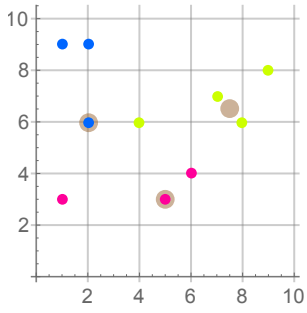
Minimalni element matrice sličnosti $R_3 \in \mathbb{R}^{3\times 3}$ je $r_{1,2} = 2.5$ i on pokazuje da treba spojiti prvi i drugi klaster: $\{1\}$ i $\{3, 4\}$ particije $\Pi^{(3)}$. Tako dobivamo particiju $\Pi^{(2)} = \{\{1, 3, 4\}, \{8\}\}$ s centrima $c_i \in \{3, 8\}$ i vrijednosti funkcije cilja $\mathcal{F}_1(\Pi^{(2)}) = 3$.

Rezultate možemo provjeriti primjenom programskog sustava *Mathematica*.

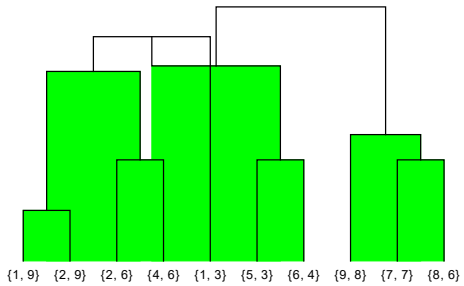
```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A = {1, 3, 4, 8};
        Agglomerate[A, DistanceFunction -> ManhattanDistance,
                       Linkage -> "Median"]
```

Primjer 5.10. *Primijenimo Algoritam 1 na podacima iz Primjera 4.19, str. 81 (vidi također Sliku 5.11a), gdje je također primijenjena ℓ_1-metrička funkcija.*

(a) Skup iz Primjera 4.19



(b) Dendrogram



Slika 5.11: Djelovanje Algoritma 1 primjenom ℓ_1-metričke funkcije i sličnosti definirane s (5.14)

Primjenom niže navedenog programa dobivamo dendrogram prikazan na Slici 5.11b. Usporedite dobivene rezultate s rezultatima iz Primjera 4.19.

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A={{1,9},{2,9},{2,6},{1,3},{5,3},{6,4},{4,6},{7,7},{8,6},{9,8}};
        DendrogramPlot[A, Linkage -> "Median",
        DistanceFunction -> ManhattanDistance, HighlightLevel -> 3,
        LeafLabels -> (# &)]
```

5.3.2 Sličnost definirana pomoću minimalne udaljenosti

Sličnost dva klastera definirat ćemo pomoću minimalne udaljenosti njihovih elemenata (5.2) u ℓ_1-metrici

$$D_1(A, B) = \min_{a \in A,\, b \in B} \|a - b\|_1. \tag{5.15}$$

Primjer 5.11. *Ponovo promatramo skup $\mathcal{A} = \{1, 3, 4, 8\}$ iz Primjera 5.3. Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.15).

Krenimo od početne particije $\Pi^{(4)} = \{\{1\},\{3\},\{4\},\{8\}\}$. Minimalni element matrice sličnosti $R_4 \in \mathbb{R}^{4\times 4}$ je $r_{2,3} = 1$ i on pokazuje da najprije treba spojiti drugi i treći klaster: $\{3\}$ i $\{4\}$ particije $\Pi^{(4)}$. Tako dobivamo particiju $\Pi^{(3)} = \{\{1\},\{3,4\},\{8\}\}$. Primijetite da u ovom slučaju više ne treba računati centre klastera.

Minimalni element matrice sličnosti $R_3 \in \mathbb{R}^{3\times 3}$ particije $\Pi^{(3)}$ je $r_{1,2} = 2$ i on pokazuje da treba spojiti prvi i drugi klaster: $\{1\}$ i $\{3,4\}$ particije $\Pi^{(3)}$. Tako dobivamo particiju $\Pi^{(2)} = \{\{1,3,4\},\{8\}\}$ koja se podudara s onom dobivenom u Primjeru 5.3 gdje smo koristili sličnost definiranu s (5.1).

$$R_4 = \begin{bmatrix} 0 & 2 & 3 & 7 \\ 2 & 0 & \boxed{1} & 5 \\ 3 & 1 & 0 & 4 \\ 7 & 5 & 4 & 0 \end{bmatrix}, \qquad R_3 = \begin{bmatrix} 0 & \boxed{2} & 7 \\ 2 & 0 & 4 \\ 7 & 4 & 0 \end{bmatrix}.$$

Rezultate možemo provjeriti primjenom programskog sustava *Mathematica*

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A = {1, 3, 4, 8};
        Agglomerate[A, DistanceFunction -> ManhattanDistance]
```

Primjer 5.12. *Zadan je skup* $\mathcal{A} = \{1,2,4,8,9,10,12,16,18,20\}$ *(vidi Sliku 5.12). Na njemu ćemo provesti Algoritam 1 primjenom sličnosti definirane s* (5.15). *Tijek algoritma može se pratiti u Tablici 5.2.*

j	$\Pi^{(j)}$	$\min R_j$	$\mathcal{F}_1(\Pi^{(j)})$
9	$\{\{1,2\},\{4\},\{8\},\{9\},\{10\},\{12\},\{16\},\{18\},\{20\}\}$	$(R_9)_{34} = 1$	1
8	$\{\{1,2\},\{4\},\{8,9\},\{10\},\{12\},\{16\},\{18\},\{20\}\}$	$(R_8)_{34} = 1$	2
7	$\{\{1,2\},\{4\},\{8,9,10\},\{12\},\{16\},\{18\},\{20\}\}$	$(R_7)_{12} = 2$	3
6	$\{\{1,2,4\},\{8,9,10\},\{12\},\{16\},\{18\},\{20\}\}$	$(R_6)_{23} = 2$	5
5	$\{\{1,2,4\},\{8,9,10,12\},\{16\},\{18\},\{20\}\}$	$(R_5)_{34} = 2$	8
4	$\{\{1,2,4\},\{8,9,10,12\},\{16\},\{18,20\}\}$	$(R_4)_{34} = 2$	10
3	$\{\{1,2,4\},\{8,9,10,12\},\{16,18,20\}\}$	$(R_3)_{12} = 4$	12
2	$\{\{1,2,4,8,9,10,12\},\{16,18,20\}\}$	$--$	30

Tablica 5.2: Tijek Algoritma 1 na skupu $\mathcal{A}$ iz Primjera 5.11 primjenom sličnosti definirane s (5.15)

Slika 5.12: Skup $\mathcal{A}$ iz Primjera 5.12

Primjer 5.13. *Slično kao u prethodnom Primjeru 5.10 primijenimo Algoritam 1 na podacima iz Primjera 4.19, str. 81 (vidi također Sliku 5.13a), gdje će također biti primijenjena ℓ_1-metrička funkcija i princip povezivanja zadan s* (5.15).

Primjenom niže navedenog programa dobivamo odgovarajući dendrogram. Usporedite dobivene rezultate s rezultatima iz Primjera 4.19 i Primjera 5.10.

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= A={{1,9},{2,9},{2,6},{1,3},{5,3},{6,4},{4,6},{7,7},{8,6},{9,8}};
        DendrogramPlot[A, Linkage -> "Single",
        DistanceFunction -> ManhattanDistance, HighlightLevel -> 2,
        LeafLabels -> (# &)]
```

5.4 Korištenje programskog sustava *Mathematica*

Aglomerativni hijerarhijski algoritmi opisani u ovom poglavlju mogu se implementirati pomoću programskog sustava *Mathematica*. U tu svrhu najprije treba uključiti paket `HierarchicalClustering` naredbom

```
In[1]:= Needs["HierarchicalClustering`"]
```

Nakon definiranja skupa $\mathcal{A}$, možemo pozvati odgovarajući *Mathematica*-modul

```
In[2]:= A={2, 4, 8, 10, 12};
        Agglomerate[A]
```

Dobivamo

```
Out[3]= Cluster[Cluster[2, 4, 4, 1, 1],
        Cluster[Cluster[8, 10, 4, 1, 1], 12, 4, 2, 1], 16, 2, 3]
```

U naredbi `Agglomerate[]` nismo naveli nikakvu dodatnu opciju. To znači da je „po defaultu" korištena LS-kvazimetrička funkcija

`DistanceFunction -> SquaredEuclideanDistance`,

a udaljenost između klastera definirana je kao minimalna udaljenost njihovih elemenata (5.2).

Oznaka

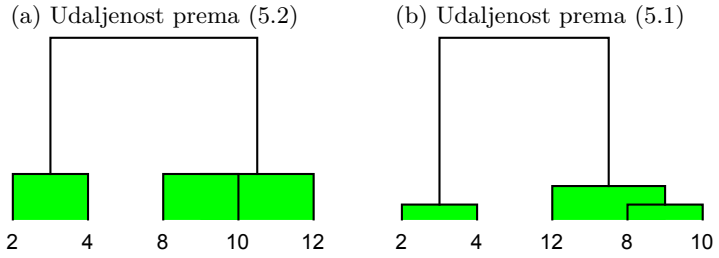
$$Cluster[\pi_1, \pi_2, D, m_1, m_2]$$

definira klaster dobiven spajanjem klastera π_1 s m_1 elemenata i klastera π_2 s m_2 elemenata, a udaljenost njihovih centara je D.

`Out[3]` znači sljedeće (vidi također Sliku 5.13:

- Najprije se spajaju jednočlani klasteri $\{2\}$ i $\{4\}$, čija je (5.2) udaljenost $2^2 = 4$;
- Nakon toga spajaju se jednočlani klasteri $\{8\}$ i $\{10\}$, čija je (5.2) udaljenost $2^2 = 4$. Tako dobiveni klaster $\{8, 10\}$ spaja se s jednočlanim klasterom $\{12\}$, čija je (5.2) udaljenost $2^2 = 4$;
- Na kraju spaja se dvočlani klaster $\{2, 4\}$ s tročlanim klasterom $\{8, 10, 12\}$, čija je (5.2) međusobna udaljenost jednaka $4^2 = 16$.

Pri tome se može dogoditi da u nekom koraku imamo više ravnopravnih mogućnosti za povezivanje klastera. Na te mogućnosti *Mathematica* također upozorava.



Slika 5.13: Dendrogrami skupa $\mathcal{A} = \{2, 4, 8, 10, 12\}$

Ako udaljenost između klastera želimo definirati kao udaljenost njihovih centroida, tj. pomoću (5.1), onda treba koristiti opciju `Linkage -> "Centroid"`, tj.

```
In[4]:= A={2, 4, 8, 10, 12};
        Agglomerate[A, Linkage -> "Centroid"]
```

U tom slučaju dobivamo

```
Out[4]= Cluster[Cluster[2, 4, 4, 1, 1],
        Cluster[12, Cluster[8, 10, 4, 1, 1], 9, 1, 2], 49, 2, 3]
```

Udaljenost klastera prilikom korištenja programskog sustava *Mathematica* može se još definirati i kao prosječna udaljenost (5.4) (`Linkage->"Average"`) ili kao Wardova udaljenost (5.12) (`Linkage->"Ward"`).

Aglomerativni hijerarhijski algoritam možemo provesti i korištenjem ℓ_1-metričke funkcije, koja se poziva opcijom

`DistanceFunction -> ManhattanDistance`.

I u tom slučaju mogu se koristiti različite mogućnosti za definiranje udaljenosti između klastera, tj. povezivanja (`Linkage`).

Poglavlje 6

Odabir najprikladnijeg broja klastera: Indeksi

Neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \colon i = 1, \ldots, m\} \subset \mathbb{R}^n$ skup podataka, a $\mathcal{P}(\mathcal{A}; k)$ skup svih njegovih k-particija $\Pi = \{\pi_1, \ldots, \pi_k\}$ s k klastera $\pi_1, \ldots, \pi_k$. Ako uvedemo neku kvazimetričku funkciju $d \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, onda svakom klasteru možemo pridružiti njegov centar

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k, \tag{6.1}$$

i definirati kriterijsku funkciju cilja

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a). \tag{6.2}$$

Pri tome je sljedeće važno pitanje ostalo otvoreno:

> U koliko bi klastera bilo najprihvatljivije grupirati promatrani skup podataka $\mathcal{A}$?

ili

> Kako za promatrani skup podataka $\mathcal{A}$ odabrati particiju s najprikladnijim brojem klastera?

Odgovor na ova pitanja jedan je od najsloženijih problema klaster analize. O tome postoji brojna stručna literatura [6, 11, 14, 19, 30, 31, 51, 53, 60, 68, 73, 78], a obično se rješava ispitivanjem različitih pokazatelja koje jednostavno nazivamo indeksi.

U nekim jednostavnim slučajevima broj klastera u particiji određen je samom prirodom problema. Primjerice, prirodno je studente grupirati u $k = 5$ klastera prema postignutom uspjehu na studiju, ali na pitanje u koliko bi klastera bilo najprihvatljivije grupirati skup svih kukaca ili u koliko skupina treba grupirati privredne subjekte neke administrativne jedinice nije lako dati odgovor.

Ako broj klastera u koji treba grupirati skup $\mathcal{A}$ nije unaprijed poznat, prirodno bi bilo potražiti particiju koja se sastoji od klastera koji su interno što kompaktniji, a eksterno što bolje međusobno razdvojeni. Za takvu particiju reći ćemo da ima najprikladniji broj klastera.

U ovom udžbeniku razmotrit ćemo primjenu samo nekoliko najpoznatijih indeksa, koji pretpostavljaju korištenje LS-kvazimetričke funkcije.

6.1 Pokazatelj vrijednosti funkcije cilja

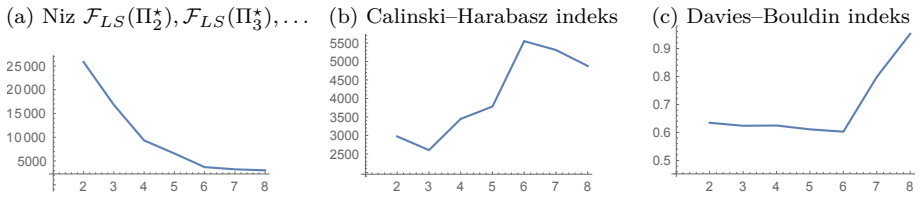
Poznato je [6, 31, 68, 73] da vrijednost funkcije cilja ne raste povećanjem broja klastera u particiji, tj. niz funkcijskih vrijednosti na optimalnim particijama

$$\mathcal{F}_{LS}(\Pi_2^\star) \geq \mathcal{F}_{LS}(\Pi_3^\star) \geq \cdots \geq \mathcal{F}_{LS}(\Pi_k^\star) \geq \cdots$$

monotono je padajuć. Zato ima smisla optimalnu particiju s najprikladnijim brojem klastera smatrati onom za koju vrijednost funkcije cilja naglo padne. To naravno nije egzaktan kriterij, ali zajedno s nekim drugim može ukazati na traženu particiju s najprikladnijim brojem klastera.

Primjer 6.1. *Neka je $\mathcal{A}$ skup iz Primjera 3.2, str. 25, a $\Pi_2^\star, \Pi_3^\star, \ldots$ njegove optimalne particije. Pokušajmo odrediti najprikladniji broj klastera skupa $\mathcal{A}$ promatrajući samo niz funkcijskih vrijednosti $\mathcal{F}_{LS}(\Pi_2^\star), \mathcal{F}_{LS}(\Pi_3^\star), \ldots$.*

Na Slici 6.1 a prikazan je graf tih funkcijskih vrijednosti kao funkcija od broja klastera. Može se primijetiti da se nagli pad vrijednosti funkcije cilja dogodio u slučaju particije s 4 i u slučaju particije s 6 klastera. To bi trebalo značiti da particiju s najprikladnijim brojem klastera treba tražiti između particija $\Pi_4^\star$ i $\Pi_6^\star$.



Slika 6.1: Izbor particije s najprikladnijim brojem klastera iz Primjera 3.2

6.2 Calinski–Harabasz indeks

Ovaj indeks predložili su T. Calinski i J. Harabasz u svom radu *„A dendrite method for cluster analysis"* objavljenom 1974. godine u časopisu Communications in Statistics. Nakon toga Calinski–Harabasz (CH) indeks doživio je brojna usavršavanja i prilagođavanja (vidi primjerice [6, 60, 78]).

CH indeks definira se tako da interno kompaktnija particija čiji su klasteri dobro međusobno razdvojeni ima veću CH vrijednost.

Ako prilikom određivanja optimalne k-particije $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ koristimo LS-kvazimetričku funkciju, onda funkciju cilja (6.2) možemo zapisati

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|c_j - a\|_2^2. \tag{6.3}$$

Vrijednost funkcija $\mathcal{F}_{LS}$ na optimalnoj particiji $\Pi^\star$ pokazuje ukupno „rasipanje" elemenata svih klastera $\pi_1^\star, \ldots, \pi_k^\star$ te particije do njihovih centroida $c_1^\star, \ldots, c_k^\star$. Kao što smo ranije primijetili, što je vrijednost funkcije $\mathcal{F}_{LS}$ manja, time je „rasipanje" manje, što znači da su klasteri interno kompaktniji.

Zato ćemo pretpostaviti da je CH-indeks optimalne particije $\Pi^\star$ obrnuto proporcionalan vrijednosti funkcije cilja $\mathcal{F}_{LS}(\Pi^\star)$.

Kao što smo primijetili ranije u Poglavlju 3, str. 31, odnosno str. 48, prilikom traženja optimalne particije, osim minimizacije funkcije $\mathcal{F}_{LS}$, možemo potražiti maksimum odgovarajuće dualne funkcije

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j \|c_j - c\|_2^2, \quad m_j = |\pi_j|, \tag{6.4}$$

pri čemu je $c = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \|x - a^i\|_2^2 = \frac{1}{m} \sum_{i=1}^{m} a^i$ centroid čitavog skupa $\mathcal{A}$.

Vrijednost funkcija $\mathcal{G}$ na particiji $\Pi^\star$ pokazuje ukupnu težinsku razdvojenost centroida $c_1^\star, \ldots, c_k^\star$ klastera $\pi_1^\star, \ldots, \pi_k^\star$. Što je vrijednost funkcije $\mathcal{G}$ veća, time su i LS-udaljenosti centroida $c_j^\star$ do centroida čitavog skupa $c^\star$ veće, pri čemu udaljenosti ponderiramo s brojem elemenata u pojedinom klasteru. To znači da su i centroidi $c_j^\star$ međusobno maksimalno moguće razdvojeni.

Zato ćemo pretpostaviti da je `CH`*-indeks optimalne particije* $\Pi^\star$ *proporcionalan vrijednosti funkcije cilja* $\mathcal{G}(\Pi^\star)$.

Uzevši u obzir još i statističke razloge povezane s brojem m elemenata skupa $\mathcal{A}$ i brojem k klastera u particiji $\Pi^\star$, mjera interne kompaktnosti i eksterne razdvojenosti klastera optimalne particije $\Pi^\star$ u slučaju primjene LS-kvazimetričke funkcije definirana je brojem

$$\texttt{CH}(k) = \frac{\mathcal{G}(\Pi^\star)/(k-1)}{\mathcal{F}_{LS}(\Pi^\star)/(m-k)}, \tag{6.5}$$

koji nazivamo `CH`-indeks particije $\Pi^\star$.

Primjer 6.2. *Promatrajmo skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *iz Primjera 3.14, str. 40. Dobivenu LS-optimalnu* 3*-particiju* $\Pi^\star(3) = \{\{2, 4\}, \{8, 10\}, \{16\}\}$ *usporedimo s LS-optimalnom* 2*-particijom. Optimalne particije možemo potražiti metodom opisanom u t. 4.5, str. 87.*

Za LS-optimalnu 3-particiju $\Pi^\star(3)$ (vidi Sliku 6.2 b) dobili smo

$$\mathcal{F}_{LS}(\Pi^\star(3)) = 4, \quad \mathcal{G}(\Pi^\star(3)) = 116.$$

LS-optimalna 2-particija je $\Pi^\star(2) = \{\{2, 4\}, \{8, 10, 16\}\}$ (vidi Sliku 6.2 a) za koju je

$$\mathcal{F}_{LS}(\Pi^\star(2)) = 36.67, \quad \mathcal{G}(\Pi^\star(2)) = 83.33.$$

Primijetite da sukladno teoriji vrijedi $\mathcal{F}_{LS}(\Pi^\star(2)) \geq \mathcal{F}_{LS}(\Pi^\star(3))$ i također $\mathcal{G}(\Pi^\star(2)) \leq \mathcal{G}(\Pi^\star(3))$.

(a) LS-optimalna 2-particija (b) LS-optimalna 3-particija

Slika 6.2: Izbor particije s prikladnijim brojem klastera

Odgovarajući `CH`-indeksi su

$$\texttt{CH}(2) = \frac{83.33/1}{36.67/3} = 6.82, \qquad \texttt{CH}(3) = \frac{116/2}{4/2} = 29.$$

Budući da je $\texttt{CH}(3) > \texttt{CH}(2)$, particija $\Pi^\star(3)$ particija je s prihvatljivijim (prikladnijim) brojem klastera.

Primjer 6.3. *Promatrajmo ponovo skup $\mathcal{A}$ iz Primjera 3.2, str. 25. Pokušajmo odrediti najprikladniji broj klastera primjenom* `CH`*-indeksa.*

`CH`-indeks za optimalne particije $\Pi_2^\star, \ldots, \Pi_8^\star$ grafički je prikazan na Slici 6.1 b, str. 113. Budući da je `CH`-indeks definiran tako da interno kompaktnija particija čiji klasteri su bolje međusobno razdvojeni ima veću `CH` vrijednost, jasno se vidi da je $\Pi_6^\star$, particija s najprihvatljivijem brojem klastera.

6.2.1 Davies–Bouldin indeks

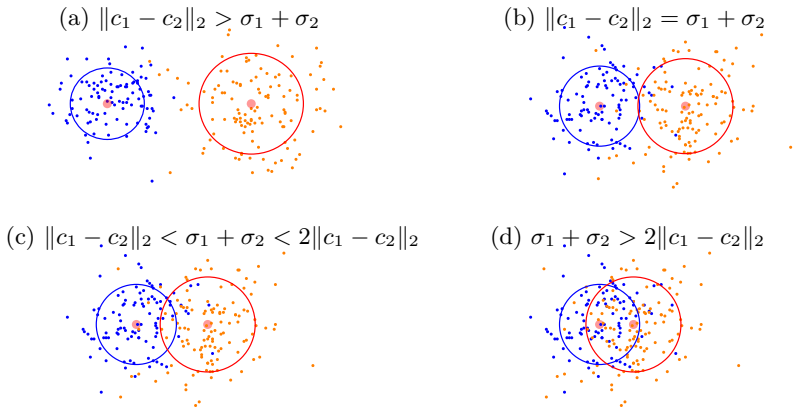
Ovaj indeks predložili su D. Davies i D. Bouldin u svom radu *„A cluster separation measure"* objavljenom 1979. godine u časopisu IEEE Transactions on Pattern Analysis and Machine Intelligence. I ovaj indeks kasnije je doživio brojne prilagodbe i usavršavanja (vidi primjerice [6, 60, 78–80]).

DB-indeks definira se tako da interno kompaktnija particija čiji su klasteri međusobno bolje razdvojeni ima manju `DB` vrijednost.

Niže navedeni koncept preuzet je iz [79]. Neka je $c \in \mathbb{R}^2$ točka u ravnini oko koje primjenom Gaussove normalne distribucije s varijancom σ^2 generirano m slučajnih točaka a^i. Ovaj skup točaka čini *sferičan skup podataka* kojeg ćemo označiti s $\mathcal{A}$.

Iz statističke literature [3] poznato je da se u krugu $K(c, \sigma)$ sa središtem u točki c i radijusom σ (standardna devijacija) nalazi oko 68% točaka skupa $\mathcal{A}$. Ovaj krug zvat ćemo glavni krug skupa podataka $\mathcal{A}$.

Pretpostavimo da su za dvije različite točke $c_1, c_2 \in \mathbb{R}^2$ i dvije različite varijance σ_1^2, σ_2^2 na prethodno opisan način generirana dva sferična skupa podataka $\mathcal{A}_1, \mathcal{A}_2$ i da su $K_1(c_1, \sigma_1)$, $K_2(c_2, \sigma_2)$ njihovi odgovarajući glavni krugovi. Radijus σ_1 prvog kruga je standardna devijacija skupa $\mathcal{A}_1$, a radijus drugog kruga σ_2 je standardna devijacija skupa $\mathcal{A}_2$.



Slika 6.3: Međusobno različiti odnosi dva sferična skupa podataka

Mogući odnosi skupova $\mathcal{A}_1, \mathcal{A}_2$ s obzirom na međusobni položaj njihovih glavnih krugova $K_1(c_1, \sigma_1)$, $K_2(c_2, \sigma_2)$ prikazani su na Slici 6.3. Na Slici 6.3 a prikazani su skupovi $\mathcal{A}_1, \mathcal{A}_2$ čiji se glavni krugovi ne sijeku i za koje vrijedi $\|c_1 - c_2\|_2 > \sigma_1 + \sigma_2$, na Slici 6.3 b prikazani su skupovi $\mathcal{A}_1, \mathcal{A}_2$ čiji se glavni krugovi dodiruju i za koje vrijedi $\|c_1 - c_2\|_2 = \sigma_1 + \sigma_2$, itd.

Dakle, možemo reći da se glavni krugovi $K_1(c_1, \sigma_1)$, $K_2(c_2, \sigma_2)$ skupova $\mathcal{A}_1, \mathcal{A}_2$ presijecaju (imaju neprazan presjek) ako vrijedi [79]

$$\|c_1 - c_2\|_2 \leq \sigma_1 + \sigma_2, \tag{6.6}$$

odnosno ako vrijedi

$$\frac{\sigma_1 + \sigma_2}{\|c_1 - c_2\|_2} > 1.$$

Promatrajmo sada optimalnu particiju $\Pi^\star$ skupa $\mathcal{A}$ s klasterima $\pi_1^\star, \ldots, \pi_k^\star$ i njihovim centroidima $c_1^\star, \ldots, c_k^\star$. Uočimo jedan od klastera $\pi_j^\star$ i razmotrimo njegov odnos prema ostalim klasterima. Primijetite da je veličinom

$$D_j := \max_{s \neq j} \frac{\sigma_j + \sigma_s}{\|c_j^\star - c_s^\star\|_2} \tag{6.7}$$

zadano najveće moguće preklapanje klastera $\pi_j^\star$ s nekim drugim klasterom. Pri tome su

$$\sigma_j := \tfrac{1}{|\pi_j^\star|} \sum_{a \in \pi_j^\star} \|c_j^\star - a\|_2^2, \qquad j = 1, \ldots, k, \tag{6.8}$$

standardne devijacije podataka klastera $\pi_1^\star, \ldots, \pi_k^\star$. Veličina

$$\tfrac{1}{k}(D_1 + \cdots + D_k), \tag{6.9}$$

prosjek je brojeva (6.7), a predstavlja još jednu mjeru interne kompaktnosti i eksterne razdvojenosti klastera u particiji. Jasno je da što je broj (6.9) manji, klasteri su kompaktniji i bolje razdvojeni. Zato se DB-indeks optimalne particije $\Pi^\star$ skupa $\mathcal{A}$ s klasterima $\pi_1^\star, \ldots, \pi_k^\star$ i njihovim centroidima $c_1^\star, \ldots, c_k^\star$ definira na sljedeći način [6, 14, 53, 60, 78, 79]

$$\texttt{DB}(k) = \frac{1}{k}\sum_{j=1}^{k} \max_{s\neq j} \frac{\sigma_j + \sigma_s}{\|c_j^\star - c_s^\star\|_2}, \quad \text{gdje je} \quad \sigma_j^2 = \frac{1}{|\pi_j^\star|}\sum_{a\in\pi_j^\star} \|c_j^\star - a\|_2^2. \tag{6.10}$$

Primjer 6.4. *Promatrajmo ponovo skup $\mathcal{A} = \{2, 4, 8, 10, 16\}$ iz Primjera 6.8. Primjenom DB-indeksa odredimo particiju s najprikladnijim brojem klastera.*

LS-optimalna 2-particija je $\Pi^\star(2) = \{\{2, 4\}, \{8, 10, 16\}\}$. Centri njenih klastera su: $c_1^\star = 3$, $c_2^\star = 11.33$, a odgovarajuće standardne devijacije: $\sigma_1 = 1$, $\sigma_2 = 3.4$. Odgovarajući DB-indeks je

$$\texttt{DB}(2) = \frac{1}{2}\left(\frac{\sigma_1+\sigma_2}{\|c_1^\star - c_2^\star\|_2} + \frac{\sigma_2+\sigma_1}{\|c_2^\star - c_1^\star\|_2}\right) = 0.58788.$$

LS-optimalna 3-particija je $\Pi^\star(3) = \{\{2, 4\}, \{8, 10\}, \{16\}\}$. Centri njenih klastera su: $c_1^\star = 3$, $c_2^\star = 9$, $c_3^\star = 16$, a odgovarajuće standardne devijacije: $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 0$. Odgovarajući DB-indeks je

$$\begin{aligned}\texttt{DB}(3) = \frac{1}{3}\Bigg(&\max\left\{\frac{\sigma_1+\sigma_2}{\|c_1^\star - c_2^\star\|_2}, \frac{\sigma_1+\sigma_3}{\|c_1^\star - c_3^\star\|_2}\right\} + \max\left\{\frac{\sigma_2+\sigma_1}{\|c_2^\star - c_1^\star\|_2}, \frac{\sigma_2+\sigma_3}{\|c_2^\star - c_3^\star\|_2}\right\} \\ &+ \max\left\{\frac{\sigma_3+\sigma_1}{\|c_3^\star - c_1^\star\|_2}, \frac{\sigma_3+\sigma_2}{\|c_3^\star - c_2^\star\|_2}\right\}\Bigg) = 0.26984.\end{aligned}$$

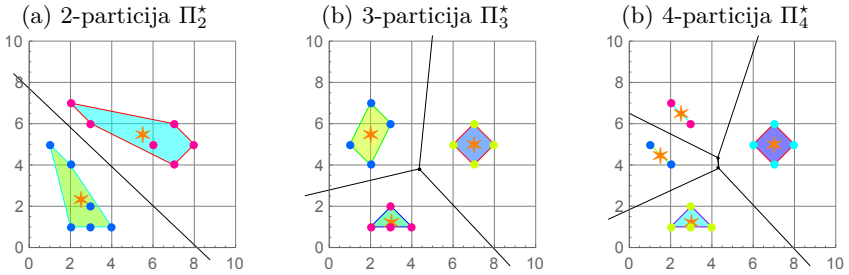
Budući da je $\texttt{DB}(3) < \texttt{DB}(2)$, particija $\Pi^\star(3)$ je particija s prihvatljivijim (prikladnijim) brojem klastera što je u suglasju s ranije dobivenim zaključkom pomoću CH-indeksa.

Primjer 6.5. *Promatrajmo ponovo skup $\mathcal{A}$ iz Primjera 3.2, str. 25. Odredimo najprikladniji broj klastera primjenom DB-indeksa.*

DB-indeks za optimalne particije $\Pi_2^\star, \ldots, \Pi_8^\star$ grafički je prikazan na Slici 6.1 c. Budući da je DB-indeks definiran tako da interno kompaktnija particija čiji su klasteri bolje međusobno razdvojeni ima manju DB vrijednost, jasno se vidi da je i prema DB-indeksu $\Pi_6^\star$ particija s najprihvatljivijem brojem klastera.

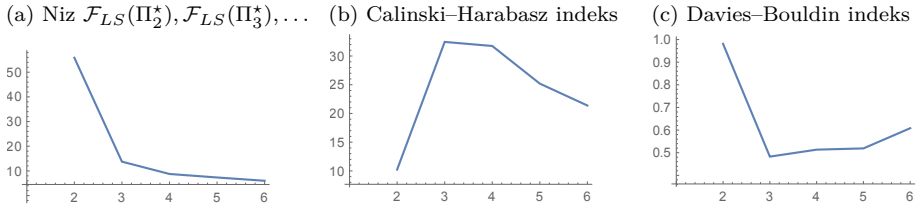
Primjer 6.6. *Potražimo particiju skupa $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 12\}$ s najprikladnijim brojem klastera pri čemu je*

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	2	3	2	2	3	4	3	6	8	7	7
y_i	5	4	6	7	1	2	1	1	5	5	4	6



Slika 6.4: Izbor LS-optimalne particije s najprikladnijim brojem klastera

Za optimalne LS-particije $\Pi_2^\star, \ldots, \Pi_6^\star$ s $2, 3, \ldots, 6$ klastera izračunat ćemo vrijednost funkcije cilja $\mathcal{F}_{LS}$, vrijednost `CH`-indeksa i vrijednost `DB`-indeksa. Na Slici 6.4 prikazane su LS-optimalna 2-particija, 3-particija i 4-particija, a na Slici 6.5 grafovi koji prikazuju vrijednost funkcije cilja i spomenutih indeksa za LS-optimalne particije.



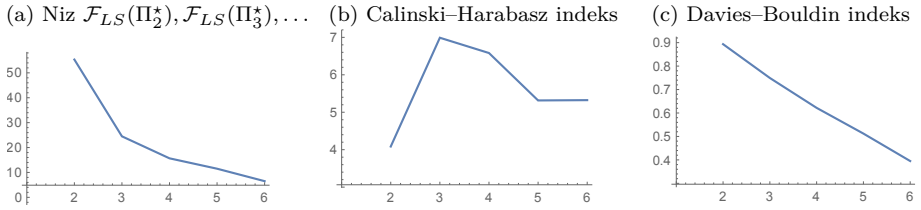
Slika 6.5: Izbor particije s najprikladnijim brojem klastera

Kao što se može vidjeti, funkcija cilja (Slika 6.5 a) ima nagli pad na 3-particiji. Također, `CH`-indeks prima najveću, a `DB`-indeks najmanju vrijednost na istoj particiji. To znači da s visokom sigurnošću možemo tvrditi da je particija s najprikladnijim brojem klastera upravo 3-particija, što je i vizualno očekivano (vidi Sliku 6.4).

Sljedeći primjer pokazuje da zaključivanje o najprikladnijem broju klastera u particiji nije uvijek jednoznačno.

Primjer 6.7. *Potražimo LS-optimalnu particiju skupa $\mathcal{A}$ iz Primjera 4.17, str. 79, s najprikladnijim brojem klastera.*

Nakon što smo pronašli LS-optimalne particije s $2, 3, 4, 5, 6$ klastera, u mogućnosti smo odrediti vrijednosti funkcije cilja $\mathcal{F}_{LS}$, vrijednosti `CH`-indeksa i `DB`-indeksa (vidi Sliku 6.6). Niz funkcijskih vrijednosti na Slici 6.6 a i `CH`-indeks na Slici 6.6 b ukazuju na 3-particiju kao particiju s najprikladnijim brojem klastera, ali to ne potvrđuje i `DB`-indeks.



Slika 6.6: Izbor particije s najprikladnijim brojem klastera

Primjer 6.8. *Za skup* $\mathcal{A} = \{1, 2, 4, 8, 9, 10, 12, 16, 18, 20\}$ *iz Primjera 5.12, str. 107, provest ćemo aglomeracijski hijerarhijski algoritam primjenom LS-kvazimetričke funkcije i udaljenosti među klasterima definirane s* (5.1)*. Pokušajmo odrediti najprikladniji broj klastera primjenom* `CH`*-indeksa i* `DB`*-indeksa.*

Za svaku particiju odredit ćemo vrijednost funkcije cilja $\mathcal{F}_{LS}$, odgovarajuću vrijednost funkcije cilja $\mathcal{G}$ te odgovarajuću vrijednost `CH`-indeksa i `DB`-indeksa. Rezultati su vidljivi u Tablici 6.1.

k	$\Pi^{(k)}$	$\mathcal{F}_{LS}$	$\mathcal{G}$	`CH`(k)	`DB`(k)
9	$\{\{1,2\},\{4\},\{8\},\{9\},\{10\},\{12\},\{16\},\{18\},\{20\}\}$	.5	389.5	97.4	.082
8	$\{\{1,2\},\{4\},\{8,9\},\{10\},\{12\},\{16\},\{18\},\{20\}\}$	1.0	389.0	111.1	.171
7	$\{\{1,2\},\{4\},\{8,9,10\},\{12\},\{16\},\{18\},\{20\}\}$	2.5	387.5	75.3	.175
6	$\{\{1,2,4\},\{8,9,10\},\{12\},\{16\},\{18\},\{20\}\}$	6.7	383.3	46.0	.195
5	$\{\{1,2,4\},\{8,9,10,12\},\{16\},\{18\},\{20\}\}$	13.4	376.6	35.1	.259
4	$\{\{1,2,4\},\{8,9,10,12\},\{16,18\},\{20\}\}$	15.4	374.6	48.6	.353
3	$\{\{1,2,4\},\{8,9,10,12\},\{16,18,20\}\}$	21.4	368.6	60.2	.374
2	$\{\{1,2,4,8,9,10,12\},\{16,18,20\}\}$	115.7	274.3	19.0	.486

Tablica 6.1: Karakteristike particija skupa $\mathcal{A}$ iz Primjera 6.8

Niz funkcijskih vrijednosti $\mathcal{F}_{LS}(\Pi_2^\star), \mathcal{F}_{LS}(\Pi_3^\star), \ldots$ ukazuje da bi $\Pi_2^\star$ mogla biti particija s najprikladnijim brojem klastera. `CH` indeks također ukazuje na particiju $\Pi_2^\star$ kao najprikladniju particiju s manjim brojem klastera, ali apsolutnu prednost daje particiji $\Pi_8^\star$. `DB` indeks ne daje prijedlog particije s najprikladnijim brojem klastera. Dakle, najmanje bismo pogriješili ako bismo particiju $\Pi_2^\star$ ili particiju $\Pi_8^\star$ proglasili particijom s najprikladnijim brojem klastera.

Općenito, treba reći da navedeni indeksi daju prihvatljive zaključke ako su podaci takvi da se uklapaju u pretpostavke na osnovi kojih su indeksi konstruirani. Pri tome, kao što pokazuju i prethodno navedeni primjeri, zaključivanje o najprikladnijem broju klastera na skupu s relativno malenim

brojem podataka neće biti dovoljno pouzdano.

Prethodna izračunavanja provedena su na temelju niže navedenog *Mathematica* programa. Navodimo samo program za izračun karakteristika particije s 8 klastera. Na sličan način mogu se dobiti i karakteristike drugih particija.

```
In[1]:= A = {1, 2, 4, 8, 9, 10, 12, 16, 18, 20}; m = Length[A];
        w = Table[1, {i, m}];
        cc = Total[A]/Length[A]
        pod = Table[{w[[i]], A[[i]]}, {i, m}];

In[2]:= PI = {{pod[[1]], pod[[2]]}, {pod[[3]]}, {pod[[4]], pod[[5]]},
              {pod[[6]]}, {pod[[7]]}, {pod[[8]]}, {pod[[9]]}, {pod[[10]]}}
        k = Length[PI];
        c = Table[Mean[PI[[j, All, 2]]], {j, Length[PI]}];
        ch = VCH[PI, c, cc];
        db = VDB[PI, c];
        Print["F=", ch[[1]], "; G=", ch[[2]], "; CH(", k, ")=",ch[[3]],
              "; DB(", k, ")=", db]
```

Poglavlje 7

Jedna primjena: analiza temperaturnih promjena u Osijeku

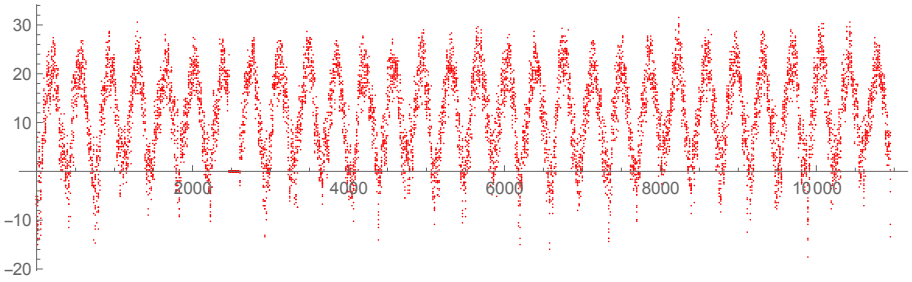
Na kraju navodimo jedan praktični primjer sa stvarnim podacima o kretanju prosječnih dnevnih temperatura u Osijeku od 1985. godine. Pri tome nemamo ambiciju davati zaključke i prijedloge vezano uz klimatske procese, već samo opisati jednu mogućnost korištenja teorije i metoda navedenih u ovom udžbeniku.

7.1 Podaci o prosječnoj dnevnoj temperaturi u Osijeku

Promatrat ćemo podatke o prosječnoj dnevnoj temperaturi u Osijeku od 1985. do 2014. godine[1] (vidi Sliku 7.1)

Radi se o $m = 10\,950$ podataka (v_i, T_i), $i = 1, \ldots, m$ s dva obilježja: vremenski trenutak v_i iskazan datumom i prosječna temperatura T_i toga dana izražena u °C. Odmah treba primijetiti da nedostaju podaci od 1. listopada 1991. godine do 16. veljače 1992. godine jer je to bilo vrijeme najžešćih napada na Osijek za vrijeme Domovinskog rata.

[1] Izvor: Republika Hrvatska - Državni hidrometeorološki zavod, Klimatološko meteorološki sektor, Zagreb, Grič 3



Slika 7.1: Kretanje prosječne dnevne temperature u Osijeku (1985. – 2014.)

Sliku 7.1 na kojoj su prikazani podaci i pripadnu linearnu regresiju možemo dobiti jednostavnim *Mathematica*-programom u kome najprije učitamo podatke iz datoteke `TOS-1985-2014.txt`

```
In[1]:= SetDirectory[NotebookDirectory[]];
        pod = Import["TOS-1985-2014.txt", "Table"];
        m = Length[pod]
        data = Table[{i, pod[[i, 4]]}, {i, m}];
        ListPlot[data, PlotStyle -> {Red, PointSize[.001]}, AspectRatio ->.3,
                  Axes -> True]
        lm = LinearModelFit[data, x, x]
```

Linearna regresija $x \mapsto 10.5267 + 0.00017032x$ pokazuje porast prosječne dnevne temperature u Osijeku. Koeficijent $k = 0.00017032$ pokazuje da je za proteklih približno 10 000 dana (promatrani period od 1985. do 2014. godine) prosječna dnevna temperatura u Osijeku porasla za 1.7°C.

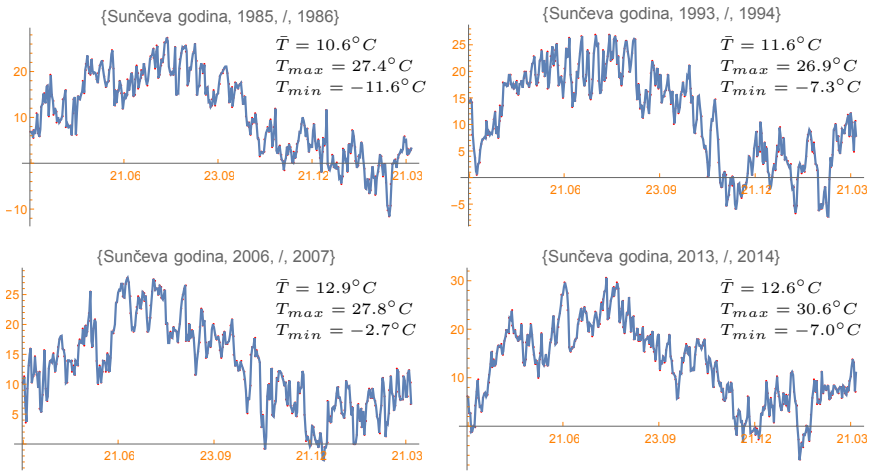
7.2 Trend kretanja prosječnih dnevnih temperatura

U cilju promatranja temperaturnih promjena tijekom spomenutog razdoblja promatrat ćemo vremenske periode tzv. „Sunčevih" ili „tropskih godina"[2], tj. godišnjih ciklusa koji počinju i završavaju pozicijom Sunca u proljetnoj točki. Pri tome pod proljetnom točkom podrazumijevamo jedno od dvaju presjecišta nebeskog ekvatora i ekliptike, gdje Sunce u prividnome godišnjem gibanju prelazi s južne na sjevernu nebesku polutku, što se događa

[2] „Hrvatska enciklopedija", Leksikografski zavod Miroslav Krleža, `www.enciklopedija.hr`, 2014.

oko 21. ožujka. To je astronomski početak proljeća. Upravo je praćenje perioda izmjene godišnjih doba bilo jedan od najvažnijih praktičkih zadataka astronomije u prošlosti.

Promatranje jednog godišnjeg ciklusa kao Sunčeve godine koja počinje 21. ožujka te godine, a završava 21. ožujka sljedeće godine pokazuje jedan kompletni životni ciklus prirode. Tako ćemo primjerice Sunčevu godinu koja počinje 21. ožujka 1985. godina, a završava 21. ožujka 1986. godine označiti jednostavno s „1985/1986". Za nekoliko takvih Sunčevih godina na Slici 7.2 prikazano je kretanje prosječnih dnevnih temperatura u gradu Osijeku. Također, na slici su navedena i neka karakteristična svojstva tih Sunčevih godina: prosječna godišnja temperatura $\bar{T}$, najviša (T_{max}) i najniža (T_{min}) prosječna dnevna temperatura u °C.



Slika 7.2: Prosječne dnevne temperature u Osijeku tijekom nekih Sunčevih godina

7.2.1 Kretanje prosječnih godišnjih temperatura

Promatrat ćemo podatke o prosječnim dnevnim temperaturama za sljedeće nizove Sunčevih godina:

$$\begin{aligned} (\text{XX}): &\quad 1985/1986, \quad 1986/1987, \ldots, 1999/2000, \\ (\text{XXI}): &\quad 2000/2001, \quad 2001/2002, \ldots, 2013/1014, \end{aligned}$$

od kojih prvi niz ima 15, a drugi 14 Sunčevih godina.

Neka je $T_1^{(s)}, \ldots, T_{365}^{(s)}$ niz prosječnih dnevnih temperatura za s-tu Sunčevu godinu. Prosječna godišnja temperatura za tu godinu računa se po jednostavnoj formuli[3]

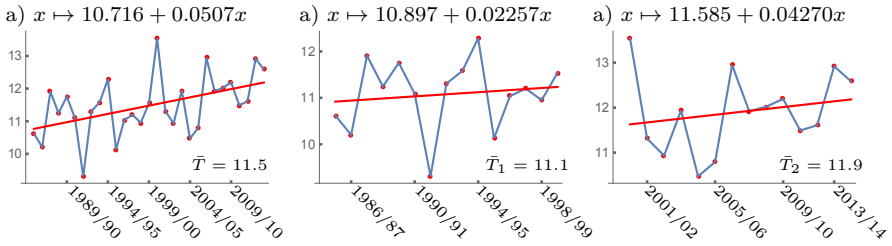
$$\bar{T}^{(s)} = \tfrac{1}{365} \sum_{i=1}^{365} T_i^{(s)}. \tag{7.1}$$

Sukladno Zadatku 2.3, str. 6, ova formula ima smisla jer se svaki pribrojnik $T_i^{(s)}$ dobije kao prosjek 24-satne temperature i-tog dana.

Razmotrimo kretanje prosječnih godišnjih temperatura za nizove (XX) i (XXI). Prosječna godišnja temperatura za period (XX), prosječna godišnja temperatura za period (XXI) i prosječna godišnja temperatura za period (XX-XXI) računaju se po jednostavnim formulama

$$\bar{T}_1 = \tfrac{1}{15} \sum_{s=1}^{15} \bar{T}^{(s)}, \quad \bar{T}_2 = \tfrac{1}{14} \sum_{s=16}^{29} \bar{T}^{(s)}, \quad \bar{T} = \tfrac{1}{29} \sum_{s=1}^{29} \bar{T}^{(s)}. \tag{7.2}$$

Budući da po pretpostavci svaka godina ima po 365 dana, sukladno Zadatku 2.3, str. 6, ove su formule korektne.



Slika 7.3: Kretanje prosječnih godišnjih temperatura u Osijeku:
a) (1985/86 – 2013/14), b) (1985/86 – 1999/00), c) (2000/01 – 2013/14)

Slika 7.3a pokazuje da je prosječna godišnja temperatura za grad Osijek u proteklih 29 godina porasla za 1.7°C. Pri tome je porast u prvom periodu (XX) izražen nešto slabije (Slika 7.3b) od porasta u periodu (XXI) (Slika 7.3c).

Burnov dijagram prosječnih dnevnih temperatura

Za daljnju analizu temperaturnih kretanja bit će nam potrebna specijalna reprezentacija periodičnih podataka. Budući da se u analizi godišnjih temperatura radi o tipičnom primjeru pojave koja pokazuje periodično

[3]Zbog jednostavnosti, podatke koji se odnose na dan 29. veljače prestupnih godina nećemo koristiti.

ponašanje s temeljnim periodom od jedne godine[4], sve podatke zgodno je prikazati pomoću Burnovog dijagrama (vidi t. , str. 17) na Slici 7.4.

Kao što je već ranije opisano u t. 2.2.5, str. 15, u tu svrhu podatke najprije treba prilagoditi za prikaz na kružnici. U cilju pojednostavljenja računskog procesa naše podatke (v_i, T_i), $i = 1, \ldots, m$ s dva obilježja (v, T) promatrat ćemo kao podatke s jednim obilježjem v s odgovarajućim težinama (temperatura T u tom trenutku). Podatke o vremenskim trenucima $v_i = (gg_i, mm_i, dd_i)$, koji se sastoje od godine (gg_i), mjeseca (mm_i) i dana (dd_i), promatrat ćemo kao podatke na jediničnoj kružnici. U tu svrhu najprije vremenski trenutak koji odgovara 1. siječnju 1985. godine proglasimo početnim trenutkom i pridružimo mu realan broj 0. Ostalim vremenskim trenucima $v_i = (gg_i, mm_i, dd_i)$ pridružujemo realne brojeve po formuli

$$v_i = gg_i + (mm_i - 1)/12 + dd_i/365. \tag{7.3}$$

Zbog pojednostavljivanja, pretpostavili smo da svaka godina ima 365 dana i 12 mjeseci.

Nadalje, transformacijom

$$t_i = 2\pi(v_i - 1985), \quad i = 1, \ldots, m, \tag{7.4}$$

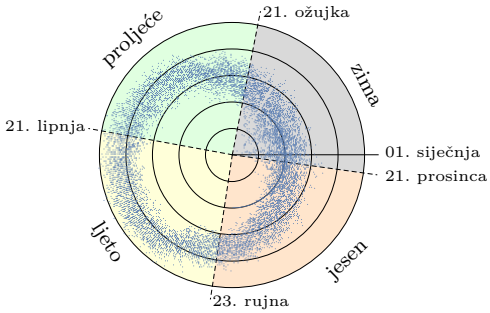
sve vremenske podatke koji pripadaju 1985. godini preslikat ćemo na interval $[0, 2\pi]$, sve vremenske podatke koji pripadaju 1986. godini preslikat ćemo na interval $[2\pi, 4\pi]$, itd. Na taj način jednoj godini pridružujemo duljinu 2π. Nakon toga brojeve t_i „namotat" ćemo na jediničnu kružnicu $K(O, 1)$ radijusa 1 sa središtem u točki O počevši od točke $(1, 0)$. Neki broj t_i pri tome pada u točku

$$a_i = a_i(\tau_i) = (\cos\tau_i, \sin\tau_i) \in K, \tag{7.5}$$

na kružnici $K(O, 1)$, gdje je τ_i duljina kružnog luka dobivena na sljedeći način

$$\begin{aligned} &Do[\\ &\quad \tau_i = t_i;\\ &\quad While[\tau_i > 2\pi, \tau_i = \tau_i - 2\pi],\\ &\quad \{i, m\}] \end{aligned}$$

[4]Ako bismo promatrali satne podatke o temperaturi na nekom mjestu, tada bi se radilo o periodičnoj pojavi koja nastaje kao superpozicija dva periodična utjecaja temeljnih perioda 1 dan i 1 godina.



Slika 7.4: Burnov dijagram prosječnih dnevnih temperatura u Osijeku (1985.–2014.)

Na taj način svakom vremenskom trenutku v_i pridružili smo jednu točku na kružnici $K(O,1)$, na kojoj točka $(1,0)$ predstavlja početak, odnosno kraj godine, a primjerice točka $(0,\frac{\pi}{2})$ predstavlja završetak prve četvrtine godine. Tako dobivamo skup točkaka na jediničnoj kružnici

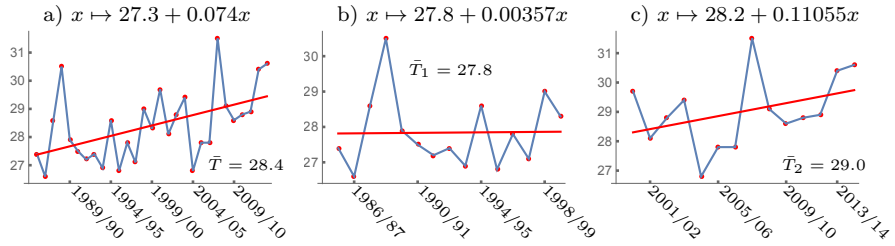
$$\mathcal{A} = \{a_i = a_i(\tau_i) = (\cos\tau_i, \sin\tau_i) \in K : i = 1, \ldots, m\}. \tag{7.6}$$

Svaka točka na Burnovom dijagramu (Slika 7.4) predstavlja prosječnu dnevnu temperaturu u nekom trenutku u godini: kutna pozicija pokazuje vremenski trenutak u godini, a udaljenost točke do ishodišta O pokazuje vrijednost prosječne temperature u °C.

7.2.2 Kretanje maksimalnih godišnjih temperatura

Za svaku od promatranih 29 Sunčevih godina odredit ćemo najvišu godišnju temperaturu i trenutak (pozicija na Burnovom dijagramu) u kome se ona postiže.

Na Slici 7.5a pokazane su te najviše godišnje temperature za grad Osijek u proteklih 29 godina. Uz prosjek od 28.4°C, primjetna je tendencija rasta: koeficijent smjera pripadnog linearnog trenda ($k = 0.074$) pokazuje da je u posljednjih tridesetak godina najviša godišnja temperatura za grad Osijek porasla za približno 2.5°C. Pri tome je porast u prvom periodu (XX) značajno slabije izražen (Slika 7.5b) u odnosu na porast u periodu (XXI) (Slika 7.5c).



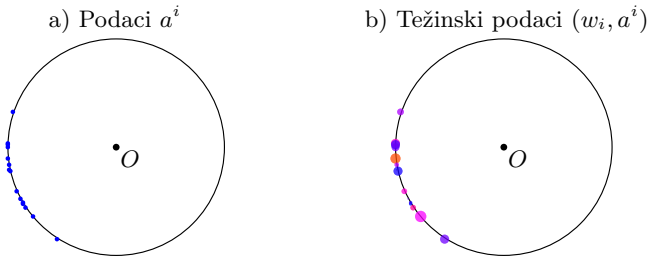
Slika 7.5: Kretanje maksimalnih godišnjih temperatura u Osijeku:
a) (1985/86 – 2013/14), b) (1985/86 – 1999/00), c) (2000/01 – 2013/14)

Vremenske lokacije maksimalnih godišnjih temperatura

Vremenske lokacije maksimalnih godišnjih temperatura odredit ćemo grupiranjem odgovarajućih podataka na jediničnoj kružnici. Neka je

$$\mathcal{A} = \{a^i = (\cos\tau_i, \sin\tau_i) : \tau_i \in [0, 2\pi],\ i = 1, \ldots, m\}$$

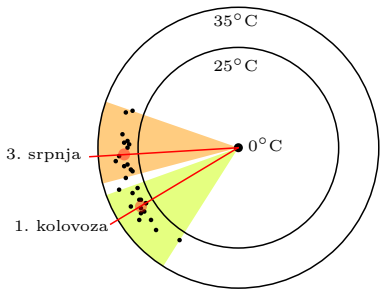
skup točaka na jediničnoj kružnici $K(O, 1)$ (vidi Sliku 7.6a). Svakom elementu skupa $\mathcal{A}$ pridružit ćemo odgovarajuću težinu $w_i > 0$. Ovako definiran skup $\mathcal{A}$ s težinama $w_i > 0$ treba grupirati u više klastera na jediničnoj kružnici.



Slika 7.6: Prikaz podataka na jediničnoj kružnici

Primjedba 7.1. Formalno gledano, skup $\mathcal{A}$ na jediničnoj kružnici predstavljen je točkicama (vidi Sliku 7.6a). Zbog jednostavnosti pretpostavimo da su težine podataka $w_i > 0$ cijeli brojevi. Ako podatku a^i pridružimo težinu $w_i > 0$, onda to možemo shvatiti kao da na mjestu točkica a^i stoji w_i točkica. Jasno je da će prilikom određivanja centra skupa podataka ili grupiranja ovakvih „otežanih" podataka veći utjecaj imati točkice s većim težinama. Na Slici 7.6b težine točkica ilustrirane su odgovarajućom veličinom i bojom.

Za svaku od 29 promatranih Sunčevih godina veličina maksimalne godišnje temperature ($T^{(i)}_{max}$) i trenutak u godini u kome se ona postiže ($t^{(i)}_{max}$) prikazan je crnim točkicama na Burnovom dijagramu na Slici 7.7. Pozicija točkice u odnosu na pozitivni smjer osi x određuje vremenski trenutak u godini, a njena udaljenost od ishodišta reprezentira maksimalnu godišnju temperaturu te godine.



Slika 7.7: Burnov dijagram vremenskih lokacija maksimalnih godišnjih temperatura

Neka je

$$\mathcal{A} = \{a^i = (\cos\tau_i, \sin\tau_i) \colon \tau_i = t^{(i)}_{max},\ i = 1, \ldots, 29\}$$

skup projekcija ovih točkica na jediničnu kružnicu $K(O, 1)$. Svakom elementu skupa $\mathcal{A}$ pridružit ćemo odgovarajuću težinu $w_i = T^{(i)}_{max} > 0$. Ovako definiran skup $\mathcal{A}$ s težinama $w_i > 0$ (kao na Slici 7.6b) treba grupirati u više klastera na jediničnoj kružnici.

Primjenom odgovarajućih indeksa (vidi t. 6, str. 111) potražit ćemo particiju skupa $\mathcal{A}$ s najprikladnijim brojem klastera. Centri ovih klastera označavat će dane u godini s najvišim godišnjim temperaturama tijekom promatranih 29 godina.

Optimalnu particiju $\Pi^\star = \{\pi^\star_1, \ldots, \pi^\star_k\}$ s centrima klastera $c^\star_1, \ldots, c^\star_k$ skupa $\mathcal{A}$ dobit ćemo minimizacijom funkcije (vidi t. 3.3.3 i t. 3.3.4, str. 44–45)

$$F(\tau_1, \ldots, \tau_k) = \sum_{i=1}^{m} w_i \min\{d_K(c_1(\tau_1), a^i(\tau_i)), \ldots, d_K(c_k(\tau_k), a^i(\tau_i))\}, \quad (7.7)$$

gdje je d_K metrička funkcija na jediničnoj kružnici (vidi t. 2.2.5, str. 15). Optimalna vrijednost funkcije cilja (7.7) može se zapisati i u obliku (vidi

t. 3.3.3, str. 44)

$$\mathcal{F}(\tau_1^\star, \ldots, \tau_k^\star) = \sum_{j=1}^{k} \sum_{a^s \in \pi_j^\star} w_s d_K(c_j^\star(\tau_j^\star), a^s(\tau_s)), \tag{7.8}$$

gdje je $c_j^\star(\tau_j^\star) = (\cos(\tau_j^\star), \sin(\tau_j^\star))$, a centri klastera definirani su formulama

$$c_j^\star = \underset{x \in [0,2\pi]}{\operatorname{argmin}} \sum_{a^i \in \pi_j^\star} w_i d_K(x, a^i(\tau_i)), \quad j = 1, \ldots, k. \tag{7.9}$$

Lokalno optimalnu particiju skupa $\mathcal{A}$, čiji elementi imaju težine w_i, možemo potražiti k-means algoritmom opisanim u t. 4.4, str. 82 uz primjenu d_K-metričke funkcije na kružnici definirane u Primjeru 2.12, str. 17. Pri tome za određivanje centara (7.9) i optimizaciju funkcije cilja (7.7) koristit ćemo globalno optimizacijsku metodu `DIRECT` (vidi [20, 21, 23, 46]).

Nakon što se odrede optimalne particije s $k = 2, 3, \ldots$ klastera primjenom `CH` i `DB`-indeksa za podatke na kružnici mogu se odrediti particije s najprihvatljivijim brojem klastera. U svim provedenim izračunima oba indeksa ukazala su na to da je particija s $k = 2$ klastera najprihvatljivija.

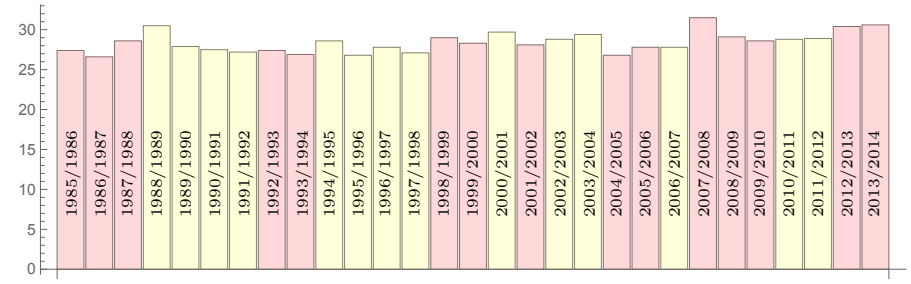
U Tablici 7.1 navedena su osnovna svojstva klastera optimalne particije: centri klastera, broj elemenata po klasteru, maksimalna godišnja temperatura Sunčevih godina u klasteru i vremenski interval u kojemu su se postigle te maksimalne godišnje temperature. Na Slici 7.7 klasteri su prikazani kružnim isječcima različitih boja, a centri klastera crvenim polupravcima.

Klasteri	Centri klastera	Datumi centara	$\lvert\pi_j^\star\rvert$	Prosj. temp. u klasteru	Širina klastera
$\pi_1^\star$	3.2019	03.07.	14	28.3°C	(11.06. – 15.07.)
$\pi_2^\star$	3.6817	01.08.	15	28.5°C	(20.07. – 28.08.)
			$F^\star = 106.48$		

Tablica 7.1: Svojstva klastera optimalne particije

Na Slici 7.8 ilustrirana su dva osnovna svojstva optimalne particije:

- maksimalna godišnja temperatura promatranih Sunčevih godina;
- sastav pojedinih klastera: svjetložutom bojom označene su sve Sunčeve godine prvog klastera $\pi_1^\star$ u kojemu su se maksimalne godišnje temperature postigle od 11. lipnja do 15. srpnja, a svjetlocrvenom Sunčeve godine drugog klastera $\pi_2^\star$ u kojemu su se maksimalne godišnje temperature postigle od 20. srpnja do 28. kolovoza.



Slika 7.8: Sastav klastera optimalne particije prema vremenskom intervalu u kojemu je postignuta maksimalna godišnja temperatura. $\pi_1^\star$: svjetložuta boja (11.06. – 15.07.), $\pi_2^\star$: svjetlocrvena boja (20.07. – 28.08.).

Primjena ℓ_1-metričke i LS-kvazimetričke funkcije

Grupiranje vremenskih lokacija maksimalnih godišnjih temperatura u promatranom razdoblju možemo pokušati napraviti i korištenjem k-means algoritma primjenom ℓ_1-metričke ili LS-kvazimetričke funkcije jer lokacije najviših godišnjih temperatura na Burnovom dijagramu zauzimaju relativno maleni kružni luk.

Neka su $t_{max}^{(i)}$, $i = 1, \ldots, 29$ vremenske lokacije na Burnovom dijagramu maksimalnih godišnjih temperatura $T_{max}^{(i)}$, $i = 1, \ldots, 29$. Nadalje, neka je $\mathcal{A} = \{a^i = t_{max}^{(i)} \in [0, 2\pi] \colon i = 1, \ldots, 29\}$ skup čijim smo elementima pridružili težine $w_i = T_{max}^{(i)} > 0$. Za ovako definiran skup potražit ćemo optimalnu particiju s najprikladnijim brojem klastera primjenom ℓ_1-metričke i LS-kvazimetričke funkcije (vidi t. 3.3, str. 39). Rezultati su prikazani u Tablici 7.2.

Klasteri	Primjena ℓ_1-metričke funkcije			Primjena LS-kvazimetričke funkcije		
	Centri	Datumi	$\lvert\pi_j^\star\rvert$	Centri	Datumi	$\lvert\pi_j^\star\rvert$
$\pi_1^\star$	3.21045	04.07.	14	3.16195	01.07.	14
$\pi_2^\star$	3.68241	01.08.	15	3.70995	03.08.	15
	$F_1^\star = 106.44$			$F_{LS}^\star = 23.29$		

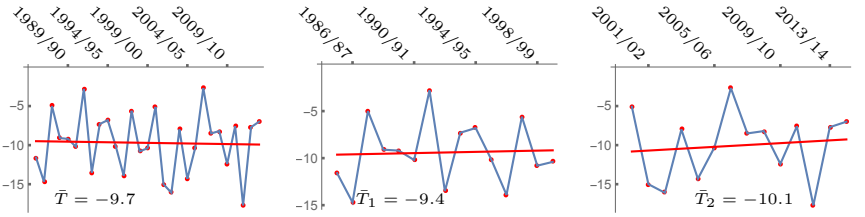
Tablica 7.2: Svojstva klastera ℓ_1-optimalnih i LS-optimalnih particija

Kao što se može vidjeti, primjena ℓ_1-metričke i LS-kvazimetričke funkcije daje rezultate vrlo slične rezultatima grupiranja na Burnovom dijagramu uz primjenu d_K-metričke funkcije (usporedi Tablicu 7.1 i Tablicu 7.2). Gotovo

potpuno podudaranje rezultata dobiva se u slučaju primjene ℓ_1-metričke funkcije.

7.2.3 Kretanje minimalnih godišnjih temperatura

Za svaku Sunčevu godinu odredit ćemo najnižu godišnju temperaturu i trenutak (pozicija na Burnovom dijagramu) u kojemu se ona postiže.



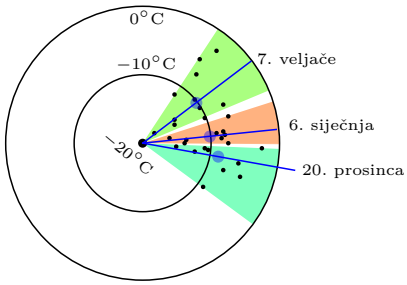
a) $x \mapsto -9.48 - 0.015x$ b) $x \mapsto -9.66 + 0.033x$ c) $x \mapsto -10.9 + 0.11868x$

Slika 7.9: Kretanje minimalnih godišnjih temperatura u Osijeku:
a) (1985/86 – 2013/14), b) (1985/86 – 1999/00), c) (2000/01 – 2013/14)

Na Slici 7.9a pokazane su te najniže godišnje temperature za grad Osijek u proteklih 29 godina. Uz prosjek od -9.7°C, primjetna je tendencija neznatnog pada: koeficijent smjera pripadnog linearnog trenda ($k = -0.015$) pokazuje da je u posljednjih tridesetak godina najniža godišnja temperatura za grad Osijek opala za približno 0.5°C iako je i u prvom (XX), a naročito u drugom periodu (XXI) prisutan porast najnižih godišnjih temperatura (vidi Sliku 7.9b i Sliku 7.9c).

Vremenske lokacije minimalnih godišnjih temperatura

Analogno kao i u slučaju traženja vremenskih lokacija maksimalnih godišnjih temperatura, za svaku Sunčevu godinu veličina najniže godišnje temperature ($T_{min}^{(i)}$) i trenutak u godini u kojemu se ona postiže ($t_{min}^{(i)}$) prikazan je na Burnovom dijagramu (plave točkice na Slici 7.10).



Slika 7.10: Burnov dijagram vremenskih lokacija minimalnih godišnjih temperatura

Neka je

$$\mathcal{A} = \{a^i = (\cos\tau_i, \sin\tau_i) \colon \tau_i = t_{min}^{(i)},\ i = 1, \ldots, 29\}$$

skup projekcija ovih točkica na jediničnu kružnicu $K(O, 1)$. Svakom elementu skupa $\mathcal{A}$ pridružit ćemo odgovarajuću težinu $w_i = |T_{min}^{(i)}| > 0$. Skup $\mathcal{A}$ čije smo elemente opskrbili odgovarajućim težinama grupirat ćemo u više klastera na jediničnoj kružnici. Primjenom odgovarajućih indeksa (vidi t. 6, str. 111) potražit ćemo particiju skupa $\mathcal{A}$ s najprikladnijim brojem klastera. Centri ovih klastera označavat će dane u godini s najnižim godišnjim temperaturama tijekom promatranih 29 godina.

Lokalno optimalnu particiju skupa $\mathcal{A}$, čiji elementi imaju težine w_i, potražit ćemo k-means algoritmom opisanim u t. 4.4, str. 82, uz primjenu d_K-metričke funkcije na kružnici definirane u Primjeru 2.12, str. 17. Pri tome ćemo za određivanje centara (7.9) i optimizaciju funkcije cilja (7.7) koristiti globalno optimizacijsku metodu `DIRECT` (vidi [20, 21, 23, 46]).

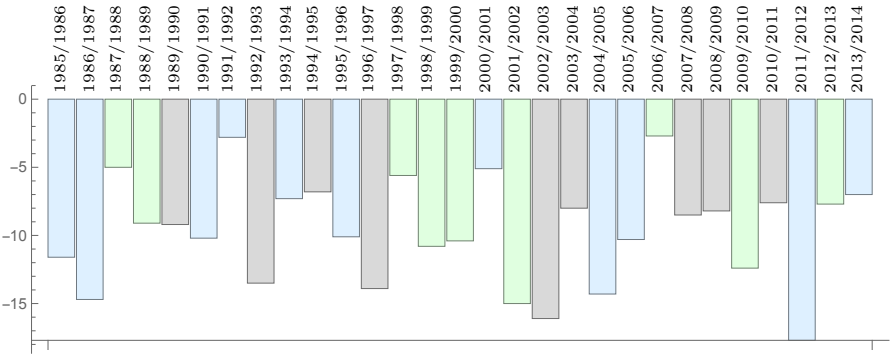
U Tablici 7.3 navedena je vrijednost funkcije cilja $F^\star$ optimalne particije te osnovna svojstva njenih klastera: centri klastera, broj elemenata po klasteru, minimalna godišnja temperatura Sunčevih godina u klasteru i vremenski interval u kojemu su se postigle te minimalne godišnje temperature. Na Slici 7.10 klasteri su prikazani kružnim isječcima različitih boja, a centri klastera plavim polupravcima.

Klasteri	Centri klastera	Datumi centara	$\lvert\pi_j^\star\rvert$	Prosj. temp. u klasteru	Širina klastera
$\pi_1^\star$	6.1065	20.12.	9	$-8.7°$C	(24.11.–28.12.)
$\pi_2^\star$	0.1020	06.01.	9	$-10.2°$C	(30.12.–18.01.)
$\pi_3^\star$	0.6478	07.02.	11	$-10.1°$C	(23.01.–27.02.)
			$F^\star = 31.95$		

Tablica 7.3: Svojstva klastera optimalne particije

Na Slici 7.11 ilustrirana su dva osnovna svojstva optimalne particije:

- minimalna godišnja temperatura promatranih Sunčevih godina;
- sastav pojedinih klastera: svjetlozelenom bojom označene su sve Sunčeve godine prvog klastera $\pi_1^\star$ u kojemu je minimalna godišnja temperatura postignuta od 24. studenog do 28. prosinca, svjetlosivom bojom označene su sve Sunčeve godine drugog klastera $\pi_2^\star$ u kojemu je minimalna godišnja temperatura postignuta od 30. prosinca do 18. siječnja, a svjetloplavom Sunčeve godine trećeg klastera $\pi_3^\star$ u kojemu je minimalna godišnja temperatura postignuta od 23. siječnja do 27. veljače.



Slika 7.11: Sastav klastera optimalne particije prema vremenskom intervalu u kojemu je postignuta minimalna godišnja temperatura. $\pi_1^\star$: svjetlozelena boja (24.11. − 28.12.), $\pi_2^\star$: svjetlosiva boja (30.12. − 18.01.), $\pi_3^\star$: svjetloplava boja (23.01. − 27.02.)

Primjena ℓ_1-metričke i LS-kvazimetričke funkcije

Grupiranje vremenskih lokacija minimalnih godišnjih temperatura u promatranom razdoblju možemo također pokušati napraviti korištenjem k-means algoritma primjenom ℓ_1-metričke ili LS-kvazimetričke funkcije (vidi t. 3.3, str. 39) jer lokacije najnižih godišnjih temperatura na Burnovom dijagramu zauzimaju relativno maleni kružni luk.

Neka su $t_{min}^{(i)}$, $i = 1, \ldots, 29$ vremenske lokacije na Burnovom dijagramu minimalnih godišnjih temperatura $T_{min}^{(i)}$, $i = 1, \ldots, 29$. Nadalje, neka je $\mathcal{A} = \{a^i = t_{min}^{(i)} \in [0, 2\pi] \colon i = 1, \ldots, 29\}$ skup čijim smo elementima pridružili težine $w_i = |T_{min}^{(i)}| > 0$. Za ovako definiran skup potražit ćemo optimalnu particiju s najprikladnijim brojem klastera primjenom ℓ_1-metričke i LS-kvazimetričke funkcije. Rezultati su prikazani u Tablici 7.4.

Klasteri	Primjena ℓ_1-metričke funkcije			Primjena LS-kvazimetričke funkcije		
	Centri	Datumi	$\lvert\pi_j^\star\rvert$	Centri	Datumi	$\lvert\pi_j^\star\rvert$
$\pi_1^\star$	6.17273	24.12.	11	6.10971	20.12.	11
$\pi_2^\star$	0.15493	09.01.	8	0.20268	12.01.	9
$\pi_3^\star$	0.67853	09.02.	10	0.71578	11.02.	9
	$F_1^\star = 31.26$			$F_{LS}^\star = 6.84$		

Tablica 7.4: Svojstva klastera optimalne particije

Kao što možemo vidjeti, primjena ℓ_1-metričke i LS-kvazimetričke funkcije daje rezultate vrlo slične rezultatima grupiranja na Burnovom dijagramu uz primjenu d_K-metričke funkcije (usporedi Tablicu 7.3 i Tablicu 7.4).

7.3 Grupiranje sličnih dana prema temperaturama

Pokažimo još na primjeru podataka o prosječnim dnevnim temperaturama u Osijeku od 1985. do 2014. godine zašto je u klaster analizi važno imati algoritme koji dobro funkcioniraju i u slučaju podataka s velikim brojem obilježja.[5] Podaci s velikim brojem obilježja pojavljuju se primjerice kod ocjenjivanja kvalitete vina, kod istraživanja ljudskog gena, itd. [47, 49].

Pokušat ćemo grupirati dane u godini prema prosječnoj dnevnoj temperaturi u gradu Osijeku od 1985. do 2014. godine. Promatrat ćemo podatke po Sunčevim godinama 1985/86,...,1990/91,1992/93,...,2013/14, gdje smo

[5] Ovaj primjer predložio je recenzent prof. dr. sc. Kristian Sabo.

isključili godinu 1991/92 zbog nepostojanja podataka. Najprije učitamo i prilagodimo podatke:

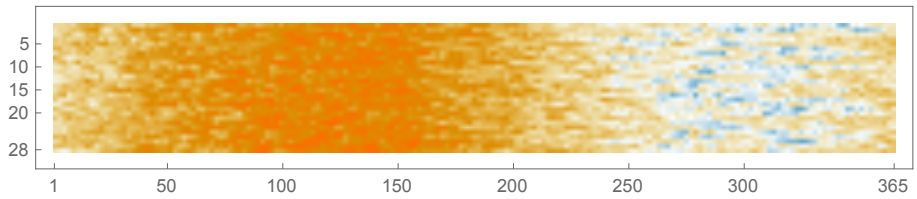
```
In[1]:= SetDirectory[NotebookDirectory[]];
        pod = Import["TOS-1985-2014.txt", "Table"];
In[2]:= mg = 365; A = {}; F = 0; c = Table[0, {j, k}]; k = 2;
        Do[
            m1 = 80 + (god1 - 1985) mg; m2 = m1 + mg - 1;
            A = Append[A, Table[pod[[i,4]],
                          {i,80+(god1-1985)*mg,80+(god1-1985)*mg+mg-1}]]
           ,{god1, 1985, 2013}]
        A = Transpose[Delete[A, {{7}}]];
```

Skup $\mathcal{A}$ ima 365 elemenata (dana u godini), a svaki element (dan) karakterizirat ćemo prosječnom dnevnom temperaturom toga dana u svakoj od promatranih Sunčevih godina. Primjerice, danu koji označavamo s „21. ožujka" pridružit ćemo vektor s 28 komponenti koje redom predstavljaju prosječne dnevne temperature toga dana kroz promatranih 28 Sunčevih godina. Dakle, promatramo skup $\mathcal{A} = \{a^i \in \mathbb{R}^{28} : i = 1, \ldots, 365\}$ preciznije opisan u Tablici 7.5. Strukturu skupa podataka $\mathcal{A}$ možemo grafički prikazati

Dan	Redni broj	Element skupa $\mathcal{A}$	Obilježja (prosječne dnevne temperature)					
			1985/86	...	1990/91	1992/93	...	2013/14
21.03.	1	a^1	T_1^1	...	T_{16}^1	T_{17}^1	...	T_{28}^1
...	...	...	...	...	...	...	...	...
31.12.	286	a^{286}	T_1^{286}	...	T_{16}^{286}	T_{17}^{286}	...	T_{28}^{286}
01.01.	287	a^{287}	T_1^{287}	...	T_{16}^{287}	T_{17}^{287}	...	T_{28}^{287}
...	...	...	...	...	...	...	...	...
20.03.	365	a^{365}	T_1^{365}	...	T_{16}^{365}	T_{17}^{365}	...	T_{28}^{365}

Tablica 7.5: Obilježja elemenata skupa $\mathcal{A}$

kao na Slici 7.12. Svaki od 365 stupaca predstavlja vektor s 28 komponenti, a veličina komponente (prosječna dnevna temperatura) identificirana je bojom (više temperature jačom narančastom, a niže jačom plavom bojom).



Slika 7.12: Skup $\mathcal{A} \in \mathbb{R}^{28}$

Za skup $\mathcal{A} \in \mathbb{R}^{28}$ iz 28-dimenzionalnog prostora (grafički prikazan na Slici 7.12) potražit ćemo particiju s najprikladnijim brojem klastera. U tu svrhu koristit ćemo gotovu *Mathematica*-naredbu `FindCluster[A,k]` za $k = 2, 3, 4, 5$. Po defaultu koristi se LS-kvazimetrička funkcija, a kao output dobivamo optimalnu particiju s k klastera. Nakon toga, za tako dobivenu particiju određujemo centroide njenih klastera, vrijednost kriterijske funkcije cilja $\mathcal{F}_{LS}$, `CH`-index i `DB`-index (vidi Tablicu 7.6).

```
In[3]:= PI = FindClusters[A, k];
        Do[
            c[[j]] = Total[PI[[j]]]/Length[PI[[j]]];
            F = F + Sum[Norm[c[[j]] - PI[[j, s]]]^2, {s, Length[PI[[j]]]}]
           ,{j, k}]
        Print["{n,m}: ", {n, m}, "; F = ", F]
        cc = Table[Sum[A[[i, s]], {i, m}]/365, {s, n}];
        Print["CH(", k ") = ", VCH[PI, c, cc, 2]]
        Print["DB(", k ") = ", VDB[PI, c, 2]]
```

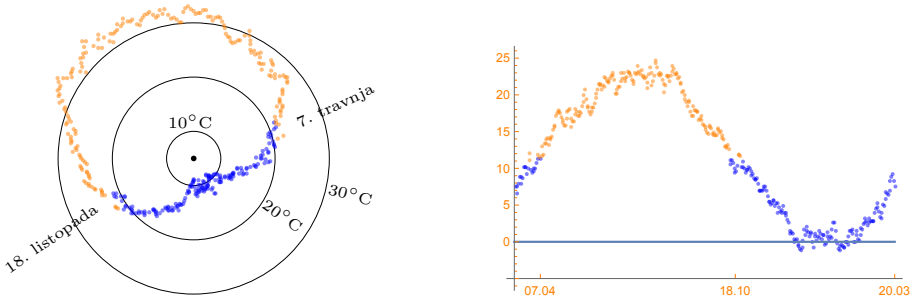
Indikator	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$\mathcal{F}_{LS}$	270 276	198 308	178 036	169 583
`CH`-index	793.7	510.1	431.8	352.7
`DB`-index	0.298	0.769	1.180	0.814

Tablica 7.6: Pokazatelji kompaktnosti i razdvojenosti optimalne particije

Prema rezultatima prikazanim u Tablici 7.6, skup $\mathcal{A}$ najprikladnije je grupirati u dva klastera (177 hladnijih i 188 toplijih dana u godini).

Na Slici 7.13 klaster hladnih dana prikazan je plavim, a klaster toplih dana narančastim točkicama.

a) Burnov dijagram b) Graf



Slika 7.13: Prosječne dnevne temperature hladnih i toplih dana kroz proteklih 28 Sunčevih godina

Na Slici 7.13a prosječne dnevne temperature hladnih i toplih dana kroz proteklih 28 Sunčevih godina prikazane su Burnovim dijagramom, a na Slici 7.13b običnim grafom.

Kao zaključak možemo reći da na bazi podataka o prosječnoj dnevnoj temperaturi u gradu Osijeku u periodu od 1985. do 2014. godine razdoblje toplih dana počinjalo je oko 7. travnja, a završavalo oko 18. listopada, dok su ostatak Sunčevih godina obilježavali hladniji dani. Interesantno je primijetiti da je pri tome u cijeloj godini samo 8 dana imalo negativnu prosječnu temperaturu, ali ne nižu od $-1°\mathrm{C}$.

7.4 Analiza najtoplijeg razdoblja za promatranu godinu

Budući da se najtoplije razdoblje u godini podudara za Sunčevu i kalendarsku godinu, u svrhu proučavanja najtoplijeg razdoblja za neku godinu promatrat ćemo kalendarske godine. Za svaku kalendarsku godinu veličine prosječnih dnevnih temperatura (T_i) i trenuci u godini u kojima su one postignute (t_i) prikazani su na Burnovom dijagramu (plave točkice na slikama 7.14, 7.15, 7.16, 7.17).

Za neku izabranu godinu neka je

$$\mathcal{A} = \{a^i = (\cos t_i, \sin t_i) \colon t_i \in [0, 2\pi],\ i = 1, \ldots, m,\ m < 180\}$$

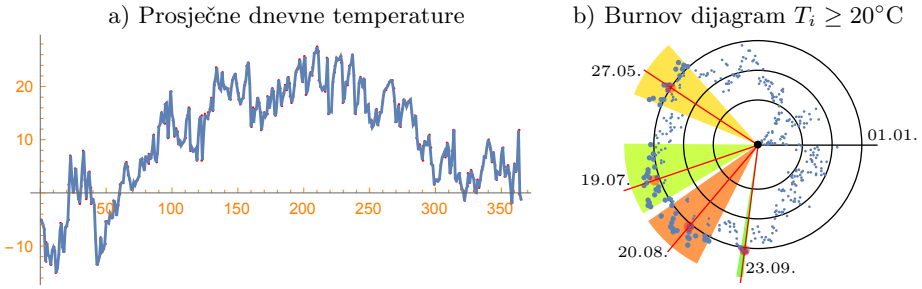
skup točkica na jediničnoj kružnici $K(O, 1)$ koje odgovaraju danima u kojima je prosječna temperatura bila $\geq 20°\mathrm{C}$. Svakom elementu skupa $\mathcal{A}$ pridružit

ćemo odgovarajuću težinu $w_i = T_i > 0$. Ovako definiran skup $\mathcal{A}$ s težinama $w_i > 0$ možemo grupirati u više klastera na jediničnoj kružnici. Na taj način za svaku promatranu godinu dobit ćemo informacije o jednom ili više najtoplijih intervala u toj godini.

Budući da se skup $\mathcal{A}$ nalazi na manje od polovine jedinične kružnice, kao što smo pokazali u t., str. 130, i t., str. 134, za grupiranje ovakvih podataka možemo promatrati obične težinske podatke s jednim obilježjem (w_i, t_i), $i = 1, \ldots, m$ i pri tome koristiti ℓ_1-metričku funkciju.

Godina 1985.

Uzmimo primjerice podatke iz 1985. godine (vidi Sliku 7.14a). Te godine bilo je 63 dana s prosječnom dnevnom temperaturom većom ili jednakom od 20°C (krupnije točkice na Burnovom dijagramu na Slici 7.14b).



Slika 7.14: Prosječne dnevne temperature u Osijeku 1985. godine

Pokazalo se da je ove podatke najprikladnije grupirati u četiri klastera čije su karakteristike prikazane u Tablici 7.7: broj elemenata klastera (koji odgovara broju dana u tom periodu s temperaturom ≥ 20), centar klastera, vremenski interval klastera s ukupnim brojem dana u intervalu klastera (bez obzira na visinu temperature) i prosječna temperatura svih dana u intervalu klastera.

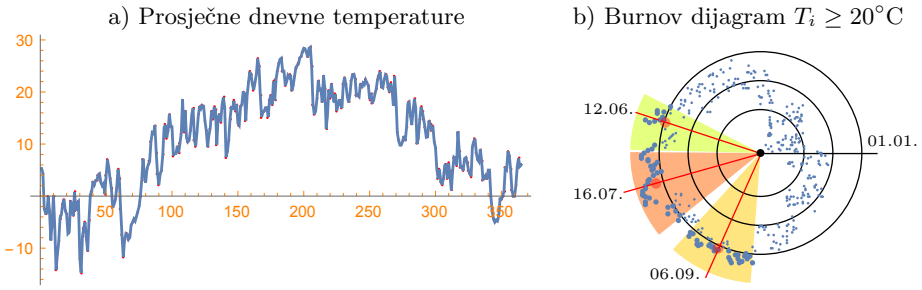
Klaster	$\lvert\pi_j\rvert$	Centar	Interval	Broj dana	Prosj.temp.
$\pi_1^\star$	15	27. svibnja	12. svibnja – 8. lipnja	28	20.3°C
$\pi_2^\star$	23	19. srpnja	30. lipnja – 1. kolovoza	33	21.6°C
$\pi_3^\star$	21	20. kolovoza	5. kolovoza – 3. rujna	30	21.1°C
$\pi_4^\star$	4	23. rujna	21. rujna – 24. rujna	4	20.8°C

Tablica 7.7: Najtopliji vremski intervali 1985. godine

Uočava se da je 1985. godina imala jedno duže, toplo i stabilno razdoblje koje je trajalo od 30. lipnja do 3. rujna s prosječnom dnevnom temperaturom svih dana u tom periodu višom od 21°C i dva kraća topla razdoblja: jedno od 12. svibnja do 8. lipnja s prosječnom dnevnom temperaturom višom od 20°C i drugo sasvim kratko od 21. do 24 rujna s prosječnom dnevnom temperaturom od 21°C.

Godina 1987.

Podaci za 1987. godinu prikazani su na Slici 7.15a. Te godine bilo je 74 dana s prosječnom dnevnom temperaturom većom ili jednakom od 20°C (krupnije točkice na Burnovom dijagramu na Slici 7.15b).



Slika 7.15: Prosječne dnevne temperature u Osijeku 1987. godine

Pokazalo se da je ove podatke najprikladnije grupirati u tri klastera čije su karakteristike navedene u Tablici 7.8.

Klaster	$\lvert\pi_j\rvert$	Centar	Interval	Broj dana	Prosj.temp.
$\pi_1^\star$	13	12. lipnja	3. lipnja – 29. lipnja	27	20.0°C
$\pi_2^\star$	33	16. srpnja	30. lipnja – 9. kolovoza	41	22.8°C
$\pi_3^\star$	28	6. rujna	18. kolovoza – 26. rujna	40	21.0°C

Tablica 7.8: Najtopliji vremenski intervali 1987. godine

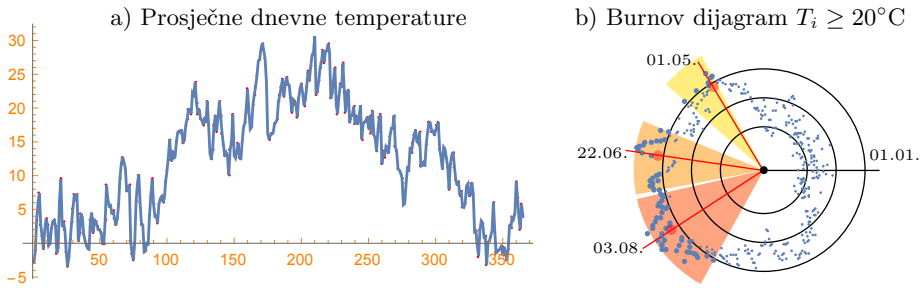
Uočava se da je i 1987. godina imala jedno dugo, toplo i stabilno razdoblje koje je trajalo od 3. lipnja do 9. kolovoza. U prvoj trećini ovog razdoblja prosječna dnevna temperatura bila je nešto viša od 20°C, a u ostatku razdoblja prosječna dnevna temperatura bila je 23°C. Drugo toplo razdoblje te godine trajalo je od 18. kolovoza do 26. rujna s visokom prosječnom dnevnom

temperaturom od 21°C.

Godina 2013.

Podaci za 2013. godinu prikazani su na Slici 7.16a. Te godine bilo je 70 dana s prosječnom dnevnom temperaturom većom ili jednakom od 20°C (krupnije točkice na Burnovom dijagramu na Slici 7.16b).

Za 2013. godinu skup podataka $\mathcal{A}$ najprikladnije je grupirati u tri klastera čije su karakteristike vidljive u Tablici 7.9.



Slika 7.16: Najtopliji vremenski intervali 2013. godine

Klaster	$\lvert\pi_j\rvert$	Centar	Interval	Broj dana	Prosj.temp.
$\pi_1^\star$	6	1. svibnja	29. travnja – 19. svibnja	21	18.5°C
$\pi_2^\star$	23	22. lipnja	8. lipnja – 1. srpnja	34	21.8°C
$\pi_3^\star$	41	3. kolovoza	13. srpnja – 1. rujna	51	23.0°C

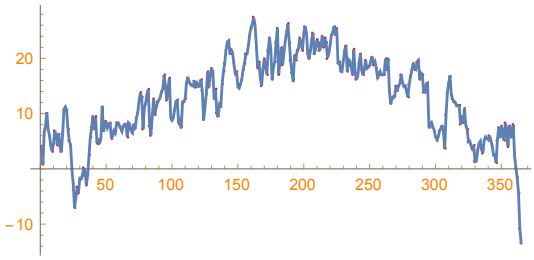
Tablica 7.9: Najtopliji vremenski intervali 2013. godine

Uočava se da je i 2013. godina imala jedno dugo, toplo i stabilno razdoblje koje je trajalo od 8. lipnja do 1. rujna s prosječnom dnevnom temperaturom između 22°C i 23°C. Drugo toplo razdoblje te godine bilo je u proljeće i trajalo je od 29. travnja do 19. svibnja s prosječnom dnevnom temperaturom od 18.5°C.

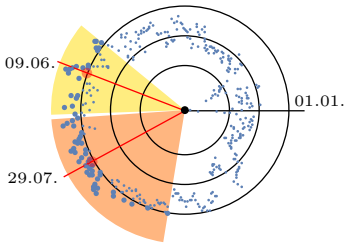
Godina 2014.

Podaci za 2014. godinu prikazani su na Slici 7.17a. Te godine bilo je 63 dana s prosječnom dnevnom temperaturom većom ili jednakom od 20°C (krupnije točkice na Burnovom dijagramu na Slici 7.17b).

a) Prosječne dnevne temperature



b) Burnov dijagram $T_i \geq 20°\text{C}$



Slika 7.17: Prosječne dnevne temperature u Osijeku 2014. godine

Za 2014. godinu skup podataka $\mathcal{A}$ najprikladnije je grupirati u dva klastera čije su karakteristike vidljive u Tablici 7.10.

Klaster	$\lvert\pi_j\rvert$	Centar	Interval	Broj dana	Prosj.temp.
$\pi_1^\star$	42	9. lipnja	21. svibnja – 2. srpnja	43	20.2°C
$\pi_2^\star$	21	29. srpnja	4. srpnja – 21. rujna	80	20.6°C

Tablica 7.10: Najtopliji vremenski intervali 2014. godine

Uočava se da je 2014. godina imala jedno dugo, toplo i stabilno razdoblje koje je trajalo od 21. svibnja do 21. rujna. Prvih 43 dana do 2. srpnja prosječna dnevna temperatura bila je 20.2°C, a nakon toga čak 80 dana prosječna dnevna temperatura bila je 20.6°C. Najviša prosječna dnevna temperatura 2014. godine bila je već 11. lipnja i iznosila je 27.4°C.

Poglavlje 8

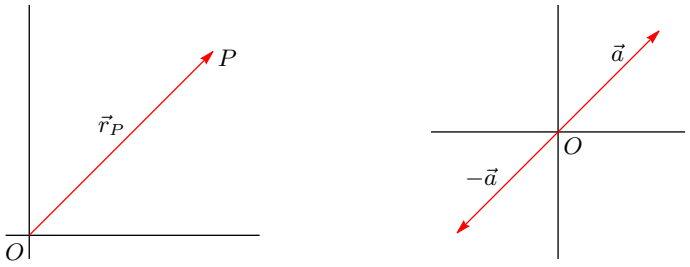
Dodatak: Vektori

8.1 Vektori u ravnini

Ako u ravninu M uvedemo pravokutni koordinatni sustav s ishodištem u točki $O \in M$, onda svakoj točki $P \in M$ pripada jedinstveni vektor $\overrightarrow{OP}$, koji zovemo radijvektor ili vektor položaja i označavamo ga s

$$\vec{r}_P = \overrightarrow{OP}.$$

Skup svih takvih radijvektora označit ćemo s $X_0(M)$. Očigledno postoji bijekcija (obostrano jednoznačno preslikavanje) između skupova M i $X_0(M)$. Nadalje, radijvektore iz $X_0(M)$ jednostavno ćemo zvati vektori.



Slika 8.1: Vektor i suprotni vektor

Svakom vektoru $\overrightarrow{OP}$ možemo pridružiti njegovu normu (modul) kao duljinu dužine $\overline{OP}$. Normu vektora $\vec{r}$ označavat ćemo s $\|\vec{r}\|$.

Primjer 8.1. *Navedimo nekoliko primjera.*
(a) Nulvektor $\vec{0}$ vektor je kojemu je $O \in M$ ujedno i početna i završna točka;

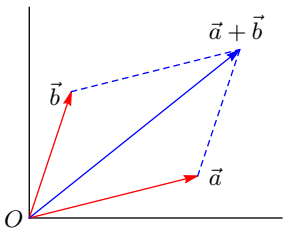
(b) Kažemo da je $\vec{e}$ jedinični vektor ako je $\|\vec{e}\| = 1$;
(c) Suprotni vektor vektora $\vec{a}$ vektor je koji ima suprotnu orijentaciju od orijentacije vektora $\vec{a}$ i označavamo ga s $(-\vec{a})$ (vidi Sliku 8.1). Za njega vrijedi $\|(-\vec{a})\| = \|\vec{a}\|$.

8.1.1 Računske operacije s vektorima

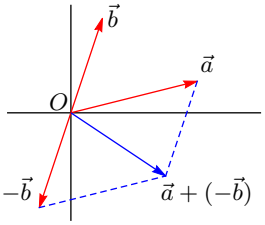
Zbrajanje vektora

Zbrajanje vektora je binarna operacija $+ : X_0(M) \times X_0(M) \to X_0(M)$. Za dva vektora $\vec{a}, \vec{b} \in X_0(M)$ definiramo novi vektor $\vec{c} := \vec{a} + \vec{b}$ pravilom paralelograma. Vektor $\vec{c}$ određen je dijagonalom paralelograma sa stranicama $\vec{a}$ i $\vec{b}$ (vidi Sliku 8.2a).

(a) Zbrajanje vektora



(b) Oduzimanje vektora



Slika 8.2: Zbrajanje vektora

Primjer 8.2. *Da bi definicija zbrajanja vektora bila potpuna, moramo također definirati i zbroj nulvektora $\vec{0}$ i nekog vektora $\vec{a}$, tako da vrijedi*

$$\vec{a} + \vec{0} = \vec{a}.$$

Primjer 8.3. *Zbrajanje vektora $\vec{a}$ i vektora $(-\vec{b})$, suprotnog vektoru $\vec{b}$, zvat ćemo „oduzimanje vektora"*

$$\vec{a} - \vec{b} := \vec{a} + (-\vec{b}),$$

i provesti također ranije spomenutim pravilom paralelograma (vidi Sliku 8.2b). Specijalno vrijedi: $\vec{a} - \vec{a} := \vec{a} + (-\vec{a}) = \vec{0}$.

Primjedba 8.1. Računska operacija zbrajanja vektora ima svojstvo zatvorenosti (ili grupoidnosti), tj. rezultat operacije zbrajanja dva vektora opet je jedan vektor. Osim toga, vrijedi

(i) za svaka tri vektora $\vec{a}, \vec{b}, \vec{c} \in X_0(M)$ vrijedi svojstvo asocijativnosti: $(\vec{a} + \vec{b}) + \vec{c} = \vec{a} + (\vec{b} + \vec{c})$;

(ii) postoji neutralni element $\vec{0}$, tako da za proizvoljni vektor $\vec{a} \in X_0(M)$ vrijedi: $\vec{a} + \vec{0} = \vec{a}$;

(iii) za svaki vektor $\vec{a} \in X_0(M)$ postoji inverzni element (suprotni vektor) $(-\vec{a})$, takav da vrijedi: $\vec{a} + (-\vec{a}) = \vec{0}$;

(iv) vrijedi zakon komutacije, tj. za svaka dva vektora $\vec{a}, \vec{b} \in X_0(M)$ vrijedi: $\vec{a} + \vec{b} = \vec{b} + \vec{a}$;

Skup svih vektora u ravnini snabdjeven računskom operacijom zbrajanja i svojstvima (i) – (iii) nazivamo aditivna grupa, a ako dodamo i svojstvo (iv), takvu strukturu nazivamo komutativna ili Abelova grupa[1] i označavamo ju s $(X_0(M), +)$.

Primjer 8.4. *U skladu s uvedenom definicijom zbrajanja vektora i Primjedbom 8.1, možemo induktivno definirati množenje vektora prirodnim brojem:*

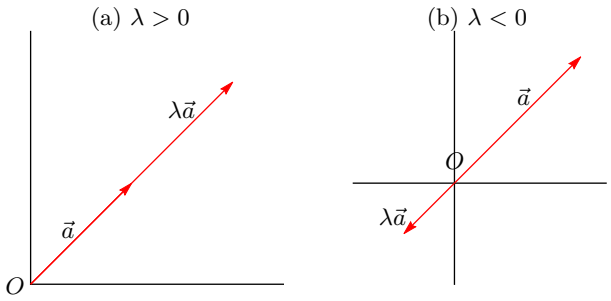
$$1 \cdot \vec{a} = \vec{a}, \qquad 2 \cdot \vec{a} = \vec{a} + \vec{a}, \qquad \ldots \qquad n \cdot \vec{a} = (n-1) \cdot \vec{a} + \vec{a}, \ldots .$$

Množenje vektora skalarom

Definicija množenja vektora prirodnim brojem navedena u Primjeru 8.4 na prirodan način proširuje se na množenje vektora realnim brojem (skalarom). Množenje vektora skalarom[2] je preslikavanje $\cdot : \mathbb{R} \times X_0(M) \to X_0(M)$.

[1] Niels Abel (1802. – 1829.), norveški matematičar.

[2] Naziv „skalar" dolazi iz fizike gdje se tradicionalno najčešće korištene fizikalne veličine dijele na vektore i skalare.



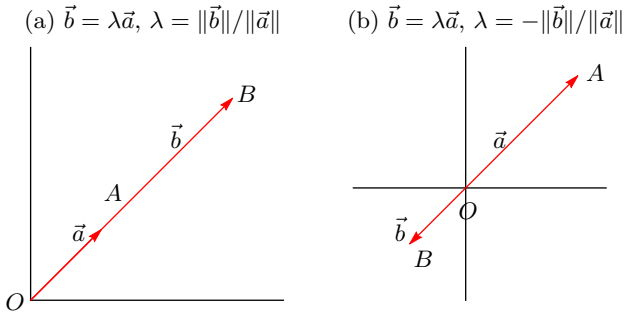
Slika 8.3: Množenje vektora skalarom

Za realni broj (skalar) $\lambda \in \mathbb{R}$ i vektor $\vec{a} \in X_0(M)$ definiramo novi vektor $\vec{b} := \lambda \cdot \vec{a}$ kao na Slici 8.3. Duljina novog vektora $\vec{b}$ definirana je s

$$\|\vec{b}\| = |\lambda| \cdot \|\vec{a}\|.$$

Primjedba 8.2. Kažemo da su dva nenul-vektora $\vec{a} = \overrightarrow{OA}$, $\vec{b} = \overrightarrow{OB}$ kolinearna ako leže na istom pravcu p koji prolazi kroz ishodište O. Obrnuto, ako vektori $\vec{a}, \vec{b}$ leže na pravcu p koji prolazi kroz ishodište O, onda postoje točke $A, B \in p$ takve da je $\vec{a} = \overrightarrow{OA}$, $\vec{b} = \overrightarrow{OB}$, pri čemu vrijedi (vidi Sliku 8.4)

$$\vec{b} = \lambda\vec{a}, \quad \text{gdje je} \quad \lambda = \begin{cases} \|\vec{b}\|/\|\vec{a}\|, & \text{ako } \vec{a}, \vec{b} \text{ imaju isti smjer} \\ -\|\vec{b}\|/\|\vec{a}\|, & \text{ako su } \vec{a}, \vec{b} \text{ suprotnog smjera} \end{cases}$$



Slika 8.4: Kolinearni vektori

Primjedba 8.3. U kontinuitetu sa svojstvima koja vrijede za zbrajanje vektora (Primjedba 8.1), navedimo i svojstva koja vrijede za množenje vektora skalarom:

(v) za svaka dva vektora $\vec{a}, \vec{b} \in X_0(M)$ i svaki skalar $\lambda \in \mathbb{R}$ vrijedi distributivnost s obzirom na vektorski faktor: $\lambda(\vec{a} + \vec{b}) = \lambda\vec{a} + \lambda\vec{b}$;

(vi) za svaka dva skalara $\lambda, \mu \in \mathbb{R}$ i svaki vektor $\vec{a} \in X_0(M)$ vrijedi distributivnost s obzirom na skalarni faktor: $(\lambda + \mu)\vec{a} = \lambda\vec{a} + \mu\vec{a}$;

(vii) za svaka dva skalara $\lambda, \mu \in \mathbb{R}$ i svaki vektor $\vec{a} \in X_0(M)$ vrijedi svojstvo kvaziasocijativnosti: $(\lambda\mu)\vec{a} = \lambda(\mu\vec{a})$;

(viii) za svaki vektor $\vec{a} \in X_0(M)$ vrijedi: $1 \cdot \vec{a} = \vec{a}$.

Skup $X_0(M)$ snabdjeven računskim operacijama zbrajanja i množenja sa skalarom, koje imaju navedenih osam svojstava nazivamo vektorski prostor i označavamo ga s $(X_0(M), +, \cdot)$.

Analogno se definiraju i vektorski prostor $(X_0(E), +, \cdot)$ u prostoru i vektorski prostor $(X_0(p), +, \cdot)$ na pravcu. Kako je $X_0(M) \subset X_0(E)$, reći ćemo da je vektorski prostor $(X_0(M), +, \cdot)$ vektorski potprostor u $(X_0(E), +, \cdot)$. Nadalje ćemo ove vektorske prostore jednostavno označavati samo s $X_0(E)$, $X_0(M)$, $X_0(p)$.

8.1.2 Linearna zavisnost i nezavisnost vektora

Razmotrimo još jednom Primjedbu 8.2. Tako definiran pojam kolinearnosti očigledno se ne može primijeniti za nulvektor $\vec{a} = \vec{0}$. S druge strane, nulvektor $\vec{0}$ očigledno je kolinearan sa svim drugim vektorima $\vec{b} \in X_0(M)$ jer možemo pisati $\vec{0} = 0 \cdot \vec{b}$. Kako bismo i taj slučaj uključili u našu algebarsku definiciju kolinearnosti, promatrat ćemo tzv. linearnu kombinaciju $\lambda\vec{a} + \mu\vec{b}$ dva proizvoljna vektora $\vec{a}, \vec{b} \in X_0(M)$.

Primijetite da je jednadžbu

$$\lambda\vec{a} + \mu\vec{b} = \vec{0}, \tag{8.1}$$

s nepoznanicama λ, μ moguće zadovoljiti na dva bitno različita načina ovisno o međusobnom položaju vektora $\vec{a}, \vec{b}$:

1. Jednadžba (8.1) uvijek ima barem tzv. trivijalno rješenje $\lambda = \mu = 0$. Ako je trivijalno rješenje i jedino rješenje jednadžbe (8.1), nijedan od vektora nije moguće prikazati pomoću onog drugog. Kažemo da vektori *nisu kolinearni*;

2. Druga mogućnost je da pored trivijalnog rješenja jednadžbe (8.1) postoji i neko netrivijalno rješenje. Pretpostavimo da je pri tome primjerice, $\lambda \neq 0$. Tada možemo pisati $\vec{a} = -\frac{\mu}{\lambda}\vec{b}$, iz čega prema Primjedbi 8.2 zaključujemo da su vektori $\vec{a}, \vec{b} \in X_0(M)$ kolinearni.

Propozicija 8.1. *Dva vektora $\vec{a}, \vec{b} \in X_0(M)$ kolinearna su onda i samo onda ako jednadžba (8.1) ima netrivijalno rješenje (barem jedan od skalara $\lambda, \mu \in \mathbb{R}$ različit je od nule).*

Obrnuto, dva vektora $\vec{a}, \vec{b} \in X_0(M)$ nisu kolinearna onda i samo onda ako jednadžba (8.1) ima samo trivijalno rješenje $\lambda = \mu = 0$.

Primjer 8.5. *Nulvektor $\vec{a} = \vec{0}$ kolinearan je s bilo kojim drugim vektorom $\vec{b} \in X_0(M)$ u smislu Propozicije 8.1. Naime, za $\lambda \neq 0$ vrijedi*

$$\lambda \vec{0} + 0\, \vec{b} = \vec{0}.$$

Sukladno Propoziciji 8.1, pojam kolinearnosti dva vektora $\vec{a}, \vec{b} \in X_0(M)$ može se poopćiti na više vektora iz vektorskog prostora $X_0(M)$. Umjesto termina „kolinearnosti", govorit ćemo o linearnoj zavisnosti vektora.

Definicija 8.1. Kažemo da je skup vektora $\vec{a}_1, \ldots, \vec{a}_n \in X_0$ linearno nezavisan ako njihova linearna kombinacija $\lambda_1 \vec{a}_1, \ldots, \lambda_n \vec{a}_n$ iščezava jedino na trivijalan način: $\lambda_1 = \cdots = \lambda_n = 0$. U protivnom, kažemo da je skup vektora $\vec{a}_1, \ldots, \vec{a}_n \in X_0$ linearno zavisan.

Zadatak 8.1. Precizirajte definiciju linearne zavisnosti skupa vektora $\vec{a}_1, \ldots, \vec{a}_n \in X_0$.

Primjer 8.6. *Skup vektora $\vec{a}_1, \ldots, \vec{a}_n$ koji sadržava nulvektor linearno je zavisan. Usporedite ovu tvrdnju s tvrdnjom iz Primjera 8.5.*

U svrhu dokaza ove tvrdnje bez smanjenja općenitosti možemo pretpostaviti da je baš prvi vektor $\vec{a}_1$ nulvektor. Tada vrijedi

$$1 \cdot \vec{0} + 0 \cdot \vec{a}_2 + \cdots + 0 \cdot \vec{a}_n = \vec{0}.$$

Na taj način pronašli smo jednu linearnu kombinaciju vektora $\vec{a}_1, \ldots, \vec{a}_n$, koja iščezava na netrivijalan način, što znači da je promatrani skup vektora linearno zavisan.

Sljedeći teorem ukazuje nam na jedan operativniji način na koji možemo ustanoviti je li neki skup vektora linearno zavisan ili nezavisan.

Teorem 8.1. *Skup vektora $\vec{a}_1, \ldots, \vec{a}_n \in X_0$ linearno je zavisan onda i samo onda ako se barem jedan od njih može prikazati kao linearna kombinacija ostalih.*

Dokaz. (Nužnost) Pretpostavimo da je skup vektora $\vec{a}_1, \ldots, \vec{a}_n \in X_0$ linearno zavisan. Prema Definiciji 8.1 to znači da postoji njihova linearna kombinacija koja iščezava na netrivijalan način. Bez smanjenja općenitosti pretpostavimo da je $\lambda_1\vec{a}_1 + \cdots + \lambda_n\vec{a}_n = \vec{0}$, a da je pri tome $\lambda_1 \neq 0$. Tada možemo pisati

$$\vec{a}_1 = \left(-\tfrac{\lambda_2}{\lambda_1}\right)\vec{a}_2 + \cdots + \left(-\tfrac{\lambda_n}{\lambda_1}\right)\vec{a}_n.$$

(Dovoljnost) Bez smanjenja općenitosti možemo pretpostaviti da je $\vec{a}_1 = \beta_2\vec{a}_2 + \cdots + \beta_n\vec{a}_n$ iz čega slijedi

$$1 \cdot \vec{a}_1 + (-\beta_2)\vec{a}_2 + \cdots + (-\beta_n)\vec{a}_n = \vec{0}.$$

Po Definiciji 8.1 to znači da su vektori $\vec{a}_1, \ldots, \vec{a}_n \in X_0$ linearno zavisni. □

Primjedba 8.4. Primijetite da su Definicija 8.1 i Teorem 8.1 neosjetljivi na to odakle dolaze vektori $\vec{a}_1, \ldots, \vec{a}_n$. Oni mogu biti iz vektorskog prostora na pravcu $X_0(p)$, iz vektorskog prostora u ravnini $X_0(M)$, iz vektorskog prostora $X_0(E)$, ali mogu biti i iz apstraktnog vektorskog prostora $\mathbb{R}^n$ svih uređenih n-torki realnih brojeva.

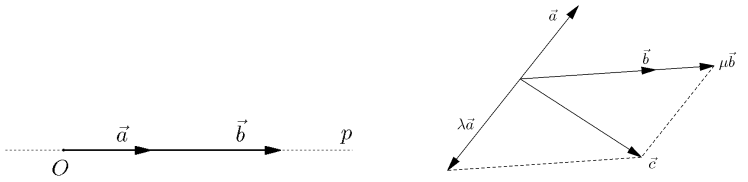
Maksimalni broj linearno nezavisnih vektora u nekom vektorskom prostoru X_0 zovemo dimenzija tog vektorskog prostora i pišemo $\dim X_0$.

Primjer 8.7. *Vrijedi (vidi Sliku 8.5)*

(i) bilo koja dva vektora $\vec{a}, \vec{b} \in X_0(p)$ na pravcu linearno su zavisna, tj. maksimalni broj linearno nezavisnih vektora u vektorskom prostoru $X_0(p)$ na pravcu je jedan ($\dim X_0(p) = 1$);

(ii) bilo koja tri vektora $\vec{a}, \vec{b}, \vec{c} \in X_0(M)$ u ravnini linearno su zavisna, tj. maksimalni broj linearno nezavisnih vektora u vektorskom prostoru $X_0(M)$ u ravnini je dva ($\dim X_0(M) = 2$);

(iii) bilo koja četiri vektora $\vec{a}, \vec{b}, \vec{c}, \vec{d} \in X_0(E)$ u prostoru E linearno su zavisna, tj. maksimalni broj linearno nezavisnih vektora u vektorskom prostoru $X_0(E)$ je tri ($\dim X_0(E) = 3$).



Slika 8.5: Maksimalni broj linearno nezavisnih vektora iz $X_0(p)$ i iz $X_0(M)$

Zadatak 8.2. Neka su $\vec{a}, \vec{b} \in X_0(M)$ dva linearno nezavisna vektora u ravnini. Pokažite da se tada svaki vektor $\vec{c} \in X_0(M)$ na jedinstven način može prikazati (rastaviti) kao linearna kombinacija vektora $\vec{a}, \vec{b}$.

Zadatak 8.3. Neka su $\vec{a}, \vec{b}, \vec{c} \in X_0(E)$ tri linearno nezavisna vektora u prostoru. Pokažite da se tada svaki vektor $\vec{d} \in X_0(E)$ na jedinstven način može prikazati (rastaviti) kao linearna kombinacija vektora $\vec{a}, \vec{b}, \vec{c}$.

8.1.3 Baza vektorskog prostora $X_0(M)$. Koordinatni sustav

Definicija 8.2. Uređen par $(\vec{e}_1, \vec{e}_2)$ linearno nezavisnih vektora iz $X_0(M)$ zovemo baza vektorskog prostora $X_0(M)$.

Neka je $\vec{a} \in X_0(M)$ proizvoljni vektor, a $(\vec{e}_1, \vec{e}_2)$ baza u $X_0(M)$. Tada vektor $\vec{a}$ na jedinstven način možemo zapisati kao linearnu kombinaciju baznih vektora (vidi Zadatak 8.3)

$$\vec{a} = a_1\vec{e}_1 + a_2\vec{e}_2.$$

Brojeve a_1, a_2 zovemo koordinate (komponente) vektora $\vec{a}$ u bazi $(\vec{e}_1, \vec{e}_2)$. Primijetite da je vektor $\vec{a}$ potpuno određen uređenim parom svojih komponenti $(a_1, a_2) \in \mathbb{R}^2$ i da zbog toga postoji bijekcija između skupa svih vektora $X_0(M)$ u ravnini M i svih uređenih parova brojeva iz $\mathbb{R}^2$, a osnovne računske operacije zbrajanja vektora i množenje vektora skalarom prirodno se prenose:

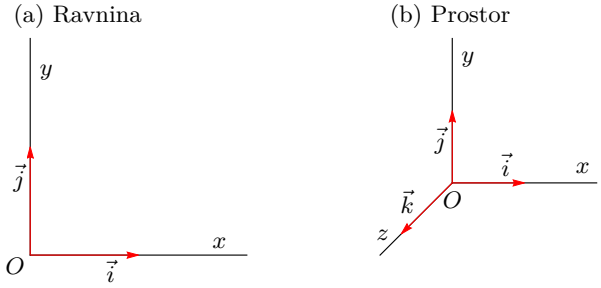
$$\vec{a} = a_1\vec{e}_1 + a_2\vec{e}_2, \qquad \vec{b} = b_1\vec{e}_1 + b_2\vec{e}_2,$$

$$\vec{a} + \vec{b} = (a_1 + b_1)\vec{e}_1 + (a_2 + b_2)\vec{e}_2, \quad \text{[zbrajanje]}$$

$$\lambda\vec{a} = (\lambda a_1)\vec{e}_1 + (\lambda a_2)\vec{e}_2. \quad \text{[množenje vektora skalarom]}$$

Definicija 8.3. Par $(O\,;(\vec{e}_1,\vec{e}_2))$ fiksne točke O i baze $(\vec{e}_1,\vec{e}_2)$ zovemo Kartezijev[3] koordinatni sustav u ravnini M.

Posebno je pogodno ako za bazu prostora $X_0(M)$ izaberemo uređen par međusobno okomitih (zatvaraju kut od 90°) i jediničnih vektora, koje obično označavamo s $(\vec{i},\vec{j})$ i zovemo ortonormirana baza u $X_0(M)$. Tako dobivamo pravokutni Kartezijev koordinatni sustav $(O;\vec{i},\vec{j})$. Pravac određen vektorom $\vec{i}$ označavamo s x i zovemo apscisa, a pravac određen vektorom $\vec{j}$ označavamo s y i zovemo ordinata (vidi Sliku 8.6a).



Slika 8.6: Pravokutni Kartezijev (Descartesov) koordinatni sustav

Zadatak 8.4. Provjerite čine li vektori $\vec{a}=3\vec{i}+2\vec{j}$, $\vec{b}=-\vec{i}+2\vec{j}$ bazu u vektorskom prostoru $X_0(M)$. Ako čine, vektor $\vec{c}=-11\vec{i}+6\vec{j}$ prikažite u toj bazi.

Rješenje: Čine, $\vec{c}=-2\vec{a}+5\vec{b}$.

Zadatak 8.5. Neka je $O\in M$ fiksna točka u ravnini M i neka točka $C\in M$ dijeli dužinu $\overline{AB}$ u omjeru $3:1$, tj. $d(A,C):d(C,B)=3:1$.

Uputa: Vektor $\overrightarrow{OC}$ prikažite kao linearnu kombinaciju vektora $\overrightarrow{OA}$ i $\overrightarrow{OB}$. Može li se sve promatrati i u prostoru E?

Rješenje: $\overrightarrow{OC}=\frac{1}{4}\overrightarrow{OA}+\frac{3}{4}\overrightarrow{OB}$.

8.1.4 Vektor kao uređeni par realnih brojeva

Kao što smo primijetili u prethodnoj točki, postoji bijekcija između skupa svih točaka u ravnini M i skupa svih uređenih parova realnih brojeva (a_1,a_2), koji ćemo označavati s $\mathbb{R}^2:=\mathbb{R}\times\mathbb{R}=\{(a_1,a_2)\colon a_1,a_2\in\mathbb{R}\}$. Zato

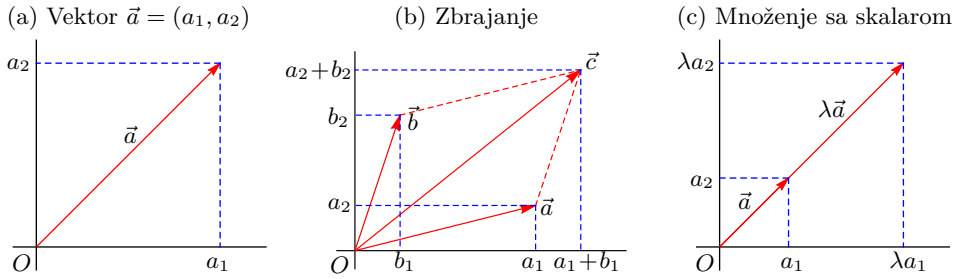
[3]Rene Descartes (1596. – 1650.), francuski filozof i matematičar. Njegovo latinizirano ime je Cartesius.

skup točaka u ravnini M možemo identificirati sa skupom svih uređenih parova realnih brojeva $\mathbb{R}^2$ i pisati $A = (a_1, a_2)$. Također, točka $A \in M$, te onda posredno i radijvektor $\vec{a} = \overrightarrow{OA}$ potpuno su određeni uređenim parom (a_1, a_2), pa također i skup svih vektora $X_0(M)$ možemo identificirati sa skupom svih uređenih parova realnih brojeva $\mathbb{R}^2$ i pisati $\vec{a} = (a_1, a_2)$.

Ranije uvedene računske operacije zbrajanja dva vektora i množenja vektora skalarom mogu se interpretirati na prostoru uređenih parova $\mathbb{R}^2$.

Razmotrimo najprije kako se pravilo paralelograma za zbrajanje dva vektora $\vec{a} = \overrightarrow{OA}$, $\vec{b} = \overrightarrow{OB}$ prenosi na prostor uređenih parova $\mathbb{R}^2$. Neka su $\vec{a} = (a_1, a_2)$, $\vec{b} = (b_1, b_2)$ vektori. Iz Slike 8.7b vidljivo je da je vektor $\vec{c} = \vec{a} + \vec{b}$ definiran s

$$\vec{c} = \vec{a} + \vec{b} = (a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2). \tag{8.2}$$



Slika 8.7: Računske operacije na $\mathbb{R}^2$

Lako je provjeriti da skup $\mathbb{R}^2$ snabdjeven računskom operacijom zbrajanja definiranom s (8.2) ima analogna svojstva (i) – (iv) iz Primjedbe 8.1.

Slično se i množenje vektora skalarom može interpretirati na prostoru uređenih parova $\mathbb{R}^2$. Neka je $\vec{a} = (a_1, a_2)$ i $\lambda \in \mathbb{R}$. Iz Slike 8.7c vidljivo je da vrijedi

$$\lambda\vec{a} = \lambda(a_1, a_2) = (\lambda a_1, \lambda a_2). \tag{8.3}$$

Općenito, neka su zadana dva vektora $\vec{a} = (a_1, a_2)$, $\vec{b} = (b_1, b_2)$ i dva skalara $\lambda, \mu \in \mathbb{R}$. Tada vrijedi

$$\lambda\vec{a} + \mu\vec{b} = \lambda(a_1, a_2) + \mu(b_1, b_2) = (\lambda a_1 + \mu b_1, \lambda a_2 + \mu b_2).$$

Zadatak 8.6. Zadani su vektori $\vec{a} = (1, 4)$, $\vec{b} = (-2, 2)$ u pravokutnom Kartezijevom koordinatnom sustavu. Nacrtajte vektore $\vec{a} + \vec{b}$ i $\vec{a} - 2\vec{b}$.

8.2 Vektori u prostoru

Slično kao u ravnini, i u vektorskom prostoru $X_0(E)$ definiramo bazu kao uređenu trojku $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ linearno nezavisnih vektora iz $X_0(E)$. Tada proizvoljni vektor $\vec{a} \in X_0(E)$ na jedinstven način možemo zapisati kao linearnu kombinaciju vektora baze

$$\vec{a} = a_1\vec{e}_1 + a_2\vec{e}_2 + a_3\vec{e}_3,$$

pri čemu su brojevi a_1, a_2, a_3 koordinate (komponente) vektora $\vec{a}$ u bazi $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$.

Kartezijev koordinatni sustav u prostoru $X_0(E)$ je par $(O\,;(\vec{e}_1, \vec{e}_2, \vec{e}_3))$ fiksne točke O i baze $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$, a pravokutni Kartezijev koordinatni sustav $(O; \vec{i}, \vec{j}, \vec{k})$ u prostoru $X_0(E)$ određen je fiksnom točkom O i uređenom trojkom međusobno okomitih i jediničnih vektora, koje označavamo s $(\vec{i}, \vec{j}, \vec{k})$ i zovemo ortonormirana baza u $X_0(E)$. Kao i u slučaju prostora $X_0(M)$, pravac određen vektorom $\vec{i}$ označava se s x i zove apscisa, pravac određen vektorom $\vec{j}$ označava se s y i zove ordinata, dok se pravac određen vektorom $\vec{k}$ označava sa z i zove aplikata (vidi Sliku 8.6b).

Primjedba 8.5. Pravila za zbrajanje vektora i množenje vektora skalarom, ako su oni zadani sa svojim koordinatama u pravokutnom Kartezijevom koordinatnom sustavu, definiraju se analogno kao u prostoru $X_0(M)$:

$$\vec{a} = a_1\vec{i} + a_2\vec{j} + a_3\vec{k}, \qquad \vec{b} = b_1\vec{i} + b_2\vec{j} + b_3\vec{k},$$

$$\vec{a} + \vec{b} = (a_1 + b_1)\vec{i} + (a_2 + b_2)\vec{j} + (a_3 + b_3)\vec{k}, \quad \text{[zbrajanje]}$$

$$\lambda\vec{a} = (\lambda a_1)\vec{i} + (\lambda a_2)\vec{j} + (\lambda a_3)\vec{k}. \quad \text{[množenje vektora skalarom]}$$

Ranije smo utvrdili da postoji bijekcija (obostrano jednoznačno preslikavanje) između skupova E i $X_0(E)$. Primijetite da također postoji bijekcija između skupa svih uređenih trojki realnih brojeva $\mathbb{R}^3$ i vektorskog prostora $X_0(E)$ jer svakoj uređenoj trojki $(a_1, a_2, a_3) \in \mathbb{R}^3$ na jedinstven način možemo pridružiti vektor $\vec{a} = a_1\vec{i} + a_2\vec{j} + a_3\vec{k}$ iz prostora $X_0(E)$ i obrnuto. Zato ćemo često po potrebi povezivati, pa neki puta i poistovjećivati pojmove: skup E, vektorski prostor $X_0(E)$ i skup svih uređenih trojki realnih brojeva $\mathbb{R}^3$.

Zadatak 8.7. Provjerite jesu li vektori:
$\vec{a} = 5\vec{i} - \vec{j} + 3\vec{k}, \quad \vec{b} = 5\vec{i} + 2\vec{j} - \vec{k}, \quad \vec{c} = -5\vec{i} - 8\vec{j} + 9\vec{k}$ linearno zavisni.

Rješenje: Jesu, $\vec{c} = 2\vec{a} - 3\vec{b}$.

8.3 Skalarni produkt

U nastavku želimo analizirati geometrijsko značenje koordinata (komponenti) nekog vektora $\vec{a} \in X_0(E)$ zadanog u pravokutnom Kartezijevom koordinatnom sustavu. Pri tome će važnu ulogu imati pojam koji uvodimo u ovom odjeljku.

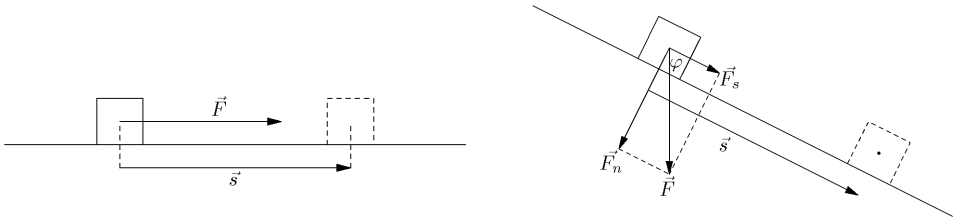
Iz tradicionalnih razloga motivacija za uvođenje pojma skalarnog produkta[4] dva vektora dolazi iz fizike. Razmotrimo fizikalnu definiciju rada sile $\vec{F}$ na putu $\vec{s}$. Ako rad obavlja sila $\vec{F}$ koja djeluje u smjeru puta $\vec{s}$, onda je rad zadan s

$$W = \|\vec{F}\| \cdot \|\vec{s}\|,$$

(pišemo jednostavno: $W = F\,s$). Ako sila $\vec{F}$ ne djeluje u smjeru puta $\vec{s}$, onda rad obavlja samo komponenta $\vec{F}_s$ sile u smjeru puta $\vec{s}$, tj.

$$\vec{F} = \vec{F}_s + \vec{F}_n,$$
$$W = \|\vec{F}_s\| \cdot \|\vec{s}\| = (F\cos\varphi)\,s = F\,s\cos\varphi,$$

gdje je φ kut između vektora sile $\vec{F}$ i vektora puta $\vec{s}$ (vidi Sliku 8.8).



Slika 8.8: Rad sile $\vec{F}$ na putu $\vec{s}$

Primijetite da je $\vec{F}_s$ ortogonalna projekcija sile $\vec{F}$ u smjeru vektora puta $\vec{s}$.

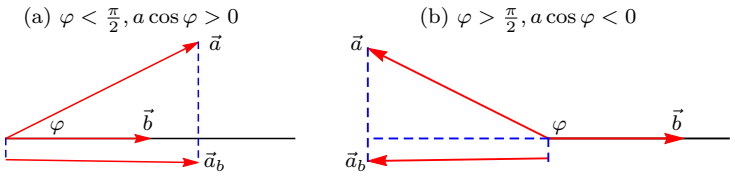
Općenito ćemo projekciju vektora $\vec{a}$ u smjeru vektora $\vec{b}$ označiti s $\vec{a}_b$.[5]

$$a_b = a\,\cos\varphi, \quad 0 \le \varphi \le \pi.$$

Primijetite da broj $a\,\cos\varphi$ može biti pozitivan ($\varphi < \frac{\pi}{2}$) ili negativan ($\varphi > \frac{\pi}{2}$) (vidi Sliku 8.9).

[4] Potrebna predznanja iz trigonometrije mogu se pronaći kod [29].

[5] U daljnjem tekstu neki puta će zbog jednostavnosti biti potrebno normu vektora $\|\vec{a}\|$ pisati jednostavno s a.



Slika 8.9: Projekcija vektora $\vec{a}$ u smjeru vektora $\vec{b}$

Definicija 8.4. Skalarni produkt u $X_0(E)$ je binarna operacija $\langle\cdot\rangle, : X_0 \times X_0 \to \mathbb{R}$ koja paru vektora $\vec{a}, \vec{b} \in X_0$ pridužuje broj (skalar), kojeg ćemo označiti s $\langle\vec{a}, \vec{b}\rangle$,

$$\langle\vec{a}, \vec{b}\rangle = \begin{cases} ab\cos\varphi, & \vec{a}, \vec{b} \neq \vec{0}, \quad 0 \leq \varphi \leq \pi, \\ 0, & \vec{a} = \vec{0} \text{ ili } \vec{b} = \vec{0}, \end{cases}$$

pri čemu je običaj da se i rezultat operacije naziva skalarni produkt.[6]

Primjedba 8.6. Primijetite da se skalarni produkt dva vektora može prikazati kao produkt norme jednog vektora i skalarne projekcije drugog vektora na prvi vektor,

$$\langle\vec{a}, \vec{b}\rangle = \|\vec{a}\|b_a = a_b\|\vec{b}\|.$$

Primjedba 8.7. Primijetite da skalarni produkt dva vektora može biti jednak nuli onda i samo onda ako je jedan od vektora nulvektor ili ako su vektori međusobno okomiti. Niže navodimo najvažnija svojstva skalarnog produkta:

1. $\langle\vec{a}, \vec{b}\rangle = \langle\vec{b}, \vec{a}\rangle$,
2. $\langle\vec{a}, \vec{a}\rangle = a^2 \geq 0 \quad$ i $\quad \langle\vec{a}, \vec{a}\rangle = 0 \iff \vec{a} = \vec{0}$,
3. $\langle\vec{a} + \vec{b}, \vec{c}\rangle = \langle\vec{a}, \vec{c}\rangle + \langle\vec{b}, \vec{c}\rangle$,
4. $\langle\lambda\,\vec{a}, \vec{b}\rangle = \lambda\,\langle\vec{a}, \vec{b}\rangle$.

Primjer 8.8. *Lako se na osnovi Definicije 8.4 vidi da također vrijedi:*

1. $(\vec{a} + \vec{b})^2 = \langle\vec{a} + \vec{b}, \vec{a} + \vec{b}\rangle = a^2 + 2\langle\vec{a}, \vec{b}\rangle + b^2$,
2. $(\vec{a} - \vec{b})^2 = \langle\vec{a} - \vec{b}, \vec{a} - \vec{b}\rangle = a^2 - 2\langle\vec{a}, \vec{b}\rangle + b^2$,
3. $(\vec{a}, \vec{b} \neq \vec{0}) \quad \cos\angle(\vec{a}, \vec{b}) = \frac{\langle\vec{a}, \vec{b}\rangle}{a\,b}, \quad \vec{a}, \vec{b} \neq \vec{0}$,

[6] engl.: scalar (dot) product, njem.: Skalarprodukt (Ineresprodukt)

4. $(\vec{a}, \vec{b} \neq \vec{0})$ $\quad \langle \vec{a}, \vec{b} \rangle = 0 \Longleftrightarrow \vec{a} \perp \vec{b}$.

Primjer 8.9. *Niže navodimo tablicu množenja (skalarni produkt) za vektore ortonormirane baze* $(\vec{i}, \vec{j}, \vec{k})$ *vektorskog prostora* $X_0(E)$*:*

$\cdot$	$\vec{i}$	$\vec{j}$	$\vec{k}$
$\vec{i}$	1	0	0
$\vec{j}$	0	1	0
$\vec{k}$	0	0	1

Zadatak 8.8. Ako je vektor $\vec{a} + 3\vec{b}$ okomit na vektor $7\vec{a} - 5\vec{b}$ i vektor $\vec{a} - 4\vec{b}$ okomit na vektor $7\vec{a} - 2\vec{b}$, odredite kut između vektora $\vec{a}$ i $\vec{b}$.

Rješenje: $\varphi_1 = \frac{\pi}{3}, \quad \varphi_2 = \frac{2}{3}\pi$.

Zadatak 8.9. Odredite kut između jediničnih vektora $\vec{e}_1$ i $\vec{e}_2$ ako se zna da su vektori $\vec{e}_1 + 2\vec{e}_2$ i $5\vec{e}_1 - 4\vec{e}_2$ međusobno okomiti.

Rješenje: $\varphi = \frac{\pi}{3}$.

Zadatak 8.10. Pokažite da je vektor $\vec{a}\,\langle \vec{b}, \vec{c} \rangle - \vec{b}\,\langle \vec{a}, \vec{c} \rangle$ okomit na vektor $\vec{c}$.

Direktnom provjerom uz korištenje tablice množenja iz Primjera 8.9 dobivamo pravilo za izračunavanje skalarnog produkta dva vektora koji su zadani sa svojim koordinatama u pravokutnom Kartezijevom koordinatnom sustavu $(O; (\vec{i}, \vec{j}, \vec{k}))$.

Teorem 8.2. *Skalarni produkt vektora* $\vec{a}, \vec{b} \in X_0(E)$*,*

$$\begin{aligned} \vec{a} &= a_1\vec{i} + a_2\vec{j} + a_3\vec{k}, \\ \vec{b} &= b_1\vec{i} + b_2\vec{j} + b_3\vec{k}, \end{aligned}$$

zadan je formulom

$$\vec{a} \cdot \vec{b} = a_1b_1 + a_2b_2 + c_1c_2. \tag{8.4}$$

Iz definicije skalarnog produkta i norme vektora korištenjem formule (8.4) dobivamo:

$$\|\vec{a}\| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{a_1^2 + a_2^2 + a_3^2}, \tag{8.5}$$

$$\cos \angle(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\,\|\vec{b}\|} = \frac{a_1b_1 + a_2b_2 + a_3b_3}{\sqrt{a_1^2 + a_2^2 + a_3^2}\sqrt{b_1^2 + b_2^2 + b_3^2}}, \quad \vec{a}, \vec{b} \neq \vec{0}. \tag{8.6}$$

Primjer 8.10. *Pokažimo da su dijagonale četverokuta $ABCD$ s vrhovima $A = (1, -2, 2)$, $B = (1, 4, 0)$, $C = (-4, 1, 1)$, $D = (-5, -5, 3)$ međusobno okomite.*

Kako je $\overrightarrow{AC} = \vec{r}_C - \vec{r}_A = -5\vec{i} + 3\vec{j} - \vec{k}$ i $\overrightarrow{BD} = \vec{r}_D - \vec{r}_B = -6\vec{i} - 9\vec{j} + 3\vec{k}$, imamo $\overrightarrow{AC} \cdot \overrightarrow{BD} = 0$.

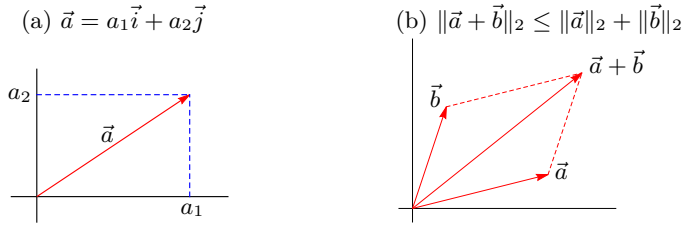
Primjer 8.11. *Zadan je trokut ABC s vrhovima $A = (-1, -2, 4)$, $B = (-4, -2, 0)$, $C = (3, -2, 1)$. Treba odrediti unutrašnji kut tog trokuta pridružen vrhu B.*

Kako je $\overrightarrow{BA} = \vec{r}_A - \vec{r}_B = 3\vec{i} + 4\vec{k}$ i $\overrightarrow{BC} = \vec{r}_C - \vec{r}_B = 7\vec{i} + \vec{k}$, dobivamo

$$\cos\beta = \frac{\overrightarrow{BA} \cdot \overrightarrow{BC}}{\|\overrightarrow{BA}\| \, \|\overrightarrow{BC}\|} = \frac{25}{\sqrt{50}\sqrt{25}} = \frac{\sqrt{2}}{2} \quad \Longrightarrow \quad \beta = \frac{\pi}{4}\,(45^\circ).$$

8.4 Norma

Pretpostavimo da je u ravninu M uveden pravokutni Kartezijev koordinatni sustav $(O; \vec{i}, \vec{j})$. Neka je $\vec{a} = a_1\vec{i} + a_2\vec{j}$ proizvoljni vektor. Geometrijski se lako vidi (vidi Sliku 8.10a) da je duljina ovog vektora $\vec{a}$ zadana s $\|\vec{a}\|_2 = \sqrt{a_1^2 + a_2^2}$. Ovu veličinu obično nazivamo euklidska norma vektora $\vec{a}$.



Slika 8.10: Euklidova norma vektora

Zadatak 8.11. Pokažite da za ovako definiranu normu vektora vrijede sljedeća osnovna svojstva:

(i) $\|\vec{a}\|_2 \geq 0 \quad \& \quad (\|\vec{a}\|_2 = 0 \Leftrightarrow \vec{a} = \vec{0})$,

(ii) $\|\lambda\vec{a}\|_2 = |\,\lambda\,| \, \|\vec{a}\|_2, \quad \lambda \in \mathbb{R}$,

(iii) $\|\vec{a} + \vec{b}\|_2 \leq \|\vec{a}\| + \|\vec{b}\|_2$.

Motivirajući se prethodnim primjerom euklidske norme, normu vektora možemo i općenito definirati:

Definicija 8.5. Neka je X_0 vektorski prostor. Funkciju $\|\cdot\| : X_0 \to \mathbb{R}_+$, koja svakom vektoru $\vec{a} \in X_0$ pridružuje nenegativni realni broj (koji ćemo označiti s $\|\vec{a}\|$) zovemo norma vektora $\vec{a}$ ako vrijedi

(i) $\|\vec{a}\| = 0 \Leftrightarrow \vec{a} = \vec{0}$ [pozitivna definitnost],

(ii) $\|\lambda\vec{a}\| = \mid \lambda \mid \|\vec{a}\|$ za svaki $\lambda \in \mathbb{R}$ i za svaki $\vec{a} \in X_0$ [homogenost],

(iii) $\|\vec{a} + \vec{b}\| \leq \|\vec{a}\| + \|\vec{b}\|$ za svaki $\vec{a}, \vec{b} \in X_0$ [nejednakost trokuta].

Vektorski prostor X_0 na kome je definirana norma naziva se normirani vektorski prostor. Najčešće korištene vektorske norme su

$$\|\vec{a}\|_1 = \mid a_1 \mid + \mid a_2 \mid + \mid a_3 \mid, \qquad [\ell_1\text{-norma (Manhattan)}]$$

$$\|\vec{a}\|_2 = \sqrt{a_1^2 + a_2^2 + a_3^2}, \qquad [\text{Euklidska } \ell_2\text{-norma}]$$

$$\|\vec{a}\|_\infty = \max\{\mid a_1 \mid, \mid a_2 \mid, \mid a_3 \mid\}, \qquad [\text{Čebiševljeva } \ell_\infty\text{-norma}]$$

$$\|\vec{a}\|_p = (|a_1|^p + |a_2|^p + |a_3|^p)^{1/p}, \quad p \geq 1. \qquad [\ell_p\text{-norma}]$$

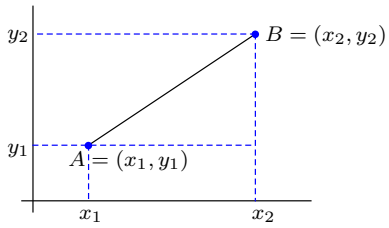
Primijetite da ℓ_p norma obuhvaća sve prethodno spomenute norme.

Zadatak 8.12. Pokažite da spomenute norme imaju sva tri svojstva navedena u Definiciji 8.5.

8.5 Udaljenost

Udaljenost dviju točaka $A = (x_1, y_1)$, $B = (x_2, y_2)$ u ravnini M u kojoj je uveden pravokutni Kartezijev koordinatni sustav možemo izračunati (vidi Sliku 8.11) po formuli

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{8.7}$$



Slika 8.11: Udaljenost točaka A, B u ravnini

Ako su $\vec{r}_A, \vec{r}_B$ odgovarajući radijvektori točaka A, B,

$$\vec{r}_A = x_1\vec{i} + y_1\vec{j}, \qquad \vec{r}_B = x_2\vec{i} + y_2\vec{j},$$

onda formulu (8.7) možemo zapisati kao

$$d_2(A,B) = \|\vec{r}_B - \vec{r}_A\|_2 = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}. \tag{8.8}$$

Na sličan način može se definirati i udaljenost dviju točaka pomoću ℓ_1-norme, ℓ_∞-norme ili općenite ℓ_p-norme sljedećim formulama:

$$d_1(A,B) = \|\vec{r}_B - \vec{r}_A\|_1 = |x_2-x_1| + |y_2-y_1|, \tag{8.9}$$

$$d_\infty(A,B) = \|\vec{r}_B - \vec{r}_A\|_\infty = \max\{|x_2-x_1|, |y_2-y_1|\}, \tag{8.10}$$

$$d_p(A,B) = \|\vec{r}_B - \vec{r}_A\|_p = (|x_2-x_1|^p + |y_2-y_1|^p)^{1/p}, \quad p \geq 1. \tag{8.11}$$

$d_p\,(p \geq 1)$ udaljenost u literaturi obično se naziva *Minkowsky udaljenost.*

Zadatak 8.13. Koji je geometrijski smisao d_1, odnosno d_∞ udaljenosti dviju točaka $A, B \in M$?

Razdaljinsku (metričku) funkciju $d : M \times M \to \mathbb{R}_+$, koja dvjema točkama A, B pridružuje njihovu udaljenost $d(A,B)$ možemo i općenito definirati kao funkciju koja ima sljedeća svojstva:

(i) $d(A,B) = 0 \quad \Leftrightarrow \quad A = B$, [pozitivna definitnost]

(ii) $d(A,B) = d(B,A)$, [simetričnost]

(iii) $d(A,B) \leq d(A,C) + d(C,B)$, $\forall A, B, C \in M$. [nejednakost trokuta]

Skup svih točaka u ravnini M, u kojoj je definirana neka razdaljinska (metrička) funkcija naziva se metrički prostor. Naravno, na sličan način može se definirati i razdaljinska (metrička) funkcija u prostoru E ili na pravcu p, ali i u apstraktnom vektorskom prostoru $\mathbb{R}^n$.

Zadatak 8.14. Jedinična „kružnica" sa središtem u ishodištu $O \in \mathbb{R}^2$ definira se kao skup $\partial K = \{T \in M : \ d(O,T) = 1\}$. Nacrtajte odgovarajuće jedinične kružnice primjenom d_1, d_2 ili d_∞ metrike.

Zadatak 8.15. Zadan je trapez $ABCD$ s vrhovima: $A = (-3,2,1)$, $B = (3,-1,4)$, $C = (5,2,-3)$. Odredite četvrti vrh D ako vrijedi $\overrightarrow{AB} = 3\,\overrightarrow{DC}$.

Rješenje: $\vec{r}_D = \vec{r}_C - \frac{1}{3}\vec{r}_B + \frac{1}{3}\vec{r}_A, \quad D = (3,3,-4)$

Zadatak 8.16. Zadan je trokut ABC s vrhovima: $A = (-3,2,1)$, $B = (3,-2,2)$, $C = (5,2,-4)$. Odredite duljinu težišnice iz vrha A.

Rješenje: $P_A = (4,0,-1), \quad \overrightarrow{AP_A} = 7\vec{i} - 2\vec{j} - 2\vec{k}, \quad d = \sqrt{57}$

Zadatak 8.17. Zadan je paralelogram $ABCD$ s vrhovima: $A = (-3,2,1)$, $B = (3,-1,4)$, $C = (5,2,-3)$, $D = (-1,5,-6)$. Izračunajte udaljenost točke A do sjecišta njegovih dijagonala.

Rješenje: $S(1,2,-1), \quad \vec{r}_S = \frac{1}{2}(\vec{r}_C - \vec{r}_A) = \frac{1}{2}(\vec{r}_D - \vec{r}_B), \quad d(A,S) = 2\sqrt{5}.$

Zadatak 8.18. Dokažite da vektor $\vec{a} = \frac{1}{2}\langle\overrightarrow{OA} + \overrightarrow{OB}\rangle$ s početkom u točki O ima vrh u polovištu dužine $\overline{AB}$.

8.6 Vektorski prostor $\mathbb{R}^n$

Skup $\mathbb{R}^n$ možemo promatrati ili kao skup uređenih n-torki realnih brojeva $(x_1, \ldots, x_n)$ koje predstavljaju točke u apstraktnom n-dimenzionalnom prostoru ili kao skup vektora iz n-dimenzionalnog vektorskog prostora $\mathbb{R}^n = \{x = (x_1, \ldots, x_n) \colon x_i \in \mathbb{R}\}$.

Za dva vektora $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ računske operacije definiramo na sljedeći način:

- zbrajanje:
$$x + y = (x_1 + y_1, \ldots, x_n + y_n);$$
- množenje sa skalarom $\lambda \in \mathbb{R}$:
$$\lambda x = (\lambda x_1, \ldots, \lambda x_n);$$
- skalarni produkt:
$$\langle x, y\rangle = x_1 y_1 + \cdots + x_n y_n.$$

Primjer 8.12. *Specijalno, budući da je* $\langle x, x\rangle = x_1^2 + \cdots + x_n^2$, *euklidsku normu vektora* $x \in \mathbb{R}^n$ *možemo pisati*

$$\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{x_1^2 + \cdots + x_n^2}.$$

Definicija 8.6. Funkciju $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$, koja svakom vektoru $x \in \mathbb{R}^n$ pridružuje nenegativni realni broj (koji ćemo označiti s $\|x\|$) zovemo norma vektora $x \in \mathbb{R}^n$ ako vrijedi

(i) $\|x\| = 0 \Leftrightarrow x = (0, \ldots, 0)$, [pozitivna definitnost]

(ii) $\|\lambda x\| = |\lambda| \|x\|$ za svaki $\lambda \in \mathbb{R}$ i za svaki $x \in \mathbb{R}^n$, [homogenost]

(iii) $\|x + y\| \leq \|x\| + \|y\|$ za svaki $x, y \in \mathbb{R}^n$. [nejednakost trokuta]

Najčešće korištene vektorske norme su

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad [\ell_1 - norma]$$

$$\|x\|_2 = \sqrt{\langle x, x \rangle} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}, \qquad [\text{Euklidska } \ell_2\text{-norma}]$$

$$\|x\|_\infty = \max_{i=1,\ldots n} |x_i|, \qquad [\text{Čebiševljeva } \ell_\infty\text{-norma}]$$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad p \geq 1. \qquad [\ell_p\text{-norma}]$$

Definicija 8.7. Funkciju $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, sa svojstvom da za svaki $x, y \in \mathbb{R}^n$ vrijedi

(i) $d(x, y) = 0 \quad \Leftrightarrow \quad x = y,$ [pozitivna definitnost]

(ii) $d(x, y) = d(y, x),$ [simetričnost]

(iii) $d(x, y) \leq d(x, z) + d(z, y), \quad \forall x, y, z \in \mathbb{R}^n,$ [nejednakost trokuta]

zovemo metrička funkcija (ili samo metrika) na $\mathbb{R}^n$, a vrijednost funkcije $d(x, y)$ za neke $x, y \in \mathbb{R}^n$ zovemo udaljenost točaka $x, y \in \mathbb{R}^n$.

Poznavanjem norme $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ moguće je definirati odgovarajuću metričku funkciju (metriku) formulom

$$d(x, y) = \|x - y\|. \tag{8.12}$$

Najčešće korištene metričke funkcije (metrike) su

$$d_1(x,y) = \|x-y\|_1 = \sum_{i=1}^{n} |x_i - y_i|, \quad \text{[Manhattan } \ell_1 \text{ udaljenost]}$$

$$d_2(x,y) = \|x-y\|_2 = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}, \quad \text{[Euklidska } \ell_2 \text{ udaljenost]}$$

$$d_\infty(x,y) = \|x-y\|_\infty = \max_{i=1,\dots n} |x_i - y_i|. \quad \text{[Čebiševljeva } \ell_\infty \text{ udaljenost]}$$

$$d_p(x,y) = \|x-y\|_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \; p \geq 1. \quad \text{[Minkowsky } \ell_p \text{ udaljenost]}$$

Više o metričkim funkcijama može se vidjeti kod [2, 13, 18, 32, 74].

Zadatak 8.19. Napišite formule za Euklidsku, Manhattan i Čebiševljevu udaljenost dviju točaka $A = (x_1, y_1)$, $B = (x_2, y_2) \in \mathbb{R}^2$. Objasnite geometrijski smisao.

Poglavlje 9

Mathematica programi i moduli

Tijekom izvođenja nastave predviđeno je ilustrirati metode i algoritme te njihova svojstva primjenom programskog sustava *Mathematica*[1], za koji Sveučilište u Osijeku redovito obnavlja licencu. S jedne strane, na taj način demonstriraju se osnove korištenja ovog programskog sustava, a s druge strane, metode i algoritmi navedeni u ovom udžbeniku postaju bliži i operativniji.

9.1 Reprezentant

Prilikom demonstriranja primjene programskog sustava *Mathematica* za definiranje i određivanje reprezentanta podataka posebno je napisan program za podatke s jednim obilježjem ($n = 1$), posebno za podatke s dva obilježja ($n = 2$), a posebno za podatke s n obilježja.

Najprije se učitavaju podaci bez težina i *metodom pokušaja i pogrešaka* pokušava odrediti što bolja aproksimacija najboljeg LS-reprezentanta i najboljeg ℓ_1-reprezentanta. Nakon toga primjenom egzaktnih formula određuju se najbolji reprezentanti. Cijeli postupak ponavlja se za slučaj podataka s težinama.

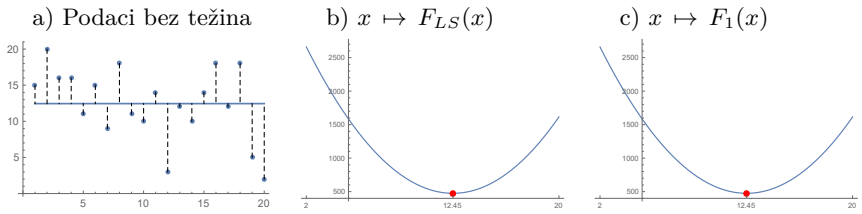
Programi su opskrbljeni odgovarajućom grafikom uz korištenje gotovih *Mathematica*-naredbi, što ima značajan utjecaj na razumijevanje gradiva. Programi se mogu modificirati izborom broja i konfiguracije skupa podataka, kao i dopunskim izračunavanjima.

[1] Svi niže navedeni programi slobodno se mogu preuzeti na adresi `http://www.mathos.unios.hr/images/homepages/scitowsk/Programi.zip`

Primjer 9.1. *Za skup podataka $\mathcal{A} \subset \mathbb{R}$ (vidi Sliku 9.1a) zadan s*

```
In[1]:= a = 1; b = 20; m = 20; SeedRandom[13];
        A = RandomInteger[{a, b}, m];
```

određuje se najbolji LS-reprezentant (centroid, aritmetička sredina) i prikazuje odgovarajuća minimizirajuća funkcija (vidi Sliku 9.1b)



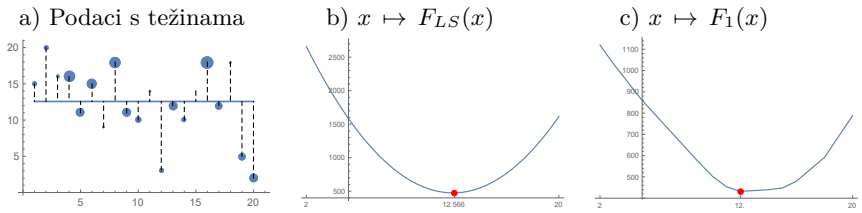
Slika 9.1: Najbolji LS-reprezentant i najbolji ℓ_1-reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}$

Također, za iste podatke određuje se najbolji ℓ_1-reprezentant (medijan) i prikazuje odgovarajuća minimizirajuća funkcija (vidi Sliku 9.1c).

Za skup podataka s težinama $\mathcal{A} \subset \mathbb{R}$ (vidi Sliku 9.2a) određuje se najbolji težinski LS-reprezentant (centroid, težinska aritmetička sredina) i prikazuje odgovarajuća minimizirajuća funkcija (vidi Sliku 9.2b).

```
In[2]:= a = 1; b = 20; m = 20; SeedRandom[13];
        A = RandomInteger[{a, b}, m];
        W = RandomInteger[{0, 10}, m];
```

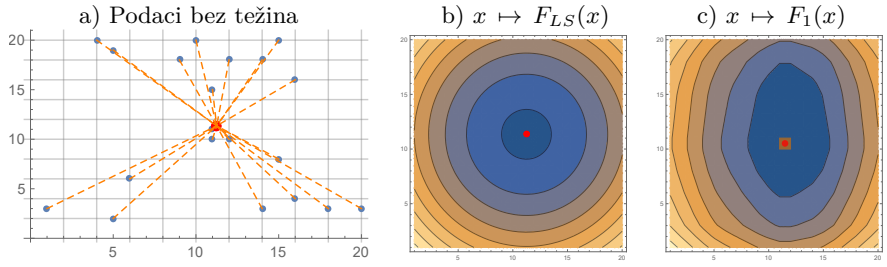
Za ove podatke određuje se najbolji težinski ℓ_1-reprezentant (težinski medijan) i prikazuje odgovarajuća minimizirajuća funkcija (vidi Sliku 9.2c).



Slika 9.2: Najbolji težinski LS-reprezentant i najbolji težinski ℓ_1-reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}$

Primjer 9.2. *Za skup podataka $\mathcal{A} \subset \mathbb{R}^2$ (vidi Sliku 9.3a) zadan s*

```
In[1]:= a = 1; b = 20; m = 20; SeedRandom[13];
        A = RandomInteger[{a, b}, {m, 2}];
```



Slika 9.3: Najbolji LS-reprezentant i najbolji ℓ_1-reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}^2$

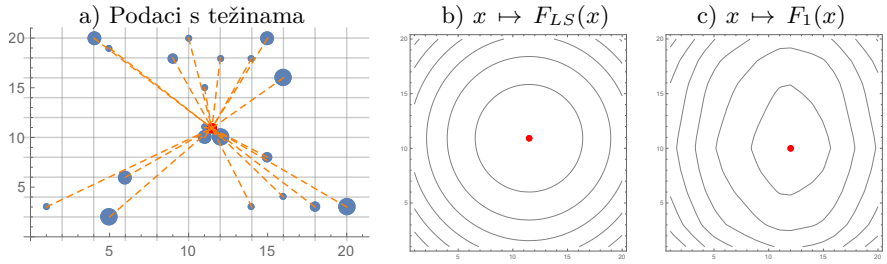
određuje se najbolji LS-reprezentant (centroid) i prikazuje `ContourPlot` *odgovarajuće minimizirajuće funkcije (vidi Sliku 9.3b)*

Također, za iste podatke određuje se najbolji ℓ_1-reprezentant (medijan) i prikazuje `ContourPlot` *odgovarajuće minimizirajuće funkcije (vidi Sliku 9.3c).*

Za skup podataka s težinama $\mathcal{A} \subset \mathbb{R}^2$ (vidi Sliku 9.4a) zadan s

```
In[2]:= a = 1; b = 20; m = 20; SeedRandom[13];
        A = RandomInteger[{a, b}, {m, 2}];
        W = RandomInteger[{2, 5}, m]
```

traži se najbolji težinski LS-reprezentant (centroid) i prikazuje `ContourPlot` *odgovarajuće minimizirajuće funkcije (vidi Sliku 9.4b).*



Slika 9.4: Najbolji težinski LS-reprezentant i najbolji težinski ℓ_1-reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}^2$

Također, za iste podatke određuje se najbolji težinski ℓ_1-reprezentant (težinski medijan) i prikazuje `ContourPlot` *odgovarajuće minimizirajuće funkcije (vidi Sliku 9.4c).*

Zadatak 9.1. Definirajte skup podataka $\mathcal{A}$ tako da ima 10% outliers (vidi Primjer 2.2, str. 5) i odredite LS-reprezentant i ℓ_1-reprezentant. Što primjećujete?

9.2 Grupiranje podataka

Kao što smo naveli u Poglavlju 3, str. 23, broj svih particija skupa $\mathcal{A} \subset \mathbb{R}^n$ jednak je Stirlingovom broju druge vrste zadanom s (3.1), str. 23, koji može biti izuzetno velik (vidi Tablicu 3.1 na str. 24). To znači da traženje optimalne particije pretraživanjem svih particija praktično nije moguće. Specijalno, samo za podatke s jednim obilježjem ($n = 1$) taj je broj znatno manji (vidi Tablicu 3.4 na str. 35), a može se izračunati po formuli (3.16), str. 35. Zato ćemo problem grupiranja podataka analizirati primjenom programskog sustava *Mathematica* posebno za podatke s jednim obilježjem, a posebno za podatke s dva obilježja. Osim toga, te slučajeve moguće je elegantno grafički obraditi. Problem grupiranja podataka s n obilježja u suštitni se ne razlikuje od problema za $n = 2$.

9.2.1 Grupiranje podataka s jednim obilježjem

Promatramo skup podataka $\mathcal{A} = \{a_i \in \mathbb{R} \colon i = 1, \ldots, m\}$ s jednim obilježjem zadan s

```
In[1]:= A = {1, 2, 6, 7, 9}; m = Length[A];
        data = Table[{A[[i]], .4}, {i, m}];
        sl1 = ListPlot[data, PlotStyle -> {Blue, PointSize[.03]},
        Axes -> {True, False}, Ticks -> {A, None}, PlotRange -> {0, 1},
        AspectRatio -> Automatic, ImageSize -> 300]
```

Zbog jednostavnosti i ilustrativnosti promatrat ćemo problem grupiranja ovakvih podataka u $k = 2$ klastera. Broj svih particija $|\mathcal{P}(\mathcal{A}, 2)|$, odnosno broj svih particija čiji se klasteri međusobno nastavljaju jedan na drugi $|\hat{\mathcal{P}}(\mathcal{A}, 2)|$, određuje se po formulama

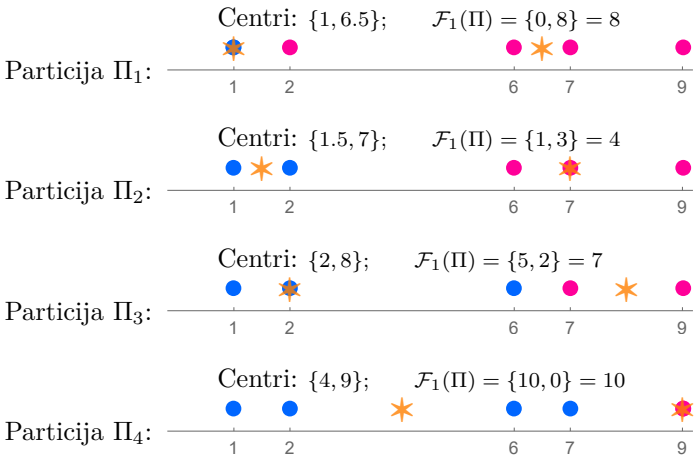
$$|\mathcal{P}(\mathcal{A}, 2)| = 2^{m-1} - 1, \qquad |\hat{\mathcal{P}}(\mathcal{A}, 2)| = \binom{m-1}{2-1}.$$

Najprije ćemo izraditi modul `Particija1[PI_, metrika_]` koji grafički prikazuje particiju $\Pi = \{\pi_1, \pi_2\}$ skupa $\mathcal{A}$ zajedno s centrima njegovih klastera uz mogućnost izbora parametra `metrika`. Za `metrika=1` modul daje centre (medijane) klastera i vrijednost funkcije cilja $\mathcal{F}_1$, a za `metrika=2` modul daje centroide (aritmetičke sredine) klastera i vrijednosti funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$.

Primjenom programskog sustava *Mathematica* korištenjem ranije spomenutog modula `Particija1[PI_, metrika_]` najprije ćemo izraditi mali jednostavni program koji ispisuje sve particije s karakteristikama vezanim uz LS-kvazimetričku funkciju ili vezanim uz ℓ_1-metričku funkciju. Nakon toga izradit ćemo mali jednostavni program koji ispisuje se sve particije čiji se

klasteri međusobno nastavljaju jedan na drugi uz primjenu LS-kvazimetričke funkcije ili ℓ_1-metričke funkcije.

Primjer 9.3. *Funkcioniranje programa i modula* `Particija1[PI_, metrika_]` *ilustrirat ćemo na skupu* $\mathcal{A} = \{1, 2, 6, 7, 9\}$. *Na Slici 9.5 prikazane su sve particije ovog skupa čiji se klasteri međusobno nastavljaju jedan na drugi uz primjenu* ℓ_1*-metričke funkcije.*



Slika 9.5: Particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$ čiji se klasteri međusobno nastavljaju jedan na drugi uz primjenu ℓ_1-metričke funkcije

9.2.2 Grupiranje podataka s dva obilježja

Promatramo skup podataka $\mathcal{A} = \{a^i \in \mathbb{R}^2 : i = 1, \ldots, m\}$ s dva obilježja zadan s (vidi Sliku 9.6)

```
In[1]:= a = 1; b = 10; m = 12; SeedRandom[2];
        A = {{2,9},{3,3},{6,5},{4,7},{5,8},{6,2},{6,7},{8,4},{8,6},{9,5}};
        tab = Table[i, {i, m}];
        slA = ListPlot[A, GridLines -> {tab,tab}, PlotStyle->{PointSize[.04]},
                       AspectRatio -> 1, ImageSize -> 200]
```

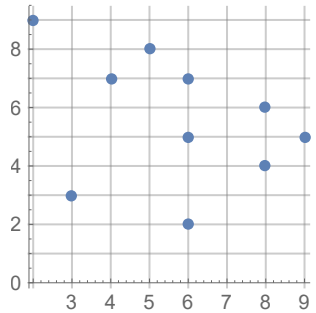
Promatramo jednu k-particiju ($k \geq 2$) skupa $\mathcal{A}$ koju ćemo analizirati uz primjenu ℓ_1-metričke funkcije (`metrika=1`) ili uz primjenu LS-kvazimetričke funkcije (`metrika=2`). Skup lista rednih brojeva svakog klastera u particiji označimo s `iPI`. Najprije ćemo izraditi modul `Particija2[A_, iPI_, metrika_]` koji grafički prikazuje particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ zajedno s centrima

njegovih klastera. Za `metrika=1` modul daje centre (medijane) klastera i vrijednost funkcije cilja $\mathcal{F}_1$, a za `metrika=2` modul daje centroide (aritmetičke sredine) klastera i vrijednosti funkcija cilja $\mathcal{F}_{LS}$.

Na taj način moći ćemo usporediti optimalnost različitih particija.

Primjer 9.4. *Funkcioniranje modula* `Particija2[A_, iPI_, metrika_]` *ilustrirat ćemo na skupu* $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1 \ldots, m\}$ *zadanom s*

i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	4	5	6	6	6	8	8	9
y_i	9	3	7	8	2	5	7	4	6	5

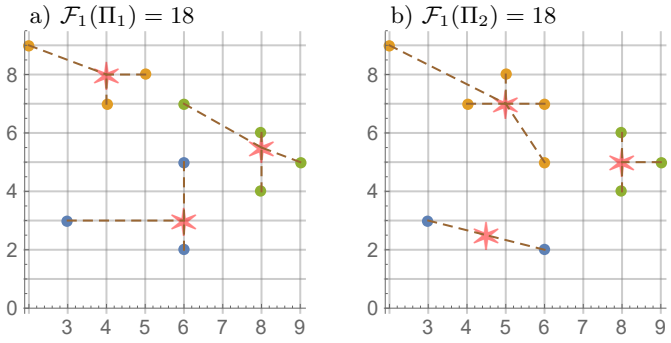


Slika 9.6: Skup $\mathcal{A} = \{(2,9),(3,3),(4,7),(5,8),(6,2),(6,5),(6,7),(8,4),(8,6),(9,5)\}$

Promatrat ćemo dvije 3*-particije* Π_1, Π_2 *koje ćemo zadati skupom lista rednih brojeva svakog klastera u particiji:*

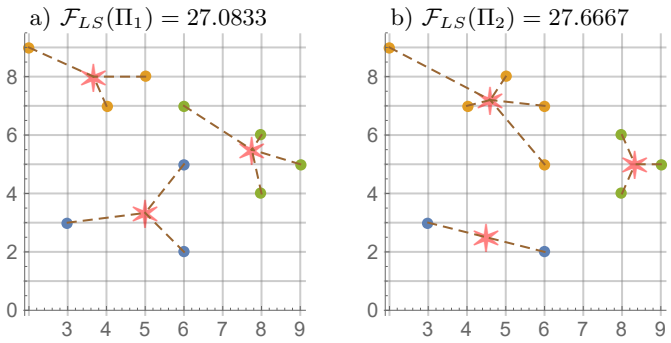
$$iPI_1 = \{\{2,3,6\},\{1,4,5\},\{7,8,9,10\}\},$$
$$iPI_2 = \{\{2,6\},\{1,3,4,5,7\},\{8,9,10\}\}.$$

Slika 9.7 prikazuje obje particije skupa $\mathcal{A}$ i njihove centre (medijane) određene uz primjenu ℓ_1-metričke funkcije. Vidi se da vrijedi $\mathcal{F}_1(\Pi_1) = \mathcal{F}_1(\Pi_2)$, što znači da su obje particije jednako blizu ℓ_1-optimalnoj.



Slika 9.7: Primjena ℓ_1-metričke funkcije

Slično, Slika 9.8 prikazuje obje particije skupa $\mathcal{A}$ i njihove centroide (aritmetičke sredine) određene uz primjenu LS-kvazimetričke funkcije. Vidi se da vrijedi $\mathcal{F}_{LS}(\Pi_1) < \mathcal{F}_{LS}(\Pi_2)$, što znači da je particija Π_1 bliže LS-optimalnoj.



Slika 9.8: Primjena LS-kvazimetričke funkcije

9.3 k-means algoritam uz primjenu LS-kvazimetričke funkcije

Kao što smo pokazali u t.4.2, str. 68, odnosno u t.4.3, str. 73, k-means algoritam za traženje LOP skupa $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ uz primjenu LS-kvazimetričke funkcije $d_{LS} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d_{LS}(x, y) = \|x - y\|_2^2$ može se opisati s dva koraka koji se iterativno ponavljaju:

`Korak A:` Pridruživanje (assignment step). Poznavanjem međusobno različitih točaka $z_1, \ldots, z_k \in \mathbb{R}^n$, skup $\mathcal{A}$ treba grupirati u k disjunktnih klastera $\pi_1, \ldots, \pi_k$ korištenjem principa minimalnih udaljenosti

$$\pi_j = \{a \in \mathcal{A} \colon \|z_j - a\|_2 \leq \|z_s - a\|_2, \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k.$$

`Korak B:` Korekcija (update step). Poznavanjem particije $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$, treba definirati centroide klastera

$$c_j = \frac{1}{W_j} \sum_{a^s \in \pi_j} w_s a^s, \quad W_j = \sum_{a^s \in \pi_j} w_s, \qquad j = 1, \ldots, k.$$

Algoritam možemo pokrenuti ili zadavanjem početne particije (tada se najprije pokreće `Korak B`) ili zadavanjem početnih centara (tada se najprije pokreće `Korak A`). Postupak se dalje ponavlja toliko dugo dok trenutna i prethodna particija ne postanu jednake (centroidi njihovih klastera tada također postanu jednaki). U svakom koraku k-means algoritma snižava se vrijednost funkcije cilja $\mathcal{F}_{LS}$ i asimptotski približava lokalno najmanjoj mogućoj vrijednosti [31, 68].

9.3.1 Podaci s jednim obilježjem

Neka je $\mathcal{A} = \{a^i \in \mathbb{R} \colon i = 1, \ldots, m\}$ skup podataka s jednim obilježjem, pri čemu svaki podatak $a^i \in \mathcal{A}$ ima odgovarajuću težinu $w_i > 0$. Neka je $z = (z_1, \ldots, z_k)$ niz (lista) međusobno različitih točaka, a `Ind` broj koji može primiti vrijednost „0" ili „1".

k-means algoritam pokrenut ćemo zadavanjem početnih centroida. `Mathematica`-modulu `WKMeans1[Pod_,z_,Ind_]` predajemo skup podataka sastavljenih od parova {težina, podatak}:

```
Pod = Table[{w[[i]], A[[i]]}, {i,Length[A]}],
```

listu `z` i pokazatelj `Ind` $\in \{0, 1\}$.

Modul `WKMeans1` nakon ispisa početnih podataka i početne slike izvodi k-means algoritam počevši s `Korakom A`. Ako je `Ind=1`, u svakoj iteraciji ispisuju se međurezultati i prikazuje odgovarajuća slika. Na kraju modul predaje lokalno optimalnu particiju, centroide njenih klastera, vrijednost kriterijske funkcije cilja $\mathcal{F}_{LS}$ i provedeni broj iteracija `IT`.

```
In[1]:= WKMeans1[Pod_, z_, Ind_] :=
        Module[{PI, tab, imin, z0, c, k=Length[z], m=Length[Pod], podaci,
        centri, IT = 0},
```

```
z0 = z;
podaci = Table[{Pod[[i, 2]], .4}, {i, m}];
centri = Table[{z0[[i]], .4}, {i, k}];
cc = Sum[Pod[[i, 1]] Pod[[i, 2]], {i, m}]/Total[Pod[[All, 1]]];
PI = Table[{}, {i, k}];
Do[
    tab = Table[Norm[Pod[[i, 2]] - z0[[j]]], {j, k}];
    imin = Ordering[tab, 1][[1]];
    PI[[imin]] = Append[PI[[imin]], Pod[[i]]]
   ,{i, m}];
        (* Početna iteracija:podaci i slike *)
FS = Table[
Sum[PI[[j,s,1]] Norm[PI[[j,s,2]] - z0[[j]]]^2, {s,Length[PI[[j]]]}]
           ,{j,k}];
Fm = Sum[
Pod[[i,1]] Min[Table[Abs[Pod[[i,2]] - z0[[j]] ]^2, {j,k}]], {i,m}];
G = Sum[
Sum[PI[[j,s,1]], {s,Length[PI[[j]]]}] Norm[cc - z0[[j]]]^2, {j,k}];
Print["\nA = ", Pod[[All, 2]], "\nw = ", Pod[[All, 1]],
 "\nPocetna pozicija: Assignment points = ", Round[z0, -.01],
 "; F0=", FS, " = ", Round[Total[FS], -.01], "; F_min =",
 Round[Fm, -.01], "; G =", Round[G, -.01]];
sl1=ListPlot[podaci,PlotStyle->{Blue,PointSize[.03]},PlotRange->{0,1},
    Axes->{True,False}, Ticks->{A, None}, AspectRatio->Automatic];
sl2=ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
    PlotStyle -> {Orange, Opacity[.8]}, PlotRange->{0,1},
    Axes -> {True, False}, AspectRatio -> Automatic];
Print[Show[sl1, sl2, ImageSize -> 300]];
                    (* Petlja *)
c = Table[0, {j, k}];
While[IT = IT + 1;
Do[j1 = k - j + 1; mj = Length[PI[[j1]]];
If[mj != 0,
c[[j1]] = Sum[PI[[j1, All, 1]][[s]]*PI[[j1, All, 2]][[s]], {s, mj}]/
           Sum[PI[[j1, All, 1]][[s]], {s, mj}],
PI = Drop[PI,{j1}]; c = Drop[c,{j1}]; z0 = Drop[z0,{j1}] ], {j,k}];
k = Length[PI];
Chop[Norm[c - z0]] != 0,
centri = Table[{c[[i]], .4}, {i, k}];
          (* Podaci i slike *)
If[Ind != 0, G = Sum[
   Sum[PI[[j,s,1]], {s,Length[PI[[j]]]}] Norm[cc - c[[j]]]^2, {j,k}];
   Fm = Sum[Pod[[i,1]] Min[Table[Abs[Pod[[i, 2]]-c[[j]] ]^2, {j,k}]]
            ,{i,m}];
   Print["Iteracija_", IT, ":  Particija: ", PI, "\nCentri: ",
      Round[c, -.01], "\nF(PI): ",Round[WFLS[PI, c], -.01],
      "; F_min=", Round[Fm, -.01], "; G =",Round[G, -.01] ];
sl1=ListPlot[Table[Table[{PI[[i,j,2]], .4}, {j,Length[PI[[i]]]}]
                  ,{i,Length[PI]}],
```

```
        PlotStyle->Table[{Hue[.3*i+.3],PointSize[.03]},{i,Length[PI]}],
          AxesOrigin->{0,0}, AspectRatio->Automatic, PlotRange -> {0,1},
          Axes -> {True, False}, Ticks -> {A, None}];
    sl2=ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
        PlotStyle -> {Orange, Opacity[.8]}, Axes -> {True, False},
        PlotRange -> {0, 1}, AspectRatio -> Automatic];
    Print[Show[sl1, sl2, ImageSize -> 300]];
         ];
    z0 = c;
    PI = Table[{}, {i, k}];
    Do[
        tab = Table[Norm[Pod[[i, 2]] - z0[[j]]], {j, k}];
        imin = Ordering[tab, 1][[1]];
        PI[[imin]] = Append[PI[[imin]], Pod[[i]]]
       ,{i, m}];
       ];
        {PI, z0, WFLS[PI, z0], IT}
       ]
```

LS-kriterijska funkcija cilja računa se u modulu `WFLS[PI_, c_]`, gdje je `PI` particija s centroidima klastera `c`.

```
In[2]:=WFLS[PI_,c_] := Module[{},
            Return[
                   Sum[PI[[j,s,1]]*Norm[PI[[j,s,2]] - c[[j]]]^2
                       ,{j,Length[PI]}, {s,Length[PI[[j]]]}]  ]
                          ];
```

Calinski - Harabasz indeks računa se u modulu `VCH[PI_, z_]`, gdje je `PI` particija s centroidima klastera `z`.

```
In[3]:= VCH[PI_, z_, cc_] := Module[{k = Length[z], F, G,m},
            m=Sum[Length[PI[[j]]], {j, k}];
             F=WFLS[PI, z];
              G=Sum[Sum[PI[[j,s,1]], {s,Length[PI[[j]]]}]*Norm[z[[j]]-cc]^2
                    ,{j,k}];
           {F,G,(m-k)*G/((k - 1) F)//N}
                        ]
```

Davies - Bouldin indeks računa se u modulu `VDB[PI_, z_]`, gdje je `PI` particija s centroidima klastera `z`.

```
In[4]:= Clear[VDB]
        VDB[PI_, z_]:=
            Module[{k = Length[z], FF = SD = Table[0, {j, k}], mj, sum = 0},
             Do[
                mj = Length[PI[[j]]];
```

```
        FF[[j]] =
            Sum[PI[[j,s,1]]*Norm[PI[[j,s,2]] - z[[j]]]^2, {s,mj}];
        SD[[j]] = Sqrt[FF[[j]]/Sum[PI[[j, s, 1]], {s, mj}]]
         ,{j,k}];
    Do[
        max = {};
     Do[
        If[i != j,
            max=Append[max, (SD[[i]]+SD[[j]])/Norm[z[[i]]-z[[j]]]]]
        ,{i,k}];
        sum = sum + Max[max]
         ,{j,k}];
    Print["St.dev. po klasterima: ", Round[SD, -.01]];
        sum/k // N
            ]
```

Izvođenje modula

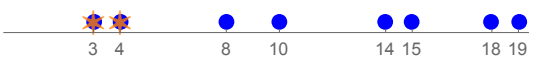
Nakon što se aktiviraju svi moduli, najprije treba učitati skup $\mathcal{A}$, odgovarajući niz težina w i početne centroide. Za Primjer 4.21, str. 83, to izgleda ovako:

```
In[1]:= A = {3, 4, 8, 10, 14, 15, 18, 19}; m = Length[A];
        w = {1, 1, 1, 3, 1, 1, 1, 1};
        z = {3, 4}; k = Length[z];
```

Samo sliku podataka i centroide možemo dobiti na sljedeći način:

```
In[2]:= podaci = Table[{A[[i]], .4}, {i, m}];
        centri = Table[{z[[j]], .4}, {j, k}];
        sl1 = ListPlot[podaci, PlotStyle -> {Blue, PointSize[.03]},
              Axes -> {True, False}, Ticks -> {A, None}, PlotRange -> {0,1},
              AspectRatio -> Automatic];
        sl2 = ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
              PlotStyle -> {Orange, Opacity[.8]}, Axes -> {True, False},
              PlotRange -> {0, 1}, AspectRatio -> Automatic];
        Show[sl1, sl2, ImageSize -> 300]
```

```
Out[3]=
```

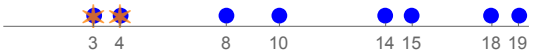


Implementaciju k-means algoritma izvodimo pozivom modula `WKMeans1` kojemu najprije predajemo *podatke* oblika `Pod=Table[{w[[i]],A[[i]]}, {i,m}]`, *centre* `z` i pokazatelj `Ind`. Ako je `Ind=0`, ispisat će se skup $\mathcal{A}$ s odgovarajućim težinama, početna pozicija s odgovarajućom slikom i rezultati: optimalna

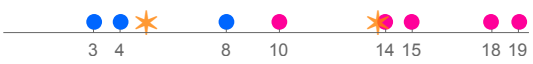
particija, centroidi njenih klastera, vrijednosti kriterijskih funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$ te odgovarajuća slika i vrijednosti DB i CH indeksa.

```
In[4]:= Pod = Table[{w[[i]], A[[i]]}, {i, m}];
        sol = WKMeans1[Pod, z, 0];
                    (* Rezultati *)
        PI = sol[[1]]; c = sol[[2]]; F1 = sol[[3]];
        cc = Sum[w[[i]] A[[i]], {i, m}]/Total[w];
        G = Sum[Sum[
        PI[[j, s, 1]], {s, Length[PI[[j]]]}] Norm[cc - c[[j]]]^2, {j, k}];
        Print["\nRjesenje:\nPI= ", PI, "\n centri = ", Round[c, -.01],
             "; c(A)=", Round[cc, -.01], "\nF=", Round[F1, -.01], "; G=",
             Round[G, -.01]];
              (* Slika *)
        centri = Table[{c[[j]], .4}, {j, k}]; tt = Union[A, c];
        sl1=ListPlot[ Table[
              Table[{PI[[i,j,2]],.4}, {j,Length[PI[[i]]]}], {i,Length[PI]}],
            PlotStyle->Table[{Hue[.3*i+.3], PointSize[.03]}, {i,Length[PI]}],
            AxesOrigin->{0,0}, AspectRatio->Automatic, PlotRange -> {0,1},
            Axes -> {True, False}, Ticks -> {A, None}];
        sl2=ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
            PlotStyle -> {Orange, Opacity[.8]}, Axes -> {True, False},
            PlotRange -> {0, 1}, AspectRatio -> Automatic];
        Print[Show[sl1, sl2, ImageSize -> 300]]
                             (* Indeksi *)
        Print["DB index: ", VDB[PI, c]]
        Print["CH index: ", VCH[PI, c, cc][[3]]]

Out[5]= A = {3,4,8,10,14,15,18,19}
        w = {1,1,1,3,1,1,1,1}
        Pocetna pozicija: Assignment points = {3.,4.};
        F0={0,766} = 766.; F_min =766.; G =519.3
```




```
        Rjesenje:
        PI= {{{1,3},{1,4},{1,8}},{{3,10},{1,14},{1,15},{1,18},{1,19}}}
        centri = {5.,13.71}; c(A)=11.1
        F=103.43; G=159.47
```



```
St.dev. po klasterima: {2.16,3.57}
DB index: 0.658061
CH index: 9.2511
```

Ako je `Ind=1`, tj. ako k-means algoritam pozovemo naredbom

```
In[6]:= sol = WKMeans1[Pod, z, 1];
```

ispisat će se numerički pokazatelji svake iteracije uz odgovarajući grafički prikaz kao na Slici 4.17, str. 84.

Ako umjesto početnih centroida algoritam želimo pokrenuti zadavanjem početne particije, tada prije pozivanja modula `WKMeans1` treba odrediti listu centroida `z` klastera početne particije.

9.3.2 Podaci s dva obilježja

Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ skup podataka, pri čemu svaki podatak $a^i \in \mathcal{A}$ ima odgovarajuću težinu $w_i > 0$ i neka je $z = (z_1, \ldots, z_k)$ lista međusobno različitih točaka. *Mathematica*-modulu `WKMeans2[Pod_,z_,Ind_]` predajemo skup podataka sastavljenih od parova {težina, podatak}:

```
Pod = Table[{w[[i]], A[[i]]}, {i,Length[A]}]
```

listu `z` i pokazatelj `Ind`.

Nakon ispisa početnih podataka i prikaza početne slike, modul `WKMeans2` izvodi k-means algoritam počevši s `Korakom A`. Ako je `Ind=1`, u svakoj iteraciji ispisuju se međurezultati i prikazuje odgovarajuća slika. Na kraju modul predaje lokalno optimalnu particiju, centroide njenih klastera, vrijednost kriterijske funkcije cilja $\mathcal{F}_{LS}$ i provedeni broj iteracija `IT`.

```
In[1]:=WKMeans2[Pod_, z_, Ind_] :=
       Module[{m=Length[Pod], k=Length[z], PI,tab,imin,F0,F1,FS,G,c, IT=1},
       cc = Sum[Pod[[i, 1]] Pod[[i, 2]], {i, m}]/Total[Pod[[All, 1]]];
       PI = Table[{}, {i, k}];
       Do[
          tab = Table[Norm[Pod[[i]][[2]] - z[[j]]]^2, {j, k}] // N;
          imin = Ordering[tab, 1][[1]];
          PI[[imin]] = Append[PI[[imin]], Pod[[i]]]
          ,{i, m}];
       PI = DeleteCases[PI, {}];
       k = Length[PI];  F0 = WFLS[PI, z];
       c = Table[Sum[PI[[j,s,1]]*PI[[j,s,2]], {s,Length[PI[[j]]]}]/
                 Sum[PI[[j, s, 1]], {s, Length[PI[[j]]]}], {j, k}];
       F1 = WFLS[PI, c];
                                (* Petlja *)
```

```
While[
 Chop[F0 - F1] != 0,
 If[Ind != 0,
 FS=Table[Sum[Norm[PI[[j, s, 2]] - c[[j]]]^2, {s, Length[PI[[j]]]}]
          ,{j,Length[PI]}];
 G=Sum[Sum[PI[[j,s,1]],{s,Length[PI[[j]]]}]*Norm[cc-c[[j]]]^2,{j,k}];
 Print["IT_", IT, ":  PI =", PI, "\n centri = ", Round[c, -.01],
 "; c(A)=", Round[cc, -.01], "\nF=", Round[FS, -.01], " = ",
 Round[F1, -.01], ";  G=", Round[G, -.01]];
     ];
IT = IT + 1;
F0 = F1;
PI = Table[{}, {i, k}];
Do[
   tab = Table[Norm[Pod[[i]][[2]] - c[[j]]]^2, {j, k}] // N;
   imin = Ordering[tab, 1][[1]];
   PI[[imin]] = Append[PI[[imin]], Pod[[i]]]
   ,{i, m}];
PI = DeleteCases[PI, {}];
k = Length[PI];
c = Table[Sum[PI[[j, s, 1]]*PI[[j, s, 2]], {s, Length[PI[[j]]]}]/
           Sum[PI[[j, s, 1]], {s, Length[PI[[j]]]}], {j, k}];
F1 = WFLS[PI, c];
      ];
 {PI, c, F1, IT}
        ]
```

Izvođenje modula

Na početku treba pokrenuti `Mathematica-package` koji omogućava potrebne grafičke prikaze

```
In[1]:= Needs["ComputationalGeometry`"]
```

Implementaciju k-means algoritma modulom `WKMeans2[Pod, z, Ind]` možemo pokrenuti zadavanjem početnih centroida na sličan način kao što smo to uradili u prethodnoj točki.

U ovom slučaju pokazat ćemo implementaciju k-means algoritma modulom `WKMeans2[Pod, z, Ind]` zadavanjem početne particije. Nakon što se aktiviraju svi moduli, treba učitati početnu particiju (čime je automatski zadan i skup $\mathcal{A}$) i odgovarajuću listu težina $\{omega\}$. Za Primjer 4.18, str. 80, to izgleda ovako:

```
In[2]:=Par={{{1,9},{2,9},{2,6}},
           {{1,3},{5,3},{6,4}},
           {{4,6},{7,7},{8,6},{9,8}}};
       omega = {{1,1,1}, {1,1,1}, {1,1,1,1}};
```

Nakon toga definiramo skup $\mathcal{A}$ s odgovarajućom listom težina w, odredimo centroid skupa $\mathcal{A}$, listu centroida z_0 početne particije i vrijednosti kriterijskih funkcija $\mathcal{F}_{LS}$ i $\mathcal{G}$ te ispisujemo početne podatke i generiramo sliku početne particije.

Modulu `WKMeans2[Pod, z, Ind]` predajemo podatke oblika

```
In[3]:= Pod = Table[{w[[i]], A[[i]]}, {i ,m}];
```

listu početnih centroida `z0` i pokazatelj `Ind`. Ako je `Ind=0`, ispisat će se početna particija, centroidi njenih klastera i početne vrijednosti kriterijskih funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$.

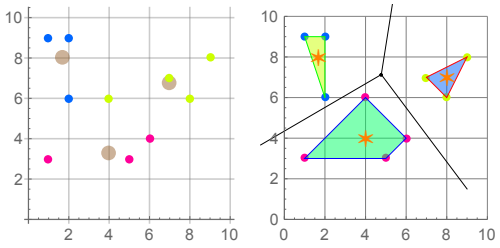
Nakon toga ispisuju se rezultati: optimalna particija, centroidi njenih klastera, vrijednosti kriterijskih funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$ i prikazuju slike početne i optimalne particije. Također ispisuju se vrijednosti `DB` i `CH` indeksa.

```
In[4]:=A = Flatten[Par, 1]; m = Length[A]; k = Length[Par];
       w = Flatten[omega, 1];
       cc = Sum[w[[i]] A[[i]], {i, m}]/Total[w];
                        (* Početne iteracija *)
       z0 = Table[Sum[omega[[j,s]] Par[[j,s]], {s, Length[Par[[j]]]}]/
                 Total[omega[[j]]], {j,k}];
       F0 = Table[Sum[omega[[j, s]] Norm[Par[[j, s]] - z0[[j]]]^2
                 ,{s,Length[Par[[j]]]}], {j,k}];
       G = Sum[Sum[omega[[j,s]], {s, Length[Par[[j]]]}] Norm[cc - z0[[j]]]^2
              ,{j,k}];
       Print["\nIT_0:  PI= ", Par, "\n centri = ", Round[z0, -.01],
             "; c(A)=", Round[cc, -.01], "\nF=", Round[F0, -.01], " = ",
             Round[Total[F0], -.01], "; G=", Round[G, -.01]];
                        (* Slika početne particije *)
       slpod=ListPlot[Table[Table[Par[[j,s]], {s, Length[Par[[j]]]}], {j,k}],
             PlotStyle->Table[{Hue[.3*i+.3], PointSize[.04]}, {i,Length[PI]}],
               Axes -> {True, False}, AxesOrigin -> {0, 0},
               AspectRatio -> Automatic, PlotRange -> {0, 1}];
       slc0=ListPlot[z0, PlotStyle -> {Brown, Opacity[.5], PointSize[.07]}];
       sl1=Show[slc0, slpod, AxesOrigin -> {0,0}, AspectRatio->Automatic,
           PlotRange->{{0,10},{0,10}}, GridLines->Automatic, ImageSize->150];
                           (* Poziv modula *)
       Pod = Table[{w[[i]], A[[i]]}, {i, m}];
       sol = WKMeans2[Pod, z0, 1];
       PI = sol[[1]]; k = Length[PI];
       c = sol[[2]]; F1 = sol[[3]]; IT = sol[[4]] - 1;
       G = Sum[Sum[
               PI[[j,s,1]], {s, Length[PI[[j]]]}]*Norm[cc - c[[j]]]^2, {j,k}];
       Print["\nRjesenje:\nPI= ", PI, "\n centri = ", Round[c, -.01],
             "; c(A)=", Round[cc, -.01], "\nF=", Round[F1, -.01], ";
             G=", Round[G, -.01]];
                              (* Slike *)
```

```
        slc = ListPlot[c, PlotMarkers -> {"\[SixPointedStar]", 15},
                       PlotStyle -> {Orange, PointSize[.025]}];
        PIP = Table[Union[PI[[j, All, 2]]], {j, Length[PI]}];
        konv = Table[ConvexHull[PIP[[j]]], {j, k}];
        slKonv = Graphics[
                 Table[{Opacity[.5], Hue[j/(k+2)], EdgeForm[{Thin, Hue[j/k]}],
                     Polygon[PIP[[j, konv[[j]]]]]}, {j, k}]];
        slVoronoi = DiagramPlot[c, LabelPoints -> False];
        slpod1 = ListPlot[Table[PI[[j, All, 2]], {j,k}],
             PlotStyle->Table[{Hue[.3*i+.3], PointSize[.04]}, {i,Length[PI]}]];
        sl2 = Show[{slVoronoi, slpod1, slKonv, slc}, AxesOrigin -> {0,0},
                      Axes -> True, PlotRange -> {{0,10}, {0,10}},
                          GridLines -> Automatic, ImageSize -> 150];
        Print[GraphicsGrid[{{sl1, sl2}}]]
         (* Indeksi *)
        Print["DB index: ", VDB[PI, c]]
        Print["CH index: ", VCH[PI, c, cc][[3]]]

Out[4]= IT_0:
         PI={{{2,6},{1,9},{2,9}},{{1,3},{6,4},{5,3}},{{4,6},{7,7},{8,6},{9,8}}}
         centri = {{1.67,8.},{4.,3.33},{7.,6.75}}; c(A)={4.5,6.1}
         F={6.67,14.67,16.75} = 38.08; G=85.32

         Rjesenje:
         PI={{{1,{2,6}},{1,{1,9}},{1,{2,9}}},
             {{1,{1,3}},{1,{6,4}},{1,{5,3}},{1,{4,6}}},
             {{1,{7,7}},{1,{8,6}},{1,{9,8}}}}
         centri = {{1.67,8.},{4.,4.},{8.,7.}}; c(A)={4.5,6.1}
         F=30.67; G=92.73
```




```
         St.dev. po klasterima: {1.49,2.24,1.15}
         DB index: 0.76257
         CH index: 10.5837
```

9.4 k-means algoritam uz primjenu ℓ_1-metričke funkcije

k-means algoritam za traženje lokalno optimalne particije skupa podataka $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ uz primjenu ℓ_1-metričke funkcije $d_1 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d_1(x, y) = \|x - y\|_1$ može se opisati s dva koraka (vidi t.4.2, str. 68, odnosno t.4.3, str. 73) koji se iterativno ponavljaju:

`Korak A:` Pridruživanje (assignment step). Poznavanjem međusobno različitih točaka $z_1, \ldots, z_k \in \mathbb{R}^n$, skup $\mathcal{A}$ treba grupirati u k disjunktnih klastera $\pi_1, \ldots, \pi_k$ korištenjem principa minimalnih udaljenosti

$$\pi_j = \{a \in \mathcal{A} : \|z_j - a\|_1 \leq \|z_s - a\|_1,\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k;$$

`Korak B:` Korekcija (update step). Za poznatu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$, treba definirati centre klastera

$$c_j = \operatorname*{med}_{a^s \in \pi_j}(w_s a^s), \qquad j = 1, \ldots, k.$$

Algoritam možemo pokrenuti ili zadavanjem početne particije (tada se najprije pokreće `Korak B`) ili zadavanjem početnih centara (tada se najprije pokreće `Korak A`). Postupak se dalje ponavlja toliko dugo dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera tada također postanu jednaki). U svakom koraku k-means algoritma snižava se vrijednost funkcije cilja $\mathcal{F}_1$ i asimptotski približava lokalno najmanjoj mogućoj vrijednosti [31, 68].

Određivanje težinskog medijana u `Koraku B` (vidi t.2.1.2, str. 6) numerički je složen i zahtjevan postupak (vidi primjerice [24, 50]). Posljednja verzija programskog sustava *Mathematica* omogućuje izračunavanja težinske aritmetičke sredine, ali i težinskog medijana skupa podataka $\mathcal{A}$ s odgovarajućim skupom težina $W > 0$ sljedećim naredbama:

```
Mean[WeightedData[A,W]];
Median[WeightedData[A,W]];
```

Ipak, iz pedagoških razloga pokazat ćemo jedno pojednostavljenje koja dozvoljava implementaciju k-means algoritma uz primjenu ℓ_1-metričke funkcije za podatke s cjelobrojnim težinama. Kao što smo pokazali u Primjeru 2.4, str. 9, težinski medijan skupa $\mathcal{A} \subset \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$, u slučaju kada su težine w_i cijeli brojevi, određuje se slično kao i medijan skupa podataka bez težina. Medijan ovakvog skupa dobije se tako da najprije sortiramo elemente skupa $\mathcal{A}$ s odgovarajućom frekvencijom pojavljivanja i nakon toga odredimo srednji element. Primjerice, medijan skupa $\mathcal{A} = \{3, 1, 4, 5, 9\}$ s težinama $w_i \in \{3, 1, 3, 2, 2\}$ srednji je element u nizu

$$\mathcal{A} = \{1, 3, 3, 3, 4, 4, 4, 5, 5, 9, 9\}.$$

Primjenom *Mathematica*-naredbe `Median[A]` dobivamo $\text{med}(A) = 4$. Zato ćemo za sada k-means algoritam uz primjenu ℓ_1-metričke funkcije konstruirati samo za podatke bez težina, ali ćemo dozvoliti da među podacima može biti i više jednakih. Kasnije, u općem slučaju u t.9.5, str. 187, dozvolit ćemo mogućnost korištenja proizvoljnih težina $w_i > 0$.

Ako težine $w_1, \ldots, w_m > 0$ nisu cijeli brojevi, množenjem svih težina istim faktorom i zaokruživanjem možemo dobiti niz aproksimativnih težina $w'_1, \ldots, w'_m > 0$ kao cijelih brojeva.

9.4.1 Podaci s jednim obilježjem

Neka je $\mathcal{A} = \{a^i \in \mathbb{R} \colon i = 1, \ldots, m\}$ skup podataka s jednim obilježjem među kojima može biti i više jednakih, neka je $z = (z_1, \ldots, z_k)$ niz (lista) međusobno različitih točaka, a `Ind` broj koji može primiti vrijednost „0" ili „1".

k-means algoritam uz primjenu ℓ_1-metričke funkcije pokrenut ćemo zadavanjem početnih centara. Modulu `WKMedian1[A_,z_,Ind_]` predajemo listu podataka $\mathcal{A}$ (među kojima može biti i jednakih), listu `z` i pokazatelj `Ind` $\in \{0, 1\}$.

Nakon ispisa početnih podataka i prikaza početne slike, modul `WKMedian1` izvodi k-means algoritam počevši s `Korakom A`. Ako je `Ind=1`, u svakoj iteraciji ispisuju se međurezultati i odgovarajuća slika. Na kraju modul predaje lokalno optimalnu particiju, centre njenih klastera, vrijednost kriterijske funkcije cilja $\mathcal{F}_1$ i provedeni broj iteracija `IT`.

```
In[1]:=WKMedian1[A_, z_, Ind_] :=
       Module[{PI, tab, imin, z0, c, k=Length[z], m=Length[A],
                podaci, centri, IT=0},
       podaci = Table[{A[[i]], .4}, {i, m}];
       z0 = z;
       centri = Table[{z0[[i]], .4}, {i, k}];
       PI = Table[{}, {i, k}];
       Do[
          tab = Table[Norm[A[[i]] - z0[[j]], 1], {j, k}];
          imin = Ordering[tab, 1][[1]];
          PI[[imin]] = Append[PI[[imin]], A[[i]]]
          ,{i,m}];
                (* Početna iteracija:podaci i slike *)
       FS=Table[Sum[Norm[PI[[j,s]] - z0[[j]], 1], {s, Length[PI[[j]]]}]
                ,{j,k}];
       Fm=Sum[Min[Table[Norm[A[[i]] - z0[[j]], 1 ], {j, k}]], {i, m}];
       Print["\nA = ", A, "\nPocetna pozicija: Assignment points = ",
             Round[z0, -.01], "; F=", Round[FS, -.01], " = ",
                Round[Total[FS], -.01], "; F_min =", Round[Fm, -.01]];
       sl1=ListPlot[podaci,PlotStyle->{Blue,PointSize[.03]},Ticks->{A,None},
            Axes->{True,False}, PlotRange->{0,1}, AspectRatio->Automatic];
```

```
sl2=ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
          PlotStyle -> {Orange, Opacity[.8]}, Axes->{True,False},
             PlotRange -> {0, 1}, AspectRatio -> Automatic];
Print[Show[sl1, sl2, ImageSize -> 300]];
                              (* Petlja *)
While[IT = IT + 1;
 c = Table[Median[PI[[j]]], {j, k}];
 k = Length[PI];
 Chop[Norm[c - z0]] != 0,
 centri = Table[{c[[i]], .4}, {i, k}];
          (* Podaci i slike *)
 If[Ind != 0,
    Print["\nIteracija_", IT, ":  Particija: ", PI, "\nCentri: ",
           N[c], "\nF(PI): ", WFLAD[PI, c] // N];
    Print["F_min=",
    Sum[Min[Table[Norm[c[[j]]-A[[i]],1], {j,k}]], {i,m}]//N];
    sl1=ListPlot[Table[
        Table[{PI[[i,j]],.4}, {j,Length[PI[[i]]]}], {i,Length[PI]}],
        PlotStyle->Table[{Hue[.3*i+.3],PointSize[.03]},{i,Length[PI]}],
        AxesOrigin -> {0, 0}, AspectRatio -> Automatic,
        PlotRange -> {0,1}, Axes -> {True,False}, Ticks -> {A,None}];
    sl2=ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
         PlotStyle -> {Orange, Opacity[.8]}, Axes -> {True, False},
             PlotRange -> {0, 1}, AspectRatio -> Automatic];
    Print[Show[sl1, sl2, ImageSize -> 300]];
     ];
                          (*  *)
 z0 = c;
 PI = Table[{}, {i, k}];
 Do[
     tab = Table[Norm[A[[i]] - z0[[j]], 1], {j, k}];
     imin = Ordering[tab, 1][[1]];
     PI[[imin]] = Append[PI[[imin]], A[[i]]]
     ,{i,m}];
      ];
  {PI, z0, WFLAD[PI, z0], IT}
         ]
```

$\mathcal{F}_1$-kriterijska funkcija cilja računa se u modulu `WFLAD[PI_, c_]`, gdje je `PI` particija s centrima klastera `c`.

```
In[2]:= WFLAD[PI_, c_] := Module[{},
          Return[
            Sum[
              Sum[Norm[c[[j]] - PI[[j, s]], 1], {s, Length[PI[[j]]]}]
                ,{j,Length[PI]}]]
                              ];
```

Izvođenje modula

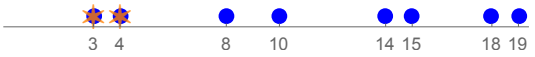
Nakon što se aktiviraju svi moduli, najprije treba učitati skup $\mathcal{A}$, odgovarajući niz težina w i početne centre z. Za Primjer 4.23, str. 85, to izgleda ovako:

```
In[1]:= A = {3, 4, 8, 10, 14, 15, 18, 19}; m = Length[A];
        w = {1, 1, 1, 3, 1, 1, 1, 1};
        z = {3, 4}; k = Length[z];
```

Samo sliku podataka i centre možemo dobiti na sljedeći način:

```
In[2]:= podaci = Table[{A[[i]], .4}, {i, m}];
        centri = Table[{z[[j]], .4}, {j, k}];
        sl1 = ListPlot[podaci, PlotStyle -> {Blue, PointSize[.03]},
              Axes->{True,False}, Ticks -> {A, None}, PlotRange->{0,1},
              AspectRatio->Automatic];
        sl2 = ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
              PlotStyle -> {Orange, Opacity[.8]}, Axes -> {True, False},
              PlotRange -> {0, 1}, AspectRatio -> Automatic];
        Show[sl1, sl2, ImageSize -> 300]
```

Out[3]=



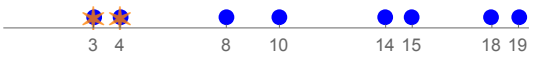
Prije poziva modula `WKMedian1` skup $\mathcal{A}$ redefiniramo tako da broj pojavljivanja nekog elementa $a^i \in \mathcal{A}$ odgovara njegovoj težini w_i. Nakon toga poziva se modul `WKMedian1` s tri argumenta: *podaci*, *centri*, `Ind`. Važno je primijetiti da zbog postupka redefiniranja skupa $\mathcal{A}$ prije svakog pokretanja modula najprije treba učitati ulazne podatke.

Ako je `Ind=0`, ispisat će se redefinirani skup $\mathcal{A}$ s odgovarajućim težinama, početna pozicija s odgovarajućom slikom i rezultati: optimalna particija, centri njenih klastera, vrijednost funkcije cilja $\mathcal{F}_1$ te odgovarajuća slika.

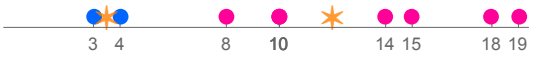
```
In[4]:=                (* Redefiniranje skupa A *)
       Do[
          If[w[[i]] > 1, set = Table[A[[i]], {j, w[[i]] - 1}];
             A = Flatten[Append[A, set], 1]
             ]
          ,{i,m}]; A = Sort[A];
       m = Length[A];
                           (* Poziv modula *)
       sol = WKMedian1[A, z, 1];
                            (* Rezultati *)
```

```
       PI=sol[[1]]; c=sol[[2]]; F=sol[[3]];
       FF=Table[Sum[Norm[PI[[j,s]] - c[[j]],1], {s,Length[PI[[j]]]}],{j,k}];
       Print["\nRjesenje:"]
       Print["LOP: ", PI, "\nCentri: ", Round[c, -.01], "; F= ",
             Round[FF, -.01], " = ", Round[Total[FF], -.01]]
       Print["Prosj.Abs.Odst. po klasterima: ",
             Round[Table[FF[[j]]/Length[PI[[j]]], {j, k}], -.01]]
                          (* Slika *)
       centri = Table[{c[[j]], .4}, {j, k}];
       sl1=ListPlot[Table[Table[{PI[[i, j]], .4}, {j, Length[PI[[i]]]}]
                         ,{i,Length[PI]}],
           PlotStyle->Table[{Hue[.3*i+.3], PointSize[.03]}, {i,Length[PI]}],
           AxesOrigin -> {0, 0}, AspectRatio -> Automatic,
           PlotRange -> {0, 1}, Axes -> {True, False}, Ticks -> {A, None}];
       sl2=ListPlot[centri, PlotMarkers -> {"\[SixPointedStar]", 20},
                PlotStyle -> {Orange, Opacity[.8]}, Axes -> {True, False},
                     PlotRange -> {0,1}, AspectRatio -> Automatic];
       Print[Show[sl1, sl2, ImageSize -> 300]];

Out[5]=A = {3,4,8,10,10,10,14,15,18,19}
       Pocetna pozicija: Assignment points = {3.,4.};
       F={0.,72.} = 72.; F_min =72.
```




```
       Rjesenje:
       LOP: {{3,4},{8,10,10,10,14,15,18,19}}
       Centri: {3.5,12.}; F= {1.,28.} = 29.
       Prosj.Abs.Odst. po klasterima: {0.5,3.5}
```



Ako je `Ind=1`, tj. ako k-means algoritam pozovemo naredbom

```
In[6]:= sol = WKMeans1[Pod, z, 1];
```

ispisat će se numerički pokazatelji svake iteracije uz odgovarajući grafički prikaz kao na Slici 4.19, str. 86.

Ako umjesto početnih centara algoritam želimo pokrenuti zadavanjem početne particije, tada prije pozivanja modula `WKMedian1` treba odrediti listu centara klastera početne particije.

9.4.2 Podaci s dva obilježja

Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ skup podataka s dva obilježja među kojima može biti i više jednakih, neka je $z = (z_1, \ldots, z_k)$ niz (lista) međusobno različitih točaka, a `Ind` broj koji može primiti vrijednost „0" ili „1".

k-means algoritam uz primjenu ℓ_1-metričke funkcije pokrenut ćemo zadavanjem početnih centara. Modulu `WKMedian2[A_,z_,Ind_]` predajemo listu podataka $\mathcal{A}$ (među kojima može biti i jednakih), listu `z` i pokazatelj `Ind` $\in \{0, 1\}$.

Nakon ispisa početnih podataka i početne slike, modul `WKMedian2` izvodi k-means algoritam počevši s `Korakom A`. Ako je `Ind=1`, u svakoj iteraciji ispisuju se međurezultati. Na kraju modul predaje lokalno optimalnu particiju, centre njenih klastera, vrijednost kriterijske funkcije cilja $\mathcal{F}_1$ i provedeni broj iteracija `IT`.

```
In[1]:=WKMedian2[A_, z_, Ind_] :=
       Module[{m=Length[A], k=Length[z], tab, imin, Pod, F0, F1, c, IT=1},
                        (* Pocetna iteracija *)
       PI = Table[{}, {i, k}];
       Do[
          tab = Table[Norm[A[[i]] - z[[j]], 1], {j, k}];
          imin = Ordering[tab, 1][[1]];
          PI[[imin]] = Append[PI[[imin]], A[[i]]]
          ,{i,m}];
       PI = DeleteCases[PI, {}];
       k = Length[PI];   F0 = WFLAD[PI, z];
       c = Table[Median[PI[[j]]], {j, k}];
       F1 = WFLAD[PI, c];
                             (* Petlja *)
       While[
        Chop[F0 - F1] != 0,
        If[Ind != 0,
           FS=Table[Sum[Norm[PI[[j,s]] - c[[j]], 1], {s, Length[PI[[j]]]}]
                    ,{j,Length[PI]}];
           Print["IT: ", IT, "  PI: ", PI, "\n centri: ", Round[c, -.01],
                 "\nF=", Round[FS, -.01], " = ", Round[F1, -.01]]
           ];
       IT = IT + 1;
       F0 = F1;  PI = Table[{}, {i, k}];
       Do[
          tab = Table[Norm[A[[i]] - c[[j]], 1], {j, k}] // N;
          imin = Ordering[tab, 1][[1]];
          PI[[imin]] = Append[PI[[imin]], A[[i]]]
          ,{i,m}];
       PI = DeleteCases[PI, {}];
       k = Length[PI];
```

```
       c = Table[Median[PI[[j]]], {j, k}];
       F1 = WFLAD[PI, c];
           ];
        {PI, c, F1, IT}
                 ]
```

Izvođenje modula

Na početku treba pokrenuti `Mathematica-package` koji omogućava potrebne grafičke prikaze

```
In[1]:= Needs["ComputationalGeometry`"]
```

Implementaciju k-means algoritam uz primjenu ℓ_1-metričke funkcije modulom `WKMedian2[Pod, z, Ind]` možemo pokrenuti zadavanjem početnih centara na sličan način kao što smo to uradili u prethodnoj točki.

U ovom slučaju pokazat ćemo implementaciju k-means algoritma modulom `WKMedian2[Pod, z, Ind]` zadavanjem početne particije. Nakon što se aktiviraju svi moduli, treba učitati početnu particiju (čime je automatski zadan i skup $\mathcal{A}$) i odgovarajuću listu težina $\{omega\}$. Za Primjer 4.19, str. 81, to izgleda ovako:

```
In[1]:= Par = {{{2,6}, {1,9}, {2,9}}, {{1 3}, {6,4}, {5,3}},
               {{4,6}, {7,7}, {8,6}, {9,8}}};
        w = {{3,1,1}, {1,1,1}, {1,1,1,1}};
```

Prije poziva modula `WKMedian2` skup $\mathcal{A}$ redefiniramo tako da broj pojavljivanja nekog elementa $a^i \in \mathcal{A}$ odgovara njegovoj težini w_i. Nakon toga poziva se modul `WKMedian2` s tri argumenta: *podaci*, *centri*, `Ind`. Važno je primijetiti da zbog postupka redefiniranja skupa $\mathcal{A}$ prije svakog pokretanja modula najprije treba učitati ulazne podatke.

Ako je `Ind=0`, ispisat će se redefinirani skup $\mathcal{A}$, početni centri i vrijednost funkcije cilja $\mathcal{F}_1$ te konstruirati odgovarajuća slika.

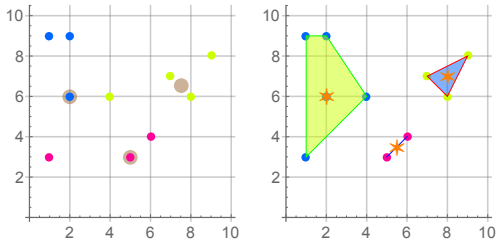
Nakon toga ispisuju se rezultati: optimalna particija s centrima njenih klastera i optimalna vrijednost funkcije cilja $\mathcal{F}_1$. Tada se također prikazuje slika početnih podataka i slika optimalne particije.

```
In[4]:=             (* Redefiniranje skupa A *)
       k = Length[Par]; ParN = Table[{}, {j, k}];
       Do[B = Par[[j]];
          Do[
             If[w[[j, s]] > 1,
                Do[B = Append[B, Par[[j, s]]], {ss, w[[j, s]] - 1}];
```

```
                ]
          ,{s,Length[Par[[j]]]};
       ParN[[j]] = B
      ,{j,k}];
                (* Centri i funkcija cilja *)
      z0 = Table[Median[ParN[[j]]], {j, k}];
      FF=Table[Sum[Norm[ParN[[j,s]]-z0[[j]],1],{s,Length[ParN[[j]]]}],{j,k}];
      Print["\nIT_0: PI=", ParN, "\nCentri: ", Round[z0, -.01],
      "\nF= ", Round[FF, -.01], " = ", Round[Total[FF], -.01]]
                     (* Pocetna slika *)
      slpod=ListPlot[Table[Table[Par[[j,s]], {s, Length[Par[[j]]]}], {j,k}],
            PlotStyle->Table[{Hue[.3*i+.3],PointSize[.04]},{i,Length[PI]}],
               Axes -> {True, False}, AxesOrigin -> {0,0},
                   AspectRatio -> Automatic, PlotRange -> {0, 1}];
      slc0=ListPlot[z0, PlotStyle -> {Brown, Opacity[.5], PointSize[.07]}];
      sl1=Show[slc0, slpod, AxesOrigin -> {0,0}, AspectRatio -> Automatic,
           PlotRange->{{0,10},{0,10}},GridLines->Automatic,ImageSize->150];
                      (* Poziv modula *)
      A = Flatten[ParN, 1]; m = Length[A];
      sol = WKMedian2[A, z0, 0];
                      (* Rezultati *)
      PI = sol[[1]]; k = Length[PI];
      c = sol[[2]]; F1 = sol[[3]]; IT = sol[[4]] - 1;
      FF = Table[{}, {j, k}];
      Do[
        FF[[j]]=Sum[Norm[PI[[j,s]]-c[[j]],1], {s,Length[PI[[j]]]}],{j,k}]
        Print["\nRjesenje:\nIT_",IT,": LOP: ",PI,"\nCentri: ",Round[c,-.01],
              "\nF= ", Round[FF, -.01], " = ",Round[Total[FF], -.01]]
        Print["Prosj.Abs.Odst. po klasterima: ",
              Table[Round[FF[[j]]/Length[PI[[j]]], -.01], {j, k}]]
                      (* Slike + Rješenje *)
      slc = ListPlot[c, PlotMarkers -> {"\[SixPointedStar]", 15},
                     PlotStyle -> {Orange, PointSize[.025]}];
      konv = Table[ConvexHull[PI[[j]]], {j, k}];
      slKonv = Graphics[
           Table[{Opacity[.5], Hue[j/(k + 2)], EdgeForm[{Thin, Hue[j/k]}],
                Polygon[PI[[j, konv[[j]]]]]}, {j, k}]];
      slpod1=ListPlot[Table[PI[[j]], {j, k}],
           PlotStyle->Table[{Hue[.3*i+.3],PointSize[.04]},{i,Length[PI]}]];
      sl2=Show[{slpod1, slKonv, slc}, AxesOrigin -> {0, 0}, Axes -> True,
               PlotRange -> {{0,10}, {0,10}}, GridLines -> Automatic,
                   AspectRatio -> Automatic, ImageSize -> 150];
      Print[GraphicsGrid[{{sl1, sl2}}]]

Out[5]=IT_0: PI={{{2,6},{1,9},{2,9},{2,6},{2,6}},
               {{1,3},{6,4},{5,3}},
               {{4,6},{7,7},{8,6},{9,8}}}
      Centri: {{2.,6.},{5.,3.},{7.5,6.5}}
      F= {7.,6.,9.} = 22.
```

```
Rjesenje:
IT_1: LOP: {{{2,6},{1,9},{2,9},{2,6},{2,6},{1,3},{4,6}},
            {{6,4},{5,3}},  {{7,7},{8,6},{9,8}}}
Centri: {{2.,6.},{5.5,3.5},{8.,7.}}
F= {13.,2.,4.} = 19.
Prosj.Abs.Odst. po klasterima: {1.86,1.,1.33}
```



Ako je `Ind=1`, tj. ako k-means algoritam uz primjenu ℓ_1-metričke funkcije pozovemo naredbom

```
In[6]:= sol = WKMedian2[Pod, z, 1];
```

ispisat će se numerički pokazatelji svake iteracije.

9.5 Opći k-means algoritam za podatke s n obilježja

Potreba za općim k-means algoritmom za podatke s n obilježja iskazana je primjerice kod primjene klaster analize za izučavanje temperaturnih kretanja u gradu Osijeku (vidi t.7, str. 121).

Kao što smo pokazali u t.4.2, str. 68, odnosno u t.4.3, str. 73, k-means algoritam za traženje LOP skupa $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$, čiji elementi imaju težine $w_i > 0$, uz primjenu kvazimetričke funkcije $d_p : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, gdje je

$$d_1(x, y) = \|x - y\|_1, \tag{9.1}$$

$$d_2(x, y) = \|x - y\|_2^2, \tag{9.2}$$

može se opisati s dva koraka koji se iterativno ponavljaju:

`Korak A:` Pridruživanje (assignment step). Poznavanjem međusobno različitih točaka $z_1, \ldots, z_k \in \mathbb{R}^n$, skup $\mathcal{A}$ treba grupirati u k disjunktnih

klastera $\pi_1, \ldots, \pi_k$ korištenjem principa minimalnih udaljenosti

if $p = 2$ **then**

$$\pi_j = \{a \in \mathcal{A}\colon \|z_j - a\|_2 \leq \|z_s - a\|_2,\ \forall s = 1, \ldots, k\},\ j = 1, \ldots, k,$$

else

$$\pi_j = \{a \in \mathcal{A}\colon \|z_j - a\|_1 \leq \|z_s - a\|_1,\ \forall s = 1, \ldots, k\},\ j = 1, \ldots, k,$$

end if

`Korak B:` Korekcija (update step). Poznavanjem particije $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$, treba definirati centre klastera

if $p = 2$ **then**

$$c_j = \frac{1}{W_j} \sum_{a^s \in \pi_j} w_s a^s, \quad W_j = \sum_{a^s \in \pi_j} w_s, \qquad j = 1, \ldots, k,$$

else

$$c_j = \operatorname*{med}_{a^s \in \pi_j} (w_s, a^s), \quad j = 1, \ldots, k,$$

end if

Primijetite da smo algoritam konstruirali tako da se može koristiti i LS-kvazimetrička i ℓ_1-metrička funkcija.

Algoritam možemo pokrenuti ili zadavanjem početne particije (tada se najprije pokreće `Korak B`) ili zadavanjem početnih centara (tada se najprije pokreće `Korak A`). Postupak se dalje ponavlja toliko dugo dok trenutna i prethodna particija ne postanu jednake (centri njihovih klastera tada također postanu jednaki). U svakom koraku k-means algoritma snižava se vrijednost funkcije cilja $\mathcal{F}$ i asimptotski približava lokalno najmanjoj mogućoj vrijednosti [31, 68].

if $p = 2$ **then**

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^s \in \pi_j} \|c_j - a\|_2^2,$$

else

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^s \in \pi_j} \|c_j - a\|_1,$$

end if

Neka je $z = (z_1, \ldots, z_k)$ niz (lista) međusobno različitih točaka, `p`$\in \{1, 2\}$ broj koji može primiti vrijednost „1" ili „2", a `Ind`$\in \{0, 1\}$ broj koji može primiti vrijednost „0" ili „1". k-means algoritam pokrenut ćemo zadavanjem početnih centara $z = (z_1, \ldots, z_k)$. Modulu `WKMeans[Pod_,z_,p_,Ind_]` predajemo skup podataka sastavljenih od parova {težina, podatak}:

```
Pod = Table[{w[[i]], A[[i]]}, {i,Length[A]}],
```

listu `z` i pokazatelje `p` $\in \{1, 2\}$ i `Ind` $\in \{0, 1\}$.

Nakon ispisa početnih podataka, Modul `WKMeans` izvodi k-means algoritam počevši s `Korakom A`. Ako je `Ind=1`, u svakoj iteraciji ispisuju se međurezultati. Na kraju modul predaje lokalno optimalnu particiju, centre njenih klastera, vrijednost kriterijske funkcije cilja $\mathcal{F}$ i provedeni broj iteracija `IT`.

```
In[1]:=WKmeans[Pod_, z_, p_, Ind_] :=
       Module[{x, PI, tab, imin, z0, c, centri, m = Length[Pod],
       k = Length[z], j1, mj},
       z0 = z;
       PI = Table[{}, {i, k}];
                (* Princip minimalnih udaljenosti *)
       Do[
          tab = Table[d[Pod[[i]][[2]], z0[[j]], p], {j, k}];
          imin = Ordering[tab, 1][[1]];
          PI[[imin]] = Append[PI[[imin]], Pod[[i]]]
          ,{i,m}];
       If[Ind != 0, Print["Particija", PI, ";  Centri: ", z0];
                    Print["F: ", WF[PI, z0, p]]];
                            (* Centri *)
       c = Table[0, {j, k}];
       While[
          Do[j1 = k - j + 1; mj = Length[PI[[j1]]];
             If[mj != 0,
              If[p == 2,
                 c[[j1]]=Mean[WeightedData[PI[[j1,All,2]], PI[[j1,All,1]]]],
                 c[[j1]]=Median[WeightedData[PI[[j1,All,2]], PI[[j1,All,1]]]]],
              PI=Drop[PI,{j1}]; c=Drop[c,{j1}]; z0=Drop[z0,{j1}] ]
             ,{j,k}];
          k = Length[PI];
          Chop[Norm[c - z0]] != 0,
          If[Ind != 0, centri = Table[{c[[i]], .5}, {i, k}];
          Print[Table[{PI[[j]], centri[[j]]}, {j, k}]];
             ];
       z0 = c;
       PI = Table[{}, {i, k}];
                (* Princip minimalnih udaljenosti *)
       Do[
```

```
      tab = Table[d[Pod[[i]][[2]], z0[[j]], p], {j, k}];
      imin = Ordering[tab, 1][[1]];
      PI[[imin]] = Append[PI[[imin]], Pod[[i]]]
      ,{i,m}];
    If[Ind != 0, Print["Particija", PI, ";  Centri: ", c];
                 Print["F: ", WF[PI, z0, p]]
      ]
        ];
     {PI, N[z0], WF[PI, z0, p]}
          ]
```

Kvazimetrička funkcija `d[apod_, bpod_, p_]` za `p` $\in \{1, 2\}$ definira se na sljedeći način

```
In[2]:=d[apod_, bpod_, p_] := Norm[apod - bpod, p]^p;
```

Kriterijska funkcija cilja računa se u modulu `WF[PI_, c_, p_]`, gdje je `PI` particija s centrima klastera `c`, a `p` $\in \{1, 2\}$.

```
In[3]:=WF[PI_, c_, p_] :=
        Sum[Sum[PI[[j,s,1]] d[PI[[j s,2]], c[[j]], p], {s,Length[PI[[j]]]}]
            ,{j,Length[PI]}];
```

Izvođenje modula

Nakon što se aktiviraju svi moduli, najprije treba učitati skup $\mathcal{A}$, odgovarajući niz težina w i početne centre. Za Primjer 4.18, str. 80, to izgleda ovako:

```
In[1]:= A={{1,9},{2,9},{2,6},{1,3},{5,3},{6,4},{4,6},{7,7},{8,6},{9,8}};
        m=Length[A];
        w = Table[1, {i, m}];
        Pod = Table[{w[[i]], A[[i]]}, {i, m}];
        cen = {{2, 8}, {5, 4}, {6, 6}};
        k = Length[cen];
```

Implementaciju k-means algoritma izvodimo pozivom modula `WKMeans` kojemu najprije predajemo *podatke* oblika `Pod=Table[{w[[i]],A[[i]]}, {i,m}]`, listu *centara* `z`, parametar `p` i pokazatelj `Ind`. Ako je `Ind=0`, ispisat će se skup $\mathcal{A}$ s odgovarajućim težinama i rezultati: optimalna particija, centri njenih klastera, vrijednost kriterijske funkcija cilja $\mathcal{F}$ i vrijednosti `DB` i `CH` indeksa (ako je `p`=2).

```
In[4]:=p=2;
       sol = WKmeans[Pod, cen, p, 0];
       Print["Broj klastera: ", k = Length[sol[[2]]] ]
       Print["Particija: ", sol[[1]]]
       Print["Centri: ", sol[[2]]]
       Print["|pi|: ", Table[Length[sol[[1, j]]], {j, k}]]
       Print["F = ", sol[[3]]//N];
       If[p == 2,
          cc = Sum[w[[i]] A[[i]], {i, m}]/Total[w];
          Print["CH = ", VCH[sol[[1]], sol[[2]], cc], "\nDB = ",
                    VDB[sol[[1]], sol[[2]]]]
           ]

Out[5]= Broj klastera = 3
        Particija: {{{1,{1,9}}, {1,{2,9}}, {1,{2,6}}},
                    {{1,{1,3}}, {1,{5,3}}, {1,{6,4}}, {1,{4,6}}},
                    {{1,{7,7}}, {1,{8,6}}, {1,{9,8}}}}
        Centri: {{1.66667,8.},{4.,4.},{8.,7.}}
        |pi|: {3,4,3}
        F = 30.6667
        CH = 19.9436
        DB = 0.308969
```

Ako je `Ind=1`, tj. ako k-means algoritam pozovemo naredbom

```
In[6]:= sol = WKMeans[Pod, z, p,1];
```

ispisat će se numerički pokazatelji svake iteracije.

Ako umjesto početnih centroida algoritam želimo pokrenuti zadavanjem početne particije, tada prije pozivanja modula `WKMeans` treba odrediti listu centroida `z` klastera početne particije.

Literatura

[1] A. M. Bagirov, J. Ugon, D. Webb, *Fast modified global k-means algorithm for incremental cluster construction*, Pattern Recognition, **44**(2011) 866–876.

[2] D. Bakić, *Linearna algebra*, Školska knjiga, Zagreb, 2008.

[3] M. Benšić, N. Šuvak, *Primijenjena statistika*, Odjel za matematiku, Sveučilište u Osijeku, 2012.

[4] M. W. Berry, Z. Drmač, E. R. Jessup, *Using linear algebra for information retrieval*, SIAM Review, **41**(1999) 335–362.

[5] M. W. Berry, J. Kogan, *Text Mining. Applications and Theory*, Wiley, 2010.

[6] J. C. Bezdek, J. Keller, R. Krisnapuram, N. R. Pal, *Fuzzy models and algorithms for pattern recognition and image processing*, Springer, 2005.

[7] D. Blanuša, *Viša matematika*, Tehnička knjiga, Zagreb, 1973.

[8] R. J. Boscovich, *De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura eius ex exemplaria etiam sensorum impressa*, Bononienci Scientiarum et Artium Znstituto Atque Academia Commentarrii, **4**(1757) 353–396.

[9] D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[10] S. Butenko, W. A. Chaovalitwongse, P. M. Pardalos, *Clustering Challenges in Biological Networks*, World Scientific, 2009.

[11] T. Calinski, J. Harabasz, *A dendrite method for cluster analysis*, Communications in Statistics, **3**(1974) 1–27.

[12] R. Cupec, R. Grbić, K. Sabo, R. Scitovski, *Three points method for searching the best least absolute deviations plane*, Applied Mathematics and Computation, **215**(2009) 983–994.

[13] L. Čaklović, *Zbirka zadataka iz linearne algebre*, Školska knjiga, Zagreb, 1992.

[14] D. Davies, D. Bouldin, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **2**(1979) 224–227.

[15] I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, In: *Proceedings of the 10-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA*, 2004, 551–556.

[16] Y. Dodge, editor, *Statistical data analysis based on the L1-norm and related methods, Proceedings of the Third International Conference on Statistical Data Analysis Based on the L1-norm and Related Methods.* Elsevier, 1997.

[17] Z. Drezner, H. W. Hamacher, *Facility Location: Applications and Theory*, Springer, 2004.

[18] N. Elezović, A. Aglić, *Linearna algebra – zbirka zadataka*, Element, Zagreb, 2003.

[19] B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.

[20] D. E. Finkel, *DIRECT Optimization Algorithm User Guide*, Center for Research in Scientific Computation. North Carolina State University, 2003, `http://www4.ncsu.edu/definkel/research/index.html`.

[21] J. M. Gablonsky, *Direct version 2.0*, Technical report, Center for Research in Scientific Computation. North Carolina State University, 2001.

[22] R. Grbić, D. Grahovac, R. Scitovski, *A method for solving the multiple ellipses detection problem*, Pattern Recognition, **60**(2016) 824–834.

[23] R. Grbić, E. K. Nyarko, R. Scitovski, *A modification of the DIRECT method for Lipschitz global optimization for a symmetric function*, Journal of Global Optimization, **57**(2013) 1193–1212.

[24] C. Gurwitz, *Weighted median algorithms for l_1 approximation*, BIT, **30**(1990) 301–310.

[25] E. M. T. Hendrix, B. G. Tóth, *Introduction to Nonlinear and Global Optimization*, Springer, 2010.

[26] C. Iyigun, A. Ben-Israel, *A generalized weiszfeld method for the multi-facility location problem*, Operations Research Letters, **38**(2010) 207–214.

[27] D. R. Jones, C. D. Perttunen, B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, Journal of Optimization Theory and Applications, **79**(1993) 157–181.

[28] D. Jukić, *Mjera i integral*, Odjel za matematiku, Sveučilište u Osijeku, 2012.

[29] D. Jukić, R. Scitovski, *Matematika I*, Odjel za matematiku, Sveučilište u Osijeku, 2004.

[30] L. Kaufman, P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, Chichester, UK, 2005.

[31] J. Kogan, *Introduction to Clustering Large and High-dimensional Data*, Cambridge University Press, New York, 2007.

[32] S. Kurepa, *Uvod u linearnu algebru*, Školska knjiga, Zagreb, 1985.

[33] S. Kurepa, *Matematička analiza II*, Tehnička knjiga, Zagreb, 1990.

[34] F. Leisch, *A toolbox for k-centroids cluster analysis*, Computational Statistics & Data Analysis, **51**(2006) 526–544.

[35] S. Lipschitz, *3000 Solved Problems in Linear Algebra*, McGraw-Hill, New York, 1989.

[36] S. Lipschitz, *Beginning Linear Algebra*, McGraw Hill, New York, 1997.

[37] R. Manger, *Strukture podataka i algoritmi*, Element, Zagreb, 2014.

[38] T. Marošević, K. Sabo, P. Taler, *A mathematical model for uniform distribution of voters per constituencies*, Croatian Operational Research Review, **4**(2013) 53–64.

[39] T. Marošević, R. Scitovski, *Multiple ellipse fitting by center-based clustering*, Croatian Operational Research Review, **6**(2015) 43–53.

[40] D. Matijević, N. Truhar, *Uvod u računarstvo*, Odjel za matematiku, Sveučilište u Osijeku, 2012.

[41] D. J. Maširević, S. Miodragović, *Geometric median in the plane*, Elemente der Mathematik, **70**(2015) 21–32.

[42] B. Mirkin, *Data clustering for Data Mining*, Chapman & Hall/CRC, 2005.

[43] A. Morales-Esteban, F. Martínez-Álvarez, S. Scitovski, R. Scitovski, *A fast partitioning algorithm using adaptive mahalanobis clustering with application to seismic zoning*, Computers & Geosciences, **73**(2014) 132–141.

[44] H. Neunzert, W. G. Eschmann, A. Blickensdörfer-Ehlers, *Analysis 2. Mit einer Einführung in die Vektor- und Matrizenrechnung*, Springer-Verlag, Berlin, 1991.

[45] J. Parajka, S. Kohnová, G. Bálint, M. Barbuc, M. Borga, P. Claps, S. C. A. Dumitrescu, E. Gaume, K. Hlavčová, R. Merz, M. Pfaundler, G. Stancalie, J. Szolgay, G. Blöschl, *Seasonal characteristics of flood regimes across the alpine–carpathian range*, Journal of Hydrology, **394**(2010) 78–89.

[46] R. Paulavičius, J. Žilinskas, *Simplicial Global Optimization*, volume X of *Series: Springer Briefs in Optimization*, Springer-Verlag, Berlin, 2014.

[47] K. Pollard, M. van der Laan, *A method to identify significant clusters in gene expression data*, In: *Proceedings of SCI, Vol. II, 318-325*, 2002.

[48] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge, 1992.

[49] L. Rotaru, *Identifying the phenotypic resemblances of the vine breeds by means of cluster analysis*, Notulae Botanicae, **37**(2009) 249–252.

[50] K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal, **50**(2008) 185–198.

[51] K. Sabo, R. Scitovski, *An approach to cluster separability in a partition*, Information Sciences, **305**(2015) 208–218.

[52] K. Sabo, R. Scitovski, P. Taler, *Uniform distribution of the number of voters per constituency on the basis of a mathematical model (in Croatian)*, Hrvatska i komparativna javna uprava, **14**(2012) 229–249.

[53] K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka - klasteri*, Osječki matematički list, **10**(2010) 149–178.

[54] K. Sabo, R. Scitovski, I. Vazler, *One-dimensional center-based l_1-clustering method*, Optimization Letters, **7**(2013) 5–22.

[55] K. Sabo, R. Scitovski, I. Vazler, M. Zekić-Sušac, *Mathematical models of natural gas consumption*, Energy Conversion and Management, **52**(2011) 1721–1727.

[56] K. Sabo, P. Taler, Z. Bertić, *Mathematics and politics: How to determine optimal constituencies in Republic of Croatia*, Croatian Operational Research News, **2(1)**(2015) 10–12.

[57] A. Schöbel, *Locating Lines and Hyperplanes: Theory and Algorithms*, Springer Verlag, Berlin, 1999.

[58] R. Scitovski, *Problemi najmanjih kvadrata. Financijska matematika*, Ekonomski fakultet, Elektrotehnički fakultet, Sveučilište u Osijeku, 1993.

[59] R. Scitovski, *Numerička matematika*, Odjel za matematiku, Sveučilište u Osijeku, 2015, 3rd edition.

[60] R. Scitovski, T. Marošević, *Multiple circle detection based on center-based clustering*, Pattern Recognition Letters, **52**(2014) 9–16, Accepted.

[61] R. Scitovski, K. Sabo, *Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters*, Knowledge-Based Systems, **57**(2014) 1–7.

[62] R. Scitovski, K. Sabo, D. Grahovac, *Globalna optimizacija*, Odjel za matematiku, 2015.

[63] R. Scitovski, S. Scitovski, *A fast partitioning algorithm and its application to earthquake investigation*, Computers & Geosciences, **59**(2013) 124–131.

[64] R. Scitovski, S.Kosanović, *Rate of change in economics research*, Economics analysis and workers management, **19**(1985) 65–75.

[65] R. Scitovski, N. Truhar, Z. Tomljanović, *Metode optimizacije*, Odjel za matematiku, Sveučilište u Osijeku, 2014.

[66] R. Scitovski, I. Vidović, D. Bajer, *A new fast fuzzy partitioning algorithm*, Expert Systems with Applications, **51**(2016) 143–150.

[67] S. Scitovski, N. Šarlija, *Cluster analysis in retail segmentation for credit scoring*, Croatian Operational Research Review, **5**(2014) 235–245.

[68] H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

[69] D. Steinley, M. J. Brusco, *Initializing k-means batch clustering: a critical evaluation of several techniques*, Journal of Classification, **24**(2007) 99–121.

[70] N. Šarlija, M. Benšić, M. Zekić-Sušac, *Comparison procedure of predicting the time to default in behavioural scoring*, Expert Systems with Applications, **36**(2009) 8778–8788.

[71] P. N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Wesley, 2006.

[72] M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research, **8**(2007) 65–102.

[73] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, Burlington, 2009, 4^{th} edition.

[74] N. Truhar, *Numerička linearna algebra*, Odjel za matematiku, Sveučilište u Osijeku, 2010.

[75] Ž. Turkalj, D. Markulak, S. Singer, R. Scitovski, *Research project grouping and ranking by using adaptive mahalanobis clustering*, Croatian Operational Research Review, **7**(2016) 81–96.

[76] I. Vazler, K. Sabo, R. Scitovski, *Weighted median of the data in solving least absolute deviations problems*, Communications in Statistics - Theory and Methods, **41:8**(2012) 1455–1465.

[77] D. Veljan, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.

[78] L. Vendramin, R. J. G. B. Campello, E. R. Hruschka, *On the comparison of relative clustering validity criteria*, In: *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 – May 2, 2009, Sparks, Nevada, USA*, SIAM, 2009, 733–744.

[79] I. Vidović, D. Bajer, R. Scitovski, *A new fusion algorithm for fuzzy clustering*, Croatian Operational Research Review, **5**(2014) 149–159.

[80] I. Vidović, R. Scitovski, *Center-based clustering for line detection and application to crop rows detection*, Computers and Electronics in Agriculture, **109**(2014) 212–220.

[81] V. Volkovich, J. Kogan, C. Nicholas, *Building initial partitions through sampling techniques*, European Journal of Operational Research, **183**(2007) 1097–1105.

Kazalo