Odjel za matematiku
Sveučilište J. J. Strossmayera u Osijeku

Matematički praktikum

Grupiranje podataka R. Scitovski, K. Sabo, I. Vazler

1 Uvod

Definicija 1. Neka je $\mathcal{A}$ skup s $m \geq 2$ elemenata i $1 \leq k \leq m$. Rastav skupa $\mathcal{A}$ na k podskupove $\pi_1, \ldots, \pi_k$, tako da bude

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

zovemo *particija skupa* $\mathcal{A}$, a skupove $\pi_1, \ldots, \pi_k$ *klasteri*. Skup svih particija skupa $\mathcal{A}$ sastavljenih od k klastera koje zadovoljavaju *(i)-(iii)* označit ćemo 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 $\mathcal{A}$.

Sinonimi: *grupiranje, segmentiranje, klasifikacija, rangiranje*
En.: cluster analysis, clustering, data mining

Može se pokazati (Veljan, 2001) da je broj svih particija skupa $\mathcal{A}$ iz Definicije 1 jednak Stirlingovom broju druge vrste

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

Primjer 1. *Broj svih particija skupa $\mathcal{A}$ koje zadovoljavaju Definiciju 1 specijalno za $m = 10, 50, 10^3, 10^6$ i $k = 2, 3, 5, 8, 10$ iznosi*

$\lvert\mathcal{P}(\mathcal{A}, k)\rvert$	$k=2$	$k=3$	$k=5$	$k=8$	$k=10$
$m=10$	511	9330	42525	750	1
$m=50$	10^{15}	10^{23}	10^{33}	10^{40}	10^{43}
$m=10^3$	10^{300}	10^{476}	10^{697}	10^{898}	10^{993}
$m=10^6$	$10^{301\,029}$	$10^{477\,120}$	$10^{698\,968}$	$10^{903\,085}$	10^{10^6}

Iz navedenog primjera vidi se da traženje optimalne particije općenito neće biti moguće provesti pretraživanjem čitavog skupa $\mathcal{P}(\mathcal{A}, k)$. Odmah teba reći da problem traženja optimalne particije spada u NP-teške probleme (Gan et al., 2007) nekonveksne optimizacije općenito nediferencijabilne funkcije više varijabli, koja najčešće posjeduje značajan broj stacionarnih točaka.

Primjene:
poljoprivreda (primjerice, razvrstavanje oranica prema plodnosti zemljišta);
biologija (primjerice, klsasifikacija kukaca u grupe)
medicina (primjerice, analiza rengenskih slika)
promet (primjerice, identifikacija prometnih "čepova")
analiza i pretraživanje teksta
analiza klimatskih kretanja
donošenje raznih odluka u tijelima državne i lokalne administracije.
definiranje izbornih sustava

Programska podrška (Sabo et al., 2011):
`http://www.mathos.hr/oml/software.htm`

2 Motivacija

$\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ – podskup realnih brojeva
$\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$ – particijaskupa $\mathcal{A}$, takva da vrijedi

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

$|\mathcal{P}(\mathcal{A}, k)| = 2^{m-1} - 1$ – broj svih ovakvih particija
$d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ – kvazimetrička funkcija

Uz primjenu LS-kvazimetričke funkcije $d(x, y) = (x - y)^2$, za klastere π_1, π_2 odredimo reprezentante – centroide

$$c_1 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_1} d(x, a_i) = \frac{1}{m_1} \sum_{a_i \in \pi_1} a_i,$$
$$c_2 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_2} d(x, a_i) = \frac{1}{m_2} \sum_{a_i \in \pi_2} a_i.$$

Mjeru kvalitete particije $\Pi = \{\pi_1, \pi_2\}$ s centroidima c_1, c_2 definirat ćemo funkcijom cilja $\mathcal{F}\colon \mathcal{P}(\mathcal{A}, 2) \to \mathbb{R}_+$

$$\mathcal{F}(\Pi) = \sum_{a_i \in \pi_1} d(c_1, a_i) + \sum_{a_i \in \pi_2} d(c_2, a_i), \tag{2}$$

koja predstavlja sumu "kvadratnog rasipanja" točaka klastera π_1 do centroida c_1 i točaka klastera π_2 do centroida c_2.

Problem: *pronaći onu particiju na kojoj funkcija cilja $\mathcal{F}$ postiže najmanju vrijednost*

Primjer 2. *Treba pronaći sve particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$, odrediti pripadne centroide i vrijednosti funkcije cilja $\mathcal{F}$.*

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
$\{1\}$	$\{3,4,8\}$	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

Obratno, za dane realne brojeve $c_1, c_2 \in \mathbb{R}$, $c_1 \neq c_2$, primjenom **principa minimalnih udaljenosti** možemo definirati particiju $\Pi = \{\pi_1, \pi_2\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_1 = \{a_i \in \mathcal{A} : d(a_i, c_1) \leq d(a_i, c_2)\},$$
$$\pi_2 = \{a_i \in \mathcal{A} : d(a_i, c_2) < d(a_i, c_1)\},$$

pri čemu treba voditi računa da svaki element skupa $\mathcal{A}$ pridružimo samo jednom klasteru. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može razmatrati kao sljedeći optimizacijski problem

$$\min_{c_1,c_2\in\mathbb{R}} F(c_1, c_2), \qquad F(c_1, c_2) = \sum_{i=1}^{m} \min\{d(c_1, a_i), d(c_2, a_i)\}, \tag{3}$$

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

Primjedba 1. Primijetite da klasteri π_1, π_2 ovise o centrima c_1, c_2 i da vrijedi

$$\mathcal{F}(\Pi) := \sum_{a_i\in\pi_1} d(c_1, a_i) + \sum_{a_i\in\pi_2} d(c_2, a_i) \leq \sum_{i=1}^{m} \min\{d(c_1, a_i), d(c_2, a_i)\} =: F(c_1, c_2), \tag{4}$$

pri čemu se jednakost postiže na lokalno optimalnoj particiji. Naime, vrijedi

$$\begin{aligned} F(c_1, c_2) &= \sum_{a_i\in\pi_1(c_1,c_2)} \min\{d(c_1, a_i), d(c_2, a_i) + \sum_{a_i\in\pi_2(c_1,c_2)} \min\{d(c_1, a_i), d(c_2, a_i) \\ &\geq \sum_{a_i\in\pi_1(c_1,c_2)} d(c_1, a_i) + \sum_{a_i\in\pi_2(c_1,c_2)} d(c_2, a_i) = \mathcal{F}(\Pi). \end{aligned}$$

Algoritam 1. (Standardni k-means algorithm)

Step 1: Inicijalizacija: $z_1 < z_2$;
$F(z_1, z_2) = \sum_{i=1}^{m} \min\{d(z_1, a_i), d(z_2, a_i)\}$

Step 2: Pridruživanje (assignment step)

$$\pi_1 = \{a_i \in \mathcal{A} : d(z_1, a_i) \leq d(z_2, a_i)\},$$
$$\pi_2 = \{a_i \in \mathcal{A} : d(z_2, a_i) < d(z_1, a_i)\},$$

Step 3: Korekcija (update step)

$$\zeta_1 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_1} d(x, a_i)$$
$$\zeta_2 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_2} d(x, a_i)$$

$$\mathcal{F}(\Pi(\pi_1, \pi_2)) = \sum_{a \in \pi_1} d(\zeta_1, a) + \sum_{a \in \pi_2} d(\zeta_2, a)$$

Primjedba 2. Step 2 i Step 3 se izmjenjuju tako dugo dok se ili centri ne poklope ili dok se particije ne poklope ili dok vrijednost funkcije cilja ne prestane opadati. Primijetite da se u Step 2 novi klasteri π_1, π_2 geometrijski mogu odrediti tako da odredimo polovište spojnice centroida. Tada svi elementi lijevo od polovišta pripadaju novom klasteru π_1, a svi elementi desno od polovišta pripadaju novom klasteru π_2.

Primjer 3. $\mathcal{A} = \{1, 2, 6, 7, 9\}$, $k = 2$,

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	4	8	19	$\{1, 2, 6\}$	$\{7, 9\}$	3	8	16
2	3	8	11	$\{1, 2\}$	$\{6, 7, 9\}$	$\frac{3}{2}$	$\frac{22}{3}$	$\frac{31}{6} \approx 5.1667$
3	$\frac{3}{2}$	$\frac{22}{3}$	5.1667	$\{1, 2\}$	$\{6, 7, 9\}$	$\frac{3}{2}$	$\frac{22}{3}$	5.1667

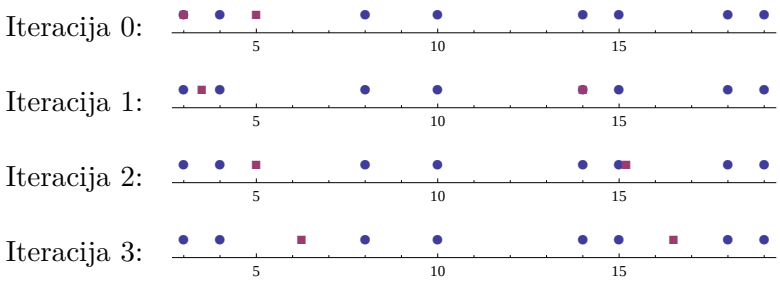
Primjer 4. $\mathcal{A} = \{0, 2, 3\}$, $k = 2$,

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	1	3	2	$\{0, 2\}$	$\{3\}$	1	3	2
1	1	3	2	$\{0, 2\}$	$\{3\}$	1	3	2

Ovo je primjer koji pokazuje da standardni k-means algoritam ne daje optimalno rješenje. Bolja particija je $\Pi_1 = \{\{0\}, \{2, 3\}\}$ jer je $F(\Pi_1) = 0.25$.

Primjer 5. $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$, $k = 2$,

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	3	5	581	$\{3, 4\}$	$\{8, 10, 14, 15, 18, 19\}$	3.5	14	94.5
2	3.5	14	78.75	$\{3, 4, 8\}$	$\{10, 14, 15, 18, 19\}$	5	15.2	64.8
3	5	15.2	62.76	$\{3, 4, 8, 10\}$	$\{14, 15, 18, 19\}$	6.25	16.5	49.75
4	6.25	16.5	49.756	$\{3, 4, 8, 10\}$	$\{14, 15, 18, 19\}$	6.25	16.5	49.75



2.1 Traženje optimalne particije

Zadatak 1. *Neka su $f, g \in C^2(\mathbb{R})$ dvije funkcije za koje vrijedi*

$$f(x) + g(x) = const, \qquad \forall x \in \mathbb{R}.$$

Pokažite da tada vrijedi

(i) Ako funkcija f u točki $x_0 \in \mathbb{R}$ postiže lokalni (odnosno globalni) minimum, onda funkcija g u toj točki postiže lokalni (odnosno globalni) maksimum;

(ii) Ako funkcija f u točki $x_0 \in \mathbb{R}$ postiže lokalni (odnosno globalni) maksimum, onda funkcija g u toj točki postiže lokalni (odnosno globalni) minimum;

Lema 1. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ skup podataka, a $\Pi = \{\pi_1, \pi_2\}$ neka particija s klasterima π_1, π_2 i centrima*

$$c_1 = \frac{1}{m_1} \sum_{a_i \in \pi_1} a_i, \quad c_2 = \frac{1}{m_2} \sum_{a_i \in \pi_2} a_i, \qquad m_1 = |\pi_1|, \quad m_2 = |\pi_2|.$$

Tada vrijedi

$$\mathcal{F}(c_1, c_2) + \mathcal{G}(c_1, c_2) = \sum_{i=1}^{m} (a_i - \bar{c})^2, \quad \bar{c} = \frac{1}{m} \sum_{i=1}^{m} a_i, \tag{5}$$

gdje je

$$\mathcal{F}(c_1, c_2) = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2, \tag{6}$$

$$\mathcal{G}(c_1, c_2) = m_1 (c_1 - \bar{c})^2 + m_2 (c_2 - \bar{c})^2. \tag{7}$$

Dokaz. Za svaki $x \in \mathbb{R}$ vrijedi

$$\sum_{a_i \in \pi_j} (a_i - x)^2 = \sum_{a_i \in \pi_j} (a_i - c_j)^2 + m_j (c_j - x)^2, \quad j = 1, 2. \tag{8}$$

Primjerice,

$$\begin{aligned} \sum_{a_i \in \pi_1} (a_i - x)^2 &= \sum_{a_i \in \pi_1} ((a_i - c_1) + (c_1 - x))^2 \\ &= \sum_{a_i \in \pi_1} (a_i - c_1)^2 + m_1 (c_1 - x)^2, \\ \text{jer je} \quad \sum_{a_i \in \pi_1} (a_i - c_1)(c_1 - x) &= (c_1 - x) \sum_{a_i \in \pi_1} (a_i - c_1) = 0. \end{aligned}$$

Ako u (8) stavimo $x = \bar{c}$ i zbrojimo obje jednakosti, dobivamo

$$\sum_{a_i \in \pi_1} (a_i - \bar{c})^2 + \sum_{a_i \in \pi_2} (a_i - \bar{c})^2 = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2 + m_1 (c_1 - \bar{c})^2 + m_2 (c_2 - \bar{c})^2.$$

Kako je

$$\sum_{a_i \in \pi_1} (a_i - \bar{c})^2 + \sum_{a_i \in \pi_2} (a_i - \bar{c})^2 = \sum_{i=1}^{m} (a_i - \bar{c})^2$$

slijedi (5). □

Teorem 1. *Neka su $F, G \in C^2(\mathbb{R}^2)$ dvije funkcije za koje vrijedi*

$$F(x, y) + G(x, y) = const, \qquad \forall (x, y) \in \mathbb{R}^2.$$

Tada vrijedi:

(i) Ako funkcija F u točki $(x^\star, y^\star) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) minimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) maksimum;

(ii) Ako funkcija F u točki $(x^\star, y^\star) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) maksimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) minimum;

Dokaz. Kako je

$$\frac{\partial F}{\partial x} + \frac{\partial G}{\partial x} = 0, \qquad \frac{\partial F}{\partial y} + \frac{\partial G}{\partial y} = 0,$$

onda se stacionarne točke funkcija F i G podudaraju. Nadalje, kako je

$$\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 G}{\partial x^2} = 0, \qquad \frac{\partial^2 F}{\partial y \partial x} + \frac{\partial^2 G}{\partial x \partial y} = 0, \qquad \frac{\partial^2 F}{\partial y^2} + \frac{\partial^2 G}{\partial y^2} = 0,$$

onda za Hessijane funkcija F, G vrijedi

$$H_F(x, y) = -H_G(x, y),$$

iz čega slijede traženi zaključci. □

Zadatak 2. *Neka su $F, G \in C(\mathbb{R})$ dvije neprekidne funkcije za koje vrijedi*

$$F(x,y) + G(x,y) = const, \qquad \forall (x,y) \in \mathbb{R}^2.$$

Pokažite da tada vrijedi:

(i) Ako funkcija F u točki $(x^\star, y^\star) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) minimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) maksimum;

(ii) Ako funkcija F u točki $(x^\star, y^\star) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) maksimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) minimum;

Korolar 1. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ skup podataka. Tada optimalnu dvočlanu particiju $\Pi^\star = \{\pi_1^\star, \pi_2^\star\}$ skupa $\mathcal{A}$ možemo tražiti tako da tražimo globalni minimum funkcije (princip minimalnih kvadratnih udaljenosti od centara klastera)*

$$\mathcal{F}(c_1, c_2) = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2,$$

gdje je

$$\pi_1 = \{a_i \in \mathcal{A} \colon |a_i - c_1| \leq |a_i - c_2|\}, \qquad \pi_2 = \{a_i \in \mathcal{A} \colon |a_i - c_2| < |a_i - c_1|\},$$

odnosno, sukladno Primjedbi 1, minimum funkcije

$$F(c_1, c_2) = \sum_{i=1}^{m} \min\{(a_i - c_1)^2, (a_i - c_2)^2\},$$

ili tako da tražimo globalni maksimum funkcije (princip maksimalne kvadratne udaljenosti centara klastera)

$$\mathcal{G}(c_1, c_2) = m_1(c_1 - \bar{c})^2 + m_2(c_2 - \bar{c})^2, \qquad \bar{c} = \tfrac{1}{m} \sum_{i=1}^{m} a_i,$$

gdje je $m_1 = |\pi_1|$, $m_2 = |\pi_2|$.

Primjedba 3. Fizikalni smisao bio bi traženje "dva težišta" diskretnog materijalnog tijela.

3 Grupiranje podataka s jednim obilježjem

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup koji na osnovi samo jednog obilježja treba grupirati u k klastera koji zadovoljavaju Definiciju 1. Primjerice, dane u godini možemo grupirati prema prosječnoj dnevnoj temperaturi izraženoj u °C. Svaki element $a_i \in \mathcal{A}$ temeljem tog obilježja reprezentirat ćemo jednim realnim brojem, kojeg ćemo također označavati s a_i. Zato ćemo nadalje govoriti o skupu podataka-realnih brojeva $\mathcal{A} = \{a_1, \ldots, a_m\}$ među kojima može biti i jednakih. 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

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_j} d(x, a_i). \tag{9}$$

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

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} d(c_j, a_i), \tag{10}$$

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

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{F}(\Pi). \tag{11}$$

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

Obratno, za dani skup centara $c_1, \ldots, c_k \in \mathbb{R}$, uz primjenu *principa minimalnih udaljenosti* možemo definirati particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k, \tag{12}$$

pri čemu treba voditi računa o tome da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći optimizacijski problem

$$\min_{c_1, \ldots, c_k \in \mathbb{R}} F(c_1, \ldots, c_k), \qquad F(c_1, \ldots, c_k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} d(c_j, a_i), \tag{13}$$

gdje je $F\colon \mathbb{R}^k \to \mathbb{R}_+$. Općenito, ova funkcija nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma (Gan et al., 2007; Iyigun and Ben-Israel, 2010; Teboulle, 2007).

Optimizacijski problem (13) u literaturi se može naći pod nazivom *k-median problem* i ekvivalentan je optimizacijskom problemu (11). Naime, vrijedi

$$\begin{aligned} F(c_1,\dots,c_k) :=& \sum_{i=1}^{m} \min\{d(c_1,a_i),\dots,d(c_k,a_i)\} \\ \geq& \sum_{j=1}^{k} \sum_{a_i\in\pi_j} \min\{d(c_1,a_i),\dots,d(c_k,a_i)\} \\ =& \sum_{j=1}^{k} \sum_{a_i\in\pi_j} d(c_j,a_i) =: \mathcal{F}(\Pi), \end{aligned} \tag{14}$$

gdje je $\Pi = \{\pi_1,\dots,\pi_k\}$,

$$\pi_j = \pi_j(c_1,\dots,c_k) = \{a\in\mathcal{A}\colon d(c_j,a) \leq d(c_s,a),\ \forall s = 1,\dots,k\},$$

odnosno

$$\pi_j = \pi_j(c_1,\dots,c_k) = \{a_i\in\mathcal{A}\colon j = \operatorname*{argmin}_{s=1,\dots,k} d(c_s,a_i)\}.$$

Jednakost u (14) vrijedi ako je Π lokalno ili globalno optimalna particija.

3.1 Kriterij najmanjih kvadrata

Ako je $d\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$, $d(a,b) = (a-b)^2$ LS-kvazimetrička funkcija, centri $c_1,\dots,c_k$ klastera $\pi_1,\dots,\pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{u\in\mathbb{R}} \sum_{a_i\in\pi_j} (a_i-u)^2 = \frac{1}{|\pi_j|}\sum_{a_i\in\pi_j} a_i, \qquad j = 1,\dots,k, \tag{15}$$

a funkcija cilja (10) s

$$\mathcal{F}(c_1,\dots,c_k) = \sum_{j=1}^{k}\sum_{a_i\in\pi_j} (c_j-a_i)^2. \tag{16}$$

Primjer 6. *Zadan je skup $\mathcal{A} = \{0,3,6,9\}$. Treba pronaći sve njegove dvočlane particije koje zadovoljavaju Definiciju 1, odrediti pripadne centre i vrijednosti kriterijske funkcije cilja $\mathcal{F}$ u smislu najmanjih kvadrata.*

Broj svih dvočlanih particija ovog skupa je $2^{m-1}-1 = 7$, a kao što se vidi iz Tablice 1 LS-optimalna particija u ovom slučaju je $\{\{0,3\},\{6,9\}\}$ jer na njoj kriterijska funkcija cilja $\mathcal{F}$ zadana s (16) postiže globalni minimum.

3.1.1 Dualni problem

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

π_1	π_2	c_1	c_2	$\mathcal{F}(c_1,c_2)$		$\mathcal{G}(c_1,c_2)$	
{0}	{3,6,9}	0	6	0+18	=18	81/4+27/4	= 27
{3}	{0,6,9}	3	5	0+42	=42	9/4+3/4	= 3
{6}	{0,3,9}	6	4	0+42	=42	9/4+3/4	= 3
{9}	{0,3,6}	9	3	0+18	=18	81/4+27/4	= 27
{0,3}	{6,9}	3/2	15/2	9/2+9/2	=9	18+18	= 36
{0,6}	{3,9}	3	6	18+18	=36	9/2+9/2	= 9
{0,9}	{3,6}	9/2	9/2	81/2+9/2	=45	0+0	= 1

Tablica 1: Particije, centri i vrijednosti funkcije cilja $\mathcal{F}$ i $\mathcal{G}$

Lema 2. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup podataka, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka particija s klasterima $\pi_1, \ldots, \pi_k$ duljine $m_1, \ldots, m_k$. Neka je nadalje*

$$c = \frac{1}{m}\sum_{i=1}^{m} a_i, \qquad c_j = \frac{1}{m_j}\sum_{a_i \in \pi_j} a_i, \quad j = 1, \ldots, k, \tag{17}$$

gdje je $m_j = |\pi_j|$. Tada vrijedi

$$\sum_{i=1}^{m}(a_i - c)^2 = \mathcal{F}(c_1, \ldots, c_k) + \mathcal{G}(c_1, \ldots, c_k), \tag{18}$$

gdje je

$$\mathcal{F}(c_1, \ldots, c_k) = \sum_{j=1}^{k}\sum_{a_i \in \pi_j}(c_j - a_i)^2, \tag{19}$$

$$\mathcal{G}(c_1, \ldots, c_k) = \sum_{j=1}^{k} m_j(c_1, \ldots, c_k)(c_j - c)^2. \tag{20}$$

Dokaz. Primijetimo najprije da za svaki $x \in \mathbb{R}$ vrijedi

$$\sum_{a_i \in \pi_j}(a_i - x)^2 = \sum_{a_i \in \pi_j}(a_i - c_j)^2 + m_j(c_j - x)^2, \quad j = 1, \ldots, k. \tag{21}$$

Naime, kako je $\sum\limits_{a_i \in \pi_j}(a_i - c_j)(c_j - x) = (c_j - x)\sum\limits_{a_i \in \pi_j}(a_i - c_j) = 0$, vrijedi

$$\begin{aligned}\sum_{a_i \in \pi_j}(a_i - x)^2 &= \sum_{a_i \in \pi_j}((a_i - c_j) + (c_j - x))^2 \\ &= \sum_{a_i \in \pi_j}(a_i - c_j)^2 + m_j(c_j - x)^2.\end{aligned}$$

Ako u (21) umjesto x stavimo $c = \frac{1}{m}\sum\limits_{i=1}^{m} a_i$ i zbrojimo sve jednakosti, dobivamo (18). □

Iz Leme 2 neposredno slijedi tvrdnja sljedećeg teorema (Dhillon et al., 2004; Späth, 1983)

Teorem 2. *Uz oznake kao u Lemi 2 vrijedi:*

$$\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A},k)} \mathcal{F}(\Pi) = \operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A},k)} \mathcal{G}(\Pi),$$
$$\operatorname*{argmin}_{c_1,\ldots,c_k\in\mathbb{R}} \mathcal{F}(c_1,\ldots,c_k) = \operatorname*{argmax}_{c_1,\ldots,c_k\in\mathbb{R}} \mathcal{G}(c_1,\ldots,c_k).$$

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije $\mathcal{F}$ zadane s (16), odnosno (19), možemo maksimizirati funkciju

$$\mathcal{G}(c_1,\ldots,c_k) = \sum_{j=1}^{k} m_j(c_1,\ldots,c_k)(c_j-c)^2. \tag{22}$$

Primjer 7. *Skup $A=\{0,3,6,9\}$ iz Primjera 6 ima 7 različitih particija i za sve njih u Tablici 1 prikazana je vrijednost kriterijske funkcije cilja $\mathcal{G}$. Kao što se vidi iz Tablice 1 funkcija $\mathcal{G}$ prima maksimalnu vrijednost na optimalnoj particiji $\{\{0,3\},\{6,9\}\}$, što je u skladu s Teoremom 1.*

Primjedba 4. Lako se može provjeriti da je veza između centra c čitavog skupa $\mathcal{A}$ i centara c_j pojedinih klastera π_j zadanih s (17) dana s

$$c = \frac{m_1}{m}c_1 + \cdots + \frac{m_k}{m}c_k.$$

Specijalno, za dva disjunktna skupa realnih brojeva $A=\{x_1,\ldots,x_p\}$, $B=\{y_1,\ldots,y_q\}$ aritmetička sredina njihove unije jednaka je ponderiranom zbroju njihovih aritmetičkih sredina, tj. vrijedi

$$\overline{A\cup B} = \frac{p}{p+q}\overline{A} + \frac{q}{p+q}\overline{B}.$$

3.2 Kriterij najmanjih apsolutnih odstupanja

Ako je $d\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$, $d(a,b)=|a-b|$ l_1-metrička funkcijadefinirana, centri $c_1,\ldots,c_k$ klastera $\pi_1,\ldots,\pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{u\in\mathbb{R}} \sum_{a_i\in\pi_j} |a_i-u| = \operatorname{med}(\pi_j), \qquad j=1,\ldots,k, \tag{23}$$

a funkcija cilja (10), odnosno (13) s

$$\mathcal{F}(c_1,\ldots,c_k) = \sum_{j=1}^{k}\sum_{a_i\in\pi_j} |c_j-a_i|. \tag{24}$$

Ako pri tome iskoristimo (25), onda za izračunavanje funkcije cilja (24) nije potrebno poznavati centre klastera (23), što može značajno ubrzati računski proces.

Zadatak 3. *Neka je $\mathcal{A}=\{a_1,\ldots,a_m\}$ konačan niz realnih brojeva. Pokažite da vrijedi*

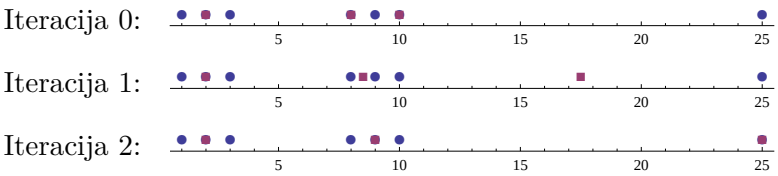
$$\sum_{i=1}^{m} |a_i - \operatorname{med}(A)| = \sum_{i=1}^{k} \left(a_{m-i+1} - a_i\right). \tag{25}$$

Primjedba 5. U slučaju izbora LAD-kriterija optimalnosti u Koraku 2 Algoritma 2 može se dogoditi da neki centroid ζ_j, može biti proizvoljan broj iz nekog intervala $[\alpha, \beta] \subset \mathbb{R}$. U tom slučaju treba uzeti $\zeta_j = \frac{\alpha+\beta}{2}$.

Primjer 8. *Zadan je skup* $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$. *Primjenom Algoritma 2 treba pronaći tročlanu particiju što bližu LS-optimalnoj.*

R.br.	z_1	z_2	z_3	$F(z_1, z_2, z_3)$	π_1	π_2	π_3	ζ_1	ζ_2	ζ_3	$\mathcal{F}(\{\pi_1, \pi_2, \pi_3\})$
1	2	8	10	228	$\{1, 2, 3\}$	$\{8, 9\}$	$\{10, 25\}$	2	8.5	17.5	115
2	2	8.5	17.5	61	$\{1, 2, 3\}$	$\{8, 9, 10\}$	$\{25\}$	2	9	25	4
3	2	9	25	4	$\{1, 2, 3\}$	$\{8, 9, 10\}$	$\{25\}$	2	9	25	4

Broj svih tročlanih particija ovog skupa je $\frac{1}{2}(3^{m-1} - 2^m + 1) = 301$. U prethodnoj tablici prikazan je tijek iterativnog postupka. Direktnom provjerom svih particija može se pokazati da je Algoritam 2 pronašao upravo optimalnu particiju. Iz ovog primjera vidljivo je da k-means algoritam u smislu LS-optimalnosti daje particiju, koja značajno ovisi o stršećem podatku, tako da upravo stršeći podatak čini zaseban klaster (vidi Sliku 1).



Slika 1: k-means algoritam za traženje lokalno optimalne particije skupa $\mathcal{A}$ iz Primjera 8

Primjer 9. *Zadan je skup* $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$ *iz Primjera 8. Primjenom Algoritma 2 treba pronaći dvočlanu particiju skupa* $\mathcal{A}$ *što bližu LAD-optimalnoj.*

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	2	15	29	$\{1, 2, 3, 8\}$	$\{9, 10, 25\}$	$[2, 3]$	10	24
2	$[2, 3]$	10	20	$\{1, 2, 3\}$	$\{8, 9, 10, 25\}$	2	$[9, 10]$	20
3	2	$[9, 10]$	20	$\{1, 2, 3\}$	$\{8, 9, 10, 25\}$	2	$[9, 10]$	20

Primjenom Algoritma 2 uz početne centre $c_1 = 2$ i $c_2 = 15$, dobivamo početnu particiju $\Pi = \{\pi_1, \pi_2\}$, $\pi_1 = \{1, 2, 3, 8\}$, $\pi_2 = \{9, 10, 25\}$. U tablici je prikazan tijek iterativnog postupka. Pri tome, centri u Koraku 2 Algoritma 2 birani su u skladu s Primjedbom 5. Direktnom provjerom može se pokazati da je algoritam pronašao upravo LAD-optimalnu particiju. U ovom slučaju stršeći podatak prirodno je pridružen drugom klasteru (vidi Sliku 2).

3.3 Grupiranje podataka s težinama

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a_1, \ldots, a_m\}$, pri čemu je svakom podatku a_i pridružena odgovarajuća težina $w_i > 0$. Kriterijska funkcija cilja (10) sada postaje

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} w_i d(c_j, a_i). \tag{26}$$

Specijalno, kod primjene kriterija LS-optimalnosti centar c_j klastera π_j određen je težinskom aritmetičkom sredinom podataka iz klastera π_j

$$c_j = \frac{1}{\kappa_j} \sum_{a_i \in \pi_j} w_i a_i, \qquad \kappa_j = \sum_{a_i \in \pi_j} w_i, \tag{27}$$

a kod primjene kriterija LAD-optimalnosti centar c_j klastera π_j određen je težinskim medijanom podataka koji pripadaju klasteru π_j (Sabo and Scitovski, 2008; Vazler et al., 2011)

$$c_j = \operatorname*{med}_{a_i \in \pi_j}(w_i, a_i). \tag{28}$$

Algoritam 2. (Standardni k-means algoritam)

Step 1: Učitati m, k, elemente skupa $\mathcal{A}$ i težine $w_1, \ldots, w_m > 0$;
Izabrati: $\min a_i \le c_1 < \cdots < c_k \le \max a_i$;

Step 2: Priduživanje (assignment step)

$$\pi_j = \{a_i \in \mathcal{A} : d(c_j, a_i) \le d(c_s, a_i), \quad s = 1, \ldots, k\}, \qquad j = 1, \ldots, k;$$

Step 3: Korekcija (update step)

$$\zeta_j = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_j} w_i d(x, a_i), \qquad j = 1, \ldots, k;$$

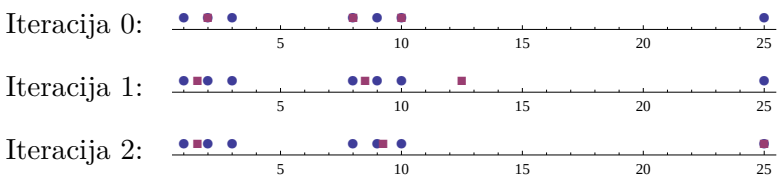
Step 2 i Step 3 se izmjenjuju tako dugo dok se ili centri ne poklope ili dok se particije ne poklope ili dok vrijednost funkcije cilja ne prestane opadati.

Primjer 10. *Podacima iz Primjera 9 dodajmo težine kako slijedi:* $w_i : 1, 1, .1, .5, .5, 1, .2$. *Grupirat ćemo skup* A *u* $k = 3$ *klastera primjenom k-means algoritma uz početne centre kao i ranije. Rezultati su vidljivi u niže navedenoj tablici.*

R.br.	z_1	z_2	z_3	F	π_1	π_2	π_3	ζ_1	ζ_2	ζ_3	$\mathcal{F}$
1	2	8	10	46.6	$\{1,2,3\}$	$\{8,9\}$	$\{10,25\}$	1,57143	8.5	12.5	38.4643
2	1.57143	8.5	12.5	34,4643	$\{1,2,3\}$	$\{8,9,10\}$	$\{25\}$	1.57143	9.5	25	2.08929
3	1.57143	9.5	25	2.08929	$\{1,2,3\}$	$\{8,9,10\}$	$\{25\}$	1.57143	9.5	25	2.08929

Prosječno težinsko kvadratno rasipanje po klasterima (varijanca): $\{0.340136, 0.6875, 0.\}$

Prosječno težinsko rasipanje po klasterima (standardna devijacija): $\{0.583212, 0.829156, 0.\}$



Slika 2: k-means algoritam za traženje lokalno optimalne particije skupa $\mathcal{A}$ iz Primjera 10

4 Grupiranje na osnovi dva obilježja

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup, koji treba na osnovi dva obilježja grupirati u k klastera koji zadovoljavaju uvjete iz Definicije 1. Primjerice, dane u godini možemo grupirati prema prosječnoj dnevnoj temperaturi izraženoj u °C i količini dnevnih padavina. Svaki element $a_i \in \mathcal{A}$ temeljem tih obilježja reprezentirat ćemo jednim vektorom $(x_i, y_i) \in \mathbb{R}^2$, kojeg ćemo označiti s $\mathbf{a}_i$. Nadalje, zbog jednostavnosti elemente skupa $\mathcal{A}$ identificirat ćemo s tim vektorima i govoriti o skupu podataka-vektora među kojima može biti i jednakih.

Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar $\mathbf{c}_j$ na sljedeći način

$$\mathbf{c}_j = c(\pi_j) := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^2} \sum_{\mathbf{a}_i \in \pi_j} d(\mathbf{x}, \mathbf{a}_i). \tag{29}$$

Kod problema grupiranja podataka u općem slučaju najčešće korištene kvazimetričke funkcije $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ su (Ben-Israel and Iyigun, 2007; Gan et al., 2007; Kogan, 2007)

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2 \qquad (\textit{least squares} \text{ udaljenost}) \tag{30}$$

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 \qquad (\text{Manhattan udaljenost}) \tag{31}$$

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})\mathbf{Q}(\mathbf{x} - \mathbf{y})^T, \quad \mathbf{Q} > 0,\ \mathbf{Q}^T = \mathbf{Q} \qquad (\text{Mahalanobisova udaljenost})$$

$$d(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{x}_i \left(\ln \frac{\mathbf{x}_i}{\mathbf{y}_i} - \mathbf{x}_i + \mathbf{y}_i \right),\ \mathbf{x}, \mathbf{y} \in \mathbb{R}_+^n\ (0 \ln 0 := 0) \quad (\text{Kullback–Leiblerova udaljenost})$$

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

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} d(\mathbf{c}_j, \mathbf{a}_i), \tag{32}$$

a d-optimalnu particiju $\Pi^\star$ tražimo rješavanjem optimizacijskog problema

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A},k)} \mathcal{F}(\Pi). \tag{33}$$

Pri tome, funkcija cilja (32) također može imati više lokalnih minimuma, koje također možemo tražiti primjenom Algoritma 2.

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

Obratno, za dani skup centara $c_1, \ldots, c_k \in \mathbb{R}^2$, uz primjenu *principa minimalnih udaljenosti* možemo definirati particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \le d(c_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k, \tag{34}$$

pri čemu treba voditi računa o tome da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru.
Voronoijev dijagram

Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći optimizacijski problem

$$\min_{c_1,\ldots,c_k\in\mathbb{R}} F(c_1,\ldots,c_k), \qquad F(c_1,\ldots,c_k)=\sum_{i=1}^{m}\min_{j=1,\ldots,k} d(c_j,a_i), \tag{35}$$

gdje je $F\colon \mathbb{R}^{2k}\to\mathbb{R}_+$. Općenito, ova funkcija nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma (Gan et al., 2007; Iyigun and Ben-Israel, 2010; Teboulle, 2007).

Optimizacijski problem (35) u literaturi se može naći pod nazivom *k-median problem* i ekvivalentan je optimizacijskom problemu (33). Naime, vrijedi

$$\begin{aligned} F(c_1,\ldots,c_k) &:= \sum_{i=1}^{m}\min\{d(c_1,a_i),\ldots,d(c_k,a_i)\} \\ &\geq \sum_{j=1}^{k}\sum_{a_i\in\pi_j}\min\{d(c_1,a_i),\ldots,d(c_k,a_i)\} \\ &= \sum_{j=1}^{k}\sum_{a_i\in\pi_j} d(c_j,a_i) =: \mathcal{F}(\Pi), \end{aligned}$$

gdje je $\Pi=\{\pi_1,\ldots,\pi_k\}$,

$$\pi_j=\pi_j(c_1,\ldots,c_k)=\{a\in\mathcal{A}\colon d(c_j,a)\leq d(c_s,a),\ \forall s=1,\ldots,k\},$$

odnosno

$$\pi_j=\pi_j(c_1,\ldots,c_k)=\{a_i\in\mathcal{A}\colon j=\operatorname*{argmin}_{s=1,\ldots,k} d(c_s,a_i)\},$$

pri čemu jednakost vrijedi ako je Π lokalno ili globalno optimalna particija.

4.1 Princip najmanjih kvadrata

Definicija 2. Neka je $\mathcal{A}=\{\mathbf{a}_i=(x_i,y_i)\in\mathbb{R}^2 : i=1,\ldots,m\}$ skup vektora iz $\mathbb{R}^2$. Kažemo da je particija $\Pi^\star=\{\pi_1^\star,\ldots,\pi_k^\star\}$ optimalna u smislu najmanjih kvadrata (skraćeno: LS-optimalna) ako je kvazimetrička funkcija $d:\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}_+$ definirana s (30), a $\Pi^\star$ rješenje optimizacijskog problema (32)–(33).

Primijetite da funkcija (30) nije metrika jer ne zadovoljava nejednakost trokuta. Centri $\mathbf{c}_1,\ldots,\mathbf{c}_k$ klastera $\pi_1,\ldots,\pi_k$ određeni su s

$$\mathbf{c}_j=\operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^2}\sum_{\mathbf{a}_i\in\pi_j}\|\mathbf{a}_i-\mathbf{u}\|_2^2=\frac{1}{|\pi_j|}\sum_{\mathbf{a}_i\in\pi_j}\mathbf{a}_i, \quad j=1,\ldots,k, \tag{36}$$

a funkcija cilja (32) s

$$\mathcal{F}(\Pi)=\sum_{j=1}^{k}\sum_{\mathbf{a}_i\in\pi_j}\|\mathbf{c}_j-\mathbf{a}_i\|_2^2 \tag{37}$$

Primjer 11. *Za skup* $\mathcal{A}=\{\mathbf{a}_1=(0,0),\mathbf{a}_2=(1,0),\mathbf{a}_3=(1,1),\mathbf{a}_4=(0,1)\}$ *odredit ćemo sve dvočlane particije, koje zadovoljavaju Definiciju 1, a nakon toga odgovarajuće centroide i vrijednosti funkcije cilja* (37) *u smislu LS-optimalnosti.*

π_1	π_2	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
$\{\mathbf{a}_1\}$	$\{\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4\}$	$\mathbf{a}_1$	$\left(\frac{2}{3}, \frac{2}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_2\}$	$\{\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_4\}$	$\mathbf{a}_2$	$\left(\frac{1}{3}, \frac{2}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_3\}$	$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_4\}$	$\mathbf{a}_3$	$\left(\frac{1}{3}, \frac{1}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_4\}$	$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$	$\mathbf{a}_4$	$\left(\frac{2}{3}, \frac{1}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_1, \mathbf{a}_2\}$	$\{\mathbf{a}_3, \mathbf{a}_4\}$	$\left(\frac{1}{2}, 0\right)$	$\left(\frac{1}{2}, 1\right)$	$\frac{1}{2} + \frac{1}{2}$	$= 1$	$\frac{1}{2} + \frac{1}{2}$	$= 1$
$\{\mathbf{a}_1, \mathbf{a}_4\}$	$\{\mathbf{a}_2, \mathbf{a}_3\}$	$\left(0, \frac{1}{2}\right)$	$\left(1, \frac{1}{2}\right)$	$\frac{1}{2} + \frac{1}{2}$	$= 1$	$\frac{1}{2} + \frac{1}{2}$	$= 1$
$\{\mathbf{a}_1, \mathbf{a}_3\}$	$\{\mathbf{a}_2, \mathbf{a}_4\}$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$1 + 1$	$= 2$	$0 + 0$	$= 0$

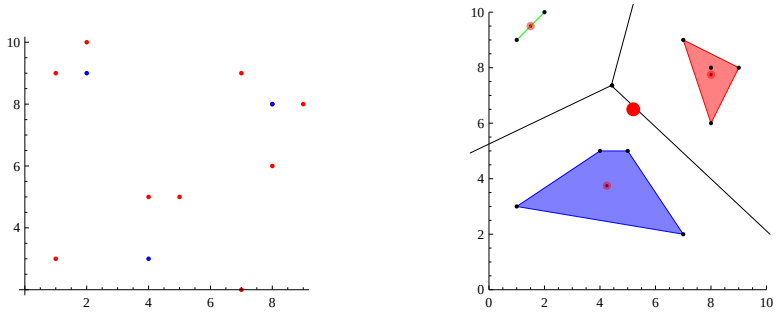
Broj svih dvočlanih particija ovog skupa je $2^{m-1} - 1 = 7$, a kao što se vidi iz tablice, dvije particije $\{\{\mathbf{a}_1, \mathbf{a}_2\}, \{\mathbf{a}_3, \mathbf{a}_4\}\}$ i $\{\{\mathbf{a}_1, \mathbf{a}_4\}, \{\mathbf{a}_2, \mathbf{a}_3\}\}$ su optimalne jer na njima kriterijska funkcija cilja (37) postiže globalni minimum (vidi sliku).

Primjer 12. *Zadan je skup* $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$

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

Uz početne centre $c_1 = (2, 9)$, $c_2 = (4, 3)$, $c_3 = (8, 8)$, primjenom LS-kvazimetričke funkcije i Algoritma 2 dobivamo lokalno optimalnu particiju $\Pi = \{\pi_1, \pi_2, \pi_3\}$ s centrima $c_1^\star = (\frac{3}{2}, \frac{19}{2})$, $c_2^\star = (\frac{17}{4}, \frac{15}{4})$, $c_3^\star = (8, \frac{31}{4})$ (vidi Sliku 3). Vrijednost funkcije cilja je $\mathcal{F} = \frac{133}{4} = 1 + \frac{51}{2} + \frac{27}{4}$.

(a) Početna particija (a) Optimalna particija



Slika 3: k-means algoritam za traženje lokalno optimalne particije skupa $\mathcal{A}$ iz Primjera 12

4.1.1 Dualni problem

Analogno, kao u jednodimenzionalnom slučaju može se pokazati da vrijedi

$$\sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{c}\|_2^2 = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_2^2 + \sum_{j=1}^{k} m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2, \tag{38}$$

gdje je $\mathbf{c} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{a}_i$ centar skupa $\mathcal{A}$, a $m_j = |\pi_j|$. Zato umjesto minimizacije funkcije $\mathcal{F}$ zadane s (37) optimalnu LS-particiju možemo tražiti maksimizacijom funkcije

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2. \tag{39}$$

Određenim prilagođavanjem (Dhillon et al., 2004) problem se svodi na poznate probleme i metode linearne algebre.

Primjer 13. *Skup $\mathcal{A}$ iz Primjera 11 ima 7 različitih particija i za sve njih u tablici je prikazana vrijednost kriterijske funkcije cilja $\mathcal{G}$. Kao što se vidi, funkcija $\mathcal{G}$ prima maksimalnu vrijednost na optimalnim particijama $\{\{\mathbf{a}_1, \mathbf{a}_2\}, \{\mathbf{a}_3, \mathbf{a}_4\}\}$ i $\{\{\mathbf{a}_1, \mathbf{a}_4\}, \{\mathbf{a}_2, \mathbf{a}_3\}\}$.*

4.2 Princip najmanjih apsolutnih odstupanja

Definicija 3. Neka je $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$ skup vektora iz $\mathbb{R}^2$. Kažemo da je particija $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ optimalna u smislu najmanjih apsolutnih odstupanja (skraćeno: LAD-optimalna) ako je kvazimetrička funkcija $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ definirana s (31), a $\Pi^\star$ rješenje optimizacijskog problema (32)–(33).

Centri $\mathbf{c}_1, \ldots, \mathbf{c}_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$\mathbf{c}_j = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^2} \sum_{\mathbf{a}_i\in\pi_j} \|\mathbf{a}_i - \mathbf{u}\|_1 = (\mathrm{med}(x), \mathrm{med}(y)) =: \mathrm{med}(\pi_j), \quad j = 1, \ldots, k, \tag{40}$$

gdje je $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, a funkcija cilja (32) zadana je s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i\in\pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_1 \tag{41}$$

Primjedba 6. Kao što smo u Odjeljku 3.3 razmatrali problem grupiranja jednodimenzionalnih težinskih podataka, slično bi mogli postupiti i u slučaju grupiranja težinskih dvodimenzionalnih i višedimenzionalnih podataka (Vazler et al., 2011).

4.3 Primjena Mahalanobis udaljenosti

Literatura

A. Ben-Israel, C. Iyigun, *Probabilistic D-clustering*, Journal of Classification **25**(2008), 5–26

D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

T. Calinski, J. Harabasz, *A dendrite method for cluster analysis*, Communications in Statistics, **3**(1974), 1–27

D. Davies, D. Bouldin, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **2**(1979), 224–227

I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA, 551–556, 2004

G. Divéki, I. Csanád, *Online facility location with facility movements*, CEJOR (2010), DOI 10.1007/s10100-010-0153-8

E. Domínguez, J. Muñoz, *Applying bio-inspired techniques to the p-median problem*, IWANN 2005; Computational Intelligence Bioinspired Syst., 8th Int. Workshop Artificial Neural Networks, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2005, pp. 67 – 74

Z. Drezner, *Facility Location: A Survey of Aplications and Methods*, Springer-Verlag, Berlin, 2004.

J. Dunn, *Well separated clusters and optimal fuzzy partitions*, Journal of Cybernetics, **4**(1974), 95–104

B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.

D. E. Finkel, C. T. Kelley, *Additive scaling and the* `DIRECT` *algorithm*, J. Glob. Optim. **36**(2006), 597-–608

C. A. Floudas, C. E. Gounaris, *A review of recent advances in global optimization*, J. Glob. Optim. **45**(2009), 3-–38

G. Gan, C Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

E. R. Hansen, G. W. Walster, *Global Optimization Using Interval Analysis*. Marcel Dekker, New York, Second Edition, Revised and Expanded, 2004.

M. Hudec, M. Vujošević, *A fuzzy system for municipalities classification*, CEJOR **18**(2010), 171–180

C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010), 207–214

C. Iyigun, *Probabilistic Distance Clustering*, Dissertation, Graduate School – New Brunswick, Rutgers, 2007

D. R. Jones, C. D. Perttunen, B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, JOTA **79**(1993), 157–181

J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

J. Kogan, C. Nicholas, M. Wiacek, *Hybrid Clustering of large high dimensional data*, In M. Castellanos and M. W. Berry (Eds.), Proceedings of the Workshop on Text Mining, SIAM, 2007.

J. Kogan, M. Teboulle, *Scaling clustering algorithms with Bregman distances*. In: M. W. Berry and M. Castellanos (Eds.), Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining, 2006.

J. Kogan, C. Nicholas, M. Wiacek, *Hybrid clustering with divergences*. In: M. W. Berry and M. Castellanos (Eds.), Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition, Springer, 2007.

C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010) 207–214

D. Littau, D. L. Boley, *Clustering very large data sets with PDDP*. In J. Kogan, C. Nicholas, M. Teboulle (eds), *Grouping Multidimensional Data: Recent Advances in Clustering*, 99–126, Springer-Verlag, New York, 2006.

S. Pan,J. S. Chen, *Two unconstrained optimization approaches for the Euclidean k-centrum location problem*, Applied Mathematics and Computation **189**(2007) 1368–1383

С. А. Пиявский, Один алгоритм отыскания абсолютного екстрмума функции, Ж. вычисл. матем. и матем. физ. **12**(1972), 888-896.

J. Reese, *Solution methods for the p-median problem: an annotated bibliography*, Published online in Wiley InterScience, Wiley, 2006.

A. M. Rodrígues-Chia, I. Espejo, Z. Drezner, *On solving the planar k-centrum problem with Euclidean distances*, EJOR, to appear

K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198

K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka. Klasteri*, Osječki matematički list **10**(2010), 149–178

K. Sabo, R. Scitovski, I. Vazler, M. Zekić-Sušac, *Mathematical models of natural gas consumption*, Energy Conversion and Management **52**(2011), 1721-1727

K. Sabo, R. Scitovski, I. Vazler, *One-dimensional center-based l_1-clustering method*, Optimization Letters (accepted)

A. Schöbel, D. Scholz, *The big cube small cube solution method for multidimensional facility location problems*, Computers & Operations Research **37**(2010), 115–122

A. Schöbel, *Locating Lines and Hyperplanes: Theory and Algorithms*, Springer Verlag, Berlin, 1999.

B. Shubert, *A sequential method seeking the global maximum of a function*, SIAM Journal on Numerical Analysis, **9**(1972), 379–388

H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

Z. Su, J. Kogan, C. Nicholas, *Constrained clustering with k-means type algorithms*, In M.W. Berry, J. Kogan (eds), *Text Mining Applications and Theory*, 81–103, Willey, Chichester, 2010.

M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102

Vazler, I., Sabo, K., Scitovski, R.: Weighted median of the data in solving least absolute deviations problems. Comm. Statist. Theory Methods (to appear)

D. Veljan, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.