M059	Izborni 4. godina	**Grupiranje podataka i primjene**	P+V+S 2+1+1	ECTS 5

Cilj predmeta. Studente upoznati s osnovnim činjenicama i rezultatima iz područja grupiranja podataka, te mogućim primjenama.

Potrebna predznanja. Linearna algebra II, Funkcije više varijabli.

Sadržaj predmeta.

1. Uvod i motivacija. Definiranje problema i osnovna svojstva. Razni primjeri iz primjena.
2. Reprezentant konačnog skupa iz R u smislu najmanjih kvadrata (LS) i u smislu najmanjih apsolutnih odstupanja (LAD). Reprezentan konačnog skupa podataka iz R^2. Kvazimetričke funkcije u R^2. Centroid, medijan i geometrijski medijan u ravnini. Reprezentan konačnog skupa podataka iz R^n: centroid, medijan, geometrijski medijan. Primjena Mahalanobis kvazimetričke funkcije. Reprezentan podataka na jediničnoj kružnici.
3. Grupiranje podataka. Motivacija na dvočlanoj particiji skupa iz R i R^2. K-means algoritam.
4. Grupiranje na osnovi jednog obilježja. LS-kriterij. Dualni problem. Transformacija podataka. LAD-kriterij. Grupiranje podataka s težinama.
5. Grupiranje na osnovi dva i više obilježja. LS-kriterij. Dualni problem. LAD-kriterij. Grupiranje podataka s težinama. Primjena Mahalanobis kvazimetričke funkcije. Svojstva: monotonost, stabilnost. Primjereni broj grupa u particiji: Indeksi.
6. Drugi geometrijski objekti kao reprezentanti podataka uz primjenu raznih kvazimetričkih funkcija: pravac, dužina, kružnica, krug. Parametarski zadana krivulja u ravnini i prostoru kao reprezentant. Grupiranje podataka u klastere čiji su centri spomenuti geometrijski objekti.
7. Metode za grupiranje podataka. K-means algoritam. Izmještanje elemenata. Korigirani algoritam k-sredina. Metoda zamjene. Metoda aglomeracije.

Očekivani ishodi učenja.

Očekuje se da nakon položenog kolegija studenti:

- budu u stanju samostalno prepoznati probleme s odgovarajućim bazama podataka gdje mogu primijeniti dobivena znanja;
- razumiju pojam reprezentanta podataka u smislu LS, LAD i Mahalanobis kvazimetričke funkcije;
- razumiju složenost optimizacijskog problema grupiranja podataka i nauče primjenu osnovnog k-means algoritma, kao i nekih drugih metoda;
- razumiju ideju raznih geometrijskih objekata kao reprezentanata podataka;
- kroz nekoliko tipičnih situacija, ovladaju metodologijom primjene grupiranja podataka.

Izvođenje nastave i vrednovanje znanja.

Predavanja i vježbe su ilustrirani gotovim programima. Vježbe su djelomično auditorne, a djelomično laboratorijske uz korištenje računala. Predavanja, vježbe i seminari su obavezni. Ispit se sastoji od pismenog i usmenog dijela, a polaže se nakon odslušanih predavanja. Prihvatljivi rezultati postignuti na kolokvijima, koje studenti pišu tijekom semestra, zamjenjuju pismeni dio ispita. Studenti mogu utjecati na ocjenu tako da tijekom semestra pišu domaće zadaće ili izrade seminarski rad. Domaće zadaće sadrže proširenje gradiva, a očekuje se samostalan i kreativan rad. Seminarski radovi shvaćaju se kao proširenje domaćih zadaća.

Može li se predmet izvoditi na engleskom jeziku: Da

Osnovna literatura:

1. K. Sabo, R. Scitovski, I. Vazler, Grupiranje podataka- klasteri, OML 10(2010), 149--178
2. J. Kogan, Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, 2007.

Dopunska literatura:

1. G. Gan, C. Ma, J. Wu, *Data clustering : theory, algorithms, and applications*, SIAM, Philadelphia, 2007.
2. B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.
3. M. Teboulle, A unified continuous optimization framework for center-based clustering methods, Journal of Machine Learning Research 8(2007), 65-10
4. C. Iyigun, Probabilistic Distance Clustering, Dissertation, Graduate School - New Brunswick, Rutgers, 2007
5. D. Bahdir, C. Iyigun, A Classification algorithm using Mahalanobis distance clustering of dana with applications on biomedical dana set, Dissertation, Graduate School of Natural and Applied Sciences, MEDU, 2011
6. H. Zha, X. He, C. Ding, H. Simon, M.Gu, *Spectral Relaxation for k-means Clustering*, Advances in Neural Information Systems, 2002.
7. H. Späth, *Cluster–Formation und- Analyse*, R. Oldenburg Verlag, München, 1983.