M062	Obavezni MR, IPM 2. semestar	**Grupiranje podataka: pristupi, metode i primjene**	P+V+S 2+1+1	ECTS 5

Cilj predmeta. Studente upoznati s osnovnim pristupima, metodama grupiranja podataka, te mogućim primjenama.

Potrebna predznanja. Preddiplomski studij

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^n: centroid, medijan, geometrijski medijan. Primjena Mahalanobis kvazimetričke funkcije. Reprezentan podataka na jediničnoj kružnici.
3. Grupiranje na osnovi jednog, dva i više obilježja. LS-kriterij. Dualni problem. Transformacija podataka. LAD-kriterij. Grupiranje podataka s težinama. Primjena Mahalanobis kvazimetričke funkcije. Svojstva: monotonost, stabilnost.
4. Metode za grupiranje podataka. K-means algoritam. EM(Expectation Maximization) algoritam. K-medoid metoda. Metoda aglomeracije.
5. Primjereni broj grupa u particiji: Indeksi.
6. Matrični pristup i primjena Ky Fan-ovog teorema.
7. Primjena teorije grafova za grupiranje podataka.
8. Vjerojatnosni i statistički aspekti grupiranja podataka.

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 razne metode za rješavanje prisutne u novijoj literaturi;
- razumiju primjenu raznih indeksa za procjenu prihvatljivog broja klastera u particiji;
- 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. A.Ben-Israel, C. Iyigun, Probabilistic distance clustering, J. Classification 25 (2008) 5–26.
2. I.Dhillon, A Unified View of Kernel k-means, Spectral Clustering and GraphCuts, UTCS Technical Report #TR-04-25, 2005.
3. H.Zha, X.He, C.Ding, H.Simon, M.Gu, *Spectral Relaxation for k-means Clustering*, Advances in Neural Information Systems, 2002.
4. A.K.Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters 31 (2010) 651–666
5. G. Gan, CMa, J.Wu, *Data clustering : theory, algorithms, and applications*, SIAM, Philadelphia, 2007.
6. J.Han, M.Kamber, A.K.H.Tung, Spatial clustering methods in dana mining: A survey B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.
7. M.Teboulle, A unified continuous optimization framework for center-based clustering methods, Journal of Machine Learning Research 8(2007), 65-10
8. C.Iyigun, Probabilistic Distance Clustering, Dissertation, Graduate School - New Brunswick, Rutgers, 2007
9. 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