M059	FIN-elective - Year 1	**Data Clustering and Applications**	L+P+S 2+1+1	ECTS 4

Course objectives. Students will be introduced to fundamental facts and results in the field of data clustering, as well as possible applications.

Course prerequisites. Undergraduate study programme in mathematics..

Syllabus.

1. Introduction and motivation. Problem statement and basic properties. Various examples from the application.
2. Representative of the finite set from R in least squares (LS) sense and in least absolute deviations (LAD) sense. Representative of the finite set from R^2. Distance-like function in R^2. Centroid, median and geometrics median in plane. Representative of the finite set from R^n: centroid, median, geometrics median. Applications of Mahalanobis distance-like function. Representative of the unit circle.
3. Data clustering. Motivation: partition of two clusters in R and R^2. K-means algorithm.
4. Clustering of the data that have only one feature. LS-criterion. Dual problem. Transformation of the data. LAD- criterion. Data clustering with weights.
5. Clustering of the data that have two or several features. LS-criterion. Dual problem. LAD-criterion. Data clustering with weights. Applications of Mahalanobis distance-like function. Properties: monotonicity, stability. Appropriate number of clusters in a partition: Indexes.
6. Other geometric objects as the representative of the data by using various distance-like functions: line, line segment, circle, disc. A curve given parametrically in plane and space as the representative. Data clustering by the aforementioned geometrics objects.
7. Data clustering methods. K-means algorithm. Displacement of elements. Corrected k-means algorithm. Replacement method. Hierarchical clustering. Matrix approach and the application of the Ky Fan theorem.

Expected learning outcomes.

After completing this course, students are expected to:

- be able to independently recognize problems with corresponding databases where they can apply the acquired knowledge;
- understand the term representative of the data by using LS, LAD and Mahalanobis distance-like functions;
- understand the complexity of data clustering and learn how to apply the basic k-means algorithm, as well as some other methods;
- understand the idea of various geometrics objects as the representative of the data;
- learn how to manage the methodology referring to the application of data clustering.

Teaching methods and student assessment.

Lectures and exercises are illustrated by ready-made software packages. Exercises are partially auditory and partially laboratory, with the use of computers. Lectures, exercises and seminars are obligatory. The exam consists of a written and an oral part, and it is taken after completion of lectures. Acceptable results achieved in mid-term exams throughout the semester replace the written part of the exam. Students may influence their final grade by doing homework or writing a seminar paper during the semester. Homework expands course contents, and students are expected to be independent and creative. Seminar papers are understood as an extension of homework.

Can the course be taught in English: Yes.

Basic literature:

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.

Recommended literature:

1. H.Zha, X.He, C.Ding, H.Simon, M.Gu, *Spectral Relaxation for k-means Clustering*, Advances in Neural Information Systems, 2002.
2. H. Späth, *Cluster–Formation und- Analyse*, R. Oldenburg Verlag, München, 1983.
3. G. Gan, CMa, J.Wu, *Data clustering : theory, algorithms, and applications*, SIAM, Philadelphia, 2007.
4. B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.
5. M.Teboulle, A unified continuous optimization framework for center-based clustering methods, Journal of Machine Learning Research 8(2007), 65-102.
6. C.Iyigun, Probabilistic Distance Clustering, Dissertation, Graduate School - New Brunswick, Rutgers, 2007