Data mining models and algorithms
Instructor: András Lukács

Term: Spring

Weeks: 1-7

Contact hours: 5

Credits: 12

Aim and scope:

Understanding the basic approach and methodology of data mining; knowledge of models, algorithms and tools to solve main tasks of data mining; planning and implementing simple data mining projects. Homeworks are an important part of this practical course. Practices cover the weekly topics in the form of Python notebooks. Main libraries: Scikit Learn, Pandas, Matplotlib, Bokeh.

Syllabus:
Knowledge discovery process, data preprocessing and exploration.
Human-computer interaction in data mining, data visualization, visual analytics.
Relational, multidimensional and stream data models. Data warehouses, Online Analytical Processing.
Feature engineering. Dimensionality reduction, singular value decomposition, Principal component analysis. Feature selection, measuring importance, wrapper and model-based methods.
Classification and regression. Performance evaluation. Class-imbalanced data. Receiver operating characteristic-curve, overfitting, sampling techniques. Decision trees, rule-based methods, nearest-neighbour classifier. Naive Bayes classifier and Bayesian networks. Perceptron, multilayer neural networks, backpropagation. Unsupervised learning. Cluster analysis, k-means and its generalizations, hierarchical and density-based methods. Model based clustering, maximum likelihood estimation, EM-method. One-class classification, outlier detection. Approximating densities. Isolation trees.
Association rule mining, frequent itemset and pattern mining. Apriori and Frequent Pattern-Growth algorithms. Constraint-based frequent pattern mining.

Grading: term mark (incorporating the solution of homeworks)

Literature:

P. N. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining (2nd edition), Pearson, 2018

J. Han, M. Kamber, J. Pei: Data Mining: Concepts and Techniques (3th edition), Elsevier, 2012

J. Friedman, T. Hastie, R. Tibshirani: The Elements of Statistical Learning (2nd edition), Springer, 2009

A. Rajaraman, J. Leskovec, J. D. Ullman: Mining of Massive Datasets, Cambridge Univ. Press, 2014