Data Science 2: Applications

View PDF
Instructor(s):

Roland Molontay

Short Description of the Course:
"What data scientists do is make discoveries while swimming in data", as described by the Harvard Business Review. In the second part of the course, we learn advanced techniques including kernel methods, recommender systems, network centrality, in addition to getting introduced to Big Data tools such as Hadoop. During the course, we will have guest lectures by data scientists from companies in the Budapest area. Students will have the option to define their data mining projects and work in teams during the semester.

Aim of the Course:
The aim of the course is to discuss advanced techniques of data mining with useful knowledge of related disciplines supporting real-world, especially bioinformatics data mining projects. By the end of the course, students will be able to analyze biological (genomic, microarray, pathway, protein, chemical) data sets using complex data mining methods.

Prerequisites:
The course requires basic knowledge in data mining. (See also the course Data Mining: Models and Algorithms) Background in probability theory, linear algebra and programming is important.

Detailed Program and Class Schedule:

- Advanced classification methods: Bagging, boosting, AdaBoost.
- More models and algorithms for classification: neural networks, linear separation methods, support vector machine (SVM).
- Random forest.
- Recommender systems. Collaborative filtering. Implicit and explicit recommendation.
- Dimensionality reduction by spectral methods, singular value decomposition, low-rank approximation.
- Search engines, web information retrieval, PageRank and network mining.
- Distributed data processing systems, data processing with Hadoop.
- Text mining, natural language processing.
- Selected topics connected to student projects (e.g. Mining biological, scientific, social media data)
- Final test.

Method of Instruction:
Handouts, presentations, IPython Notebooks, relevant research papers, web page, course mailing list and Wiki. Weekly regular office hour for consultations.

Textbooks:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Addison-Wesley, 2006.

Jure Leskovec, Anand Rajaraman, Jeff Ullman: Mining of Massive Datasets

http://www.mmds.org/

Instructors' bio:

Roland Molontay (born 1991) obtained his masters degree in applied mathematics from Budapest University of Technology and Economics (BME) in 2015. Between 2015 and 2018 he was a PhD student at BME specialized in network theory. In 2016 he was a visiting PhD student at Brown University. Currently he holds a research position at MTA-BME Stochastics Research Group and he also teaches mathematics and data science at BME for undergraduate students. He has been participating in many successful data intensive R&D projects with renowned companies (such as NOKIA or Bell Labs) throughout the years. At BME he also leads a small research group conducting data-driven educational research.