Data Science 1: Models and Algorithms

View PDF
Instructor(s):

Roland Molontay

Short Description of the Course:
Data Scientist is called "the sexiest job of the Century" by Harvard Business Review. In the first part of the course, we learn the basics of understanding data and predicting its unknown properties.

We give a general introduction to data analysis, modeling, and algorithms of data mining.The course provides a good base for its follow-up course, Data Mining Applications.Lectures are supplemented by computer exercises and student projects in small teams.

Aim of the Course:
The aim of the course is to provide a basic but comprehensive introduction to data mining. By the end of the course students will be able to choose the right algorithms for data science problems to build, implement and evaluate data mining models.

Prerequisites:
The course requires basic knowledge in calculus, probability theory, and linear algebra. Knowledge of graphs and basic algorithms is an advantage. Basic programming skills are also required.

Detailed Program and Class Schedule:

- Motivations for data mining. Examples of application domains.
- Analyzing data: preparation and exploration.
- Models and algorithms for classification.
- Introduction to the IPython Notebook and python based data mining software packages. Classification with scikit-learn.
- Basics of classification. Concepts of training and prediction. Measuring quality and comparison of classification models.
- Type of variables, measuring similarity and distances. The k-nearest neighbor classifier.
- Decision trees, naive Bayes. The concept of model over and underfitting. Midterm test.
- Basics of cluster analysis. Partitioning clustering algorithms, k-means, k-medoids.
- Hierarchical clustering algorithms.
- Introduction to frequent itemset mining. Applications for finding association rules. Level-wise algorithms, APRIORI.
- 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.