Data Science 2: Applications

Instructor(s):
Roland Molontay

Short description:
Data scientist is called "the sexiest job of the Century" by Harvard Business Review. In the second part of the course, we learn advanced supervised learning techniques including neural networks and ensemble methods together with unsupervised learning techniques (especially clustering). Students will have the option to define their data science projects and work in teams during the semester. Lectures are supplemented by problem-solving sessions, Python programming exercises and student projects in small teams.

Aim of the Course:
The aim of the course is to provide a comprehensive introduction to data science with a focus on machine learning. By the end of the course, students will be able to choose the right algorithms for data science problems to build, implement and evaluate machine learning models. Students will also be able to analyze real-world data sets using complex data science methods. The aim of the course is to provide the knowledge and skills needed to excel in a job interview for a junior data scientist position.

Prerequisites:
Basics of linear algebra (basic matrix operation, solving systems of linear equations, equations of lines and planes)
Basics of multivariate calculus (partial derivatives, gradient, finding maxima and minima of uni- and multivariate functions)
Basics of probability (Conditional probability, Bayes theorem, correlation, covariance, binomial distribution, normal distribution)
Basics of Python programming

Syllabus:

1. **Linear regression**: Parametric and nonparametric regression, kNN and Decision Tree for regression task, MSE, decomposition of MSE and variance, Bias–Variance tradeoff, the optimal solution of regression, linear regression, gradient descent, stochastic gradient descent, learning rate, regularization, polynomial regression, interpreting linear regression models.
2. **Logistic regression and SVM**: Classification by regression, sigmoid function, logistic regression, linear separability, non-linear decision boundary, logit model, maximal margin, support vectors and SVM
3. **Neural networks**: Biological motivation, activation function, perceptron and its relation to other algorithms, representing Boolean functions with neural networks, deep-learning, forward propagation, backpropagation.
4. **Ensemble learning**: Ensemble methods, bagging, metamodels, boosting and AdaBoost, gradient boosting, Random Forest, semi-supervised learning, classification of imbalanced data, SMOTE.
5. **Cluster analysis**: Concept, types, clustering algorithms, k-means algorithm, hierarchical clustering, distance of clusters, Simple-linkage and Complete-linkage clustering, DBSCAN algorithm, core border and noise points, validation of clustering (distance matrix, SSE, silhouette)
6. **Recommendation systems**: content-based recommender, collaborative filtering, user-based and k-nearest neighbors recommender, latent factor recommender system, matrix factorization.
Technologies:
**Python:** pandas, Scikit-learn, NumPy, SciPy, matplotlib, IPython
Topics: classification and regression tasks, gradient descent, ensemble methods

**Method of instructions**
Lectures (presentations) Problem solving sessions (handouts) Programming sessions (IPython notebooks)

**Requirements:**
Final exam (50%)
A sample final exam is available [here](https://example.com).
Team project (50%)

**Recommended literature:**

**Instructors' bio:**

**Roland Molontay** (born 1991) obtained his PhD degree in network and data science from Budapest University of Technology and Economics (BME). He was a visiting PhD student at Brown University in 2016. Currently he holds a research position at MTA-BME Stochastics Research Group and he also teaches mathematics and data science at BME for undergraduate and graduate students. He has been participating in many successful data intensive R&D projects with renowned companies (such as NOKIA-Bell Labs) throughout the years. He has been awarded the Gyula Farkas Memorial Prize in 2020 for his outstanding work in applied mathematics. He is the founder and leader of the Human and Social Data Science Lab at BME.