Course description

This is a basic course on Machine Learning that introduces most popular methods and approaches to knowledge extraction from data. Lectures will cover such topics as linear models, case-based methods, decision trees and ensembles. We will also have a look at the basics of neural networks. Students will learn how to preprocess data (including categorical and text cases), choose and analyse quality metrics for a particular task, validate and evaluate models. All topics will be covered with practical homework assignments on Python.

Course requirements, grading, and attendance policies

Class attendance and participation are encouraged, but not required. The course grade will be based on homework assignments (60% of the grade) and the final exam (40%).

Course Contents

1. Data types, problem statements and general approach in machine learning.
2. K nearest neighbour model. Metrics. Model selection, basic loss functions, cross-validation.
3. Linear regression model. Analytical solution and numerical approach. Gradient descent for model learning. Overfitting and regularization. Loss functions and dealing with outliers.
4. Linear classification model. Learning classification models with upper bounds on binary loss function. Quality metrics for classification problems. Multi-class and multi-label problems and their reduction to binary classification.
5. Decision trees. Greedy learning approach. Impurity functions. Connection between linear models and decision trees.
6. Bagging. Bias-variance decomposition. Analysis of bias and variance for bagging. Random forest.
7. Gradient boosting on decision trees. Model correction by fitting residuals. Modern implementations of gradient boosting.
8. Feature selection. Black-box methods and model-based methods. Model interpretation.
9. Data visualization. t-SNE method.
10. Introduction to neural networks. Backpropagation. Fully-connected and convolutional neural networks. Embeddings.

Sample tasks for course evaluation

Write down a linear classification model. How many parameters are there? How to evaluate model quality for an imbalanced dataset? How to fit this model on a large dataset? How to prepare a dataset to speed up gradient descent? How to apply cross-validation to prevent overfitting? How to use regularization here, how to choose regularization strength?

Course materials

There is no required textbook, but selected chapters from Hastie-Tibshirani-Friedman “The Elements of Statistical Learning” will be useful.

Academic integrity policy

Cheating, plagiarism, and any other violations of academic ethics at NES are not tolerated.