Tárgytematika / Course Description

Machine learning

GKNM_MSTA040

Tárgyfelelős neve / Teacher's name: dr. Takács Gábor

Félév / Semester: 2022/23/2

Beszámolási forma / Assesment: Vizsga

Tárgy heti óraszáma / Teaching hours(week): 2/2/0

Tárgy féléves óraszáma / Teaching hours(sem.): 0/0/0

OKTATÁS CÉLJA / AIM OF THE COURSE

The goal of the course is to introduce the basic concepts, models and algorithms of machine learning. In addition, the course teaches how to create real-life machine learning solutions using scikit-learn and the Python scientific stack.

TANTÁRGY TARTALMA / DESCRIPTION

- Fundamental problems of machine learning: classification, regression, clustering. Some interesting real-life applications.
- Univariate linear regression. Testing on the Baseball Players data set. Multivariate linear regression. Bias term. Testing on the Boston Housing data set. The root mean squared error and the mean absolute error metric.
- Introduction to scikit-learn. Linear regression in scikit-learn. Train-test split and cross-validation.
- Univariate logistic regression. The cross-entropy metric. Training with Newton-method. Evaluating binary classifiers. Confusion matrix. Sensitivity, specificity, precision, recall, F1-score, accuracy and balanced accuracy. ROC-curve.
- Multivariate logistic regression. Logistic regression in scikit-learn. Testing on the Wisconsin Breast Cancer data set. Multiclass logistic regression.
- The phenomenon of overfitting. L1 and L2 regularization. Ridge regression. Testing regularized logistic regression on the SMS Spam data set. Model selection. Greedy search and grid search.
- The decision stump and the decision tree model. Training with "brute force" method. Testing on the Boston Housing data set. Visualizing decision trees.
- Ensemble methods. Random forest. Extremely randomized trees. Testing on the Bostion Housing data set.
- Gradient boosted decision trees. Testing on the Boston Housing data set. Explanation generation. State of the art libraries for gradient boosting.
- Artificial neural networks. The multilayer perceptron model. Training with stochastic gradient method. Testing on the Phishing Websites data set. Multiclass case, testing on the MNIST data set.
- Feature engineering. Input normalization. Binning. Handling missing data. Demonstrating end-to-end machine learning on the Bike Sharing data set.
- Visualization methods. Principal component analysis (PCA). T-distributed stochastic neighbor embedding (t-SNE).
- Clustering methods. K-means. DBSCAN.

SZÁMONKÉRÉSI ÉS ÉRTÉKELÉSI RENDSZERE / ASSESMENT'S METHOD

The course ends with an exam, where the students solve simple machine learning problems in Python. The available time is 90 minutes. Following the computer-based part of the exam, students must validate the originality of their work through an oral exam. Scoring: 21-24: 5 (excellent), 18-20: 4 (good), 15-17: 3 (satisfactory), 12-14: 2 (pass).

KÖTELEZŐ IRODALOM / OBLIGATORY MATERIAL

- Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer, ISBN 978-0387848570.
- I. Witten, E. Frank, M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann, ISBN 978-0123748560.