Machine Learning and Data Mining

April, 2021

Outline, Dates & Deadlines

Dates

- 26. April 2021., 28. April 2021., 29. April 2021.;
- 5. May 2021., 11. May 2021., 18. May 2021., 24. May 2021., 26. May 2021., 27. May 2021., 28 May 2021.
- Deadlines: Task 1 Seminar presentation (11. May 2021.), Task 2 Problem Sheet 1 (18. May 2021.).

Outline

- Introduction and motivation (Machine Learning and Data Mining)
- Introduction to both fundamental programming concepts and the Python programming language (the programming language R)
- The Principal Components Analysis (PCA)
- The Limit Order Book (LOB)
- The Long short-term memory (LSTM) neural network

What is data mining?

Data mining

- the process of analyzing the data sets using machine learning and statistics to find insights and to detect pattern in the observed data sets
- explosive data growth, we are able to store much more data than before $\implies$ automation of the massive data sets analysis
- “We are drowning in information but starved for knowledge.” - John Naisbitt
- The main job of data mining is to extract and pick out the potentially useful, understandable and hidden information, previously unknown information/knowledge from the data set.
- Alternative names: data analysis, business intelligent, knowledge extraction, information harvesting, etc.
- the process of analyzing data sets from different prospective and summarizing the relevant knowledge

The steps in data mining

The steps in data mining

- Data cleaning (noise, outliers, missing values, duplicate data)
- Data integration
- Data selection
- Data preprocessing/ Data transformation
- Data mining: select the mining approach, choose the mining algorithm
- Pattern evaluation
- Knowledge presentation and visualization
- Use of discovered information

Basic Data Mining Tasks

- Classification: Assign each data element into the one of the predefined classes (e.g. Spam e-mail detection, etc.)
- Regression: map item into the real valued prediction variable (e.g. predict a value of savings in future)
- Clustering: similar to classification but the groups are not predefined (e.g. group people into communities from a given social network)

Python

Installation

- Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, etc.)
- https://www.anaconda.com/
- The Jupyter Notebook is an open-source web application that lets you easily write and iterate Python code for data analysis.
- https://jupyter.org/
- PyCharm
- https://www.jetbrains.com/pycharm/