DA515

PRACTICAL CASE STUDIES in DATA ANALYTICS

Course objectives:

This course aims at discussing the key principles of knowledge discovery process through various case studies arising from different application areas. The students are expected to learn the main steps to traverse when they face new data analytics problems. With each case study, the tools for cleaning, processing and altering the data shall be visited. A particular attention shall be given to data inspection, feature reduction and model selection. Each case study will be completed by a thorough discussion and interpretation of the results.

Course outline:

- Python and Important Packages
- Review of Machine Learning Methods
- Knowledge Discovery Process
- Web Crawling – Case Study: Twitter Sentiment Analysis
- Unsupervised Learning – Case Study: Customer Segmentation
- Supervised Learning – Case Study: Credit Scoring
- Dynamic Plotting – Case Study: Piketty's Capital
- Fraud Detection – Case Study: Enron Case
- Character Recognition from Images – Case Study: Google Street View

Software requirements:

We shall use Anaconda (Python 2.7). This software installs all the necessary packages on your computer. Please install it before coming to the class. For the term project and in-class project you are required to use IPython Notebook and provide adequate documentation of your work.

Detailed outline:

Preliminaries:

- Revisiting Python: tools, packages, programming environment, IPython Notebook
- Machine learning review: supervised and unsupervised learning
- Warm-up case study: Text mining on citation data of Turkish academics

Case Study 1 – Crawling the Web: Sentiment Analysis with Twitter Data

- Using an application programming interface (API)
- Inspecting the JSON format
- Simple sentiment analysis with word counts
- Plotting the word cloud

Case Study 2 - Unsupervised Learning: Customer Segmentation in Retail Industry

- Main approaches in clustering
- Features, correlations, numerical and categorical data
- Model selection
- Clustering and plotting
- Discussion

Case Study 3 - Supervised Learning: Personal Credit Scoring

- Main approaches in supervised learning
- Features, correlations, normalization and feature reduction
- Model selection
- Logistic regression and plotting

- Discussion

Case Study 4 - Dynamic Plots: Visualizing Piketty's *Capital*

- Plot types, automatic coloring, subplots
- Designing interactive elements
- Publishing dynamic plots on the web

Case Study 5 - Fraud Detection via Graph Analysis: Enron Email Dataset

- Graph construction
- Incidence matrix
- Centrality, degrees,
- Graph visualization

Case Study 6 – Character Recognition from Images

- Main approaches in image processing
- Review of classification methods
- Model selection and support vector machines
- Discussion