Modern data analysis - selected topics in Data Science

Wojciech Krzemień
**********

In the last years, with the constant development of the available computing powers, there has been a surge of interest in statistical modeling, classification, and prediction. Especially, the so-called machine learning, which combines techniques from the fields of statistics, computer science, and artificial intelligence, has attracted a lot of attention. Those methods are well suited to process large, heterogeneous data sets and can have a vast spectrum of applications such as discovery and prediction of social network trends, disease pattern searches, image recognition or particle identification for high energy physics experiments. Although many of those techniques have been known for years (e.g. neural networks) some recent results (e.g. application of unsupervised learning) brought the hope, for further progress in this domain.

The proposed course includes a selection of methods and tools of modern data analysis and the basics of Monte Carlo simulations. The main focus will be put both on the understanding of key statistical ideas which lay the foundation of those methods, and examples of algorithm implementations based on its simplified versions. Also, standard existing programming tools like scikit-learn will be described. It is assumed that the participants of the course have the basic knowledge of statistics (at the level of a typical university course) and basic programming skills. The algorithm example will be implemented in the Python programming language.

Nowoczesne metody analizy danych - Wybrane tematy
**********

W ostatnich latach, wraz z rozwojem mocy obliczeniowych komputerów, można zaobserwować znaczący wzrost zainteresowania metodami modelowania, klasyfikacji i predykcji statystycznych. Na popularności zyskały w szczególności metody uczenia maszynowego(machine learning), które łączą w sobie techniki z zakresu statystyki, informatyki i sztucznej inteligencji. Metody te dobrze nadają się do przetwarzania dużych zbiorów danych, przy czym charakter tych danych może być bardzo różnorodnych i obejmować zagadnienia typu wyszukiwanie trendów wśród użytkowników sieci społecznościowych, wyszukiwanie wzorców chorobowych, rozpoznawanie obrazów lub np. identyfikacja typów cząstek w eksperymentach wysokich energii. Mimo, że wiele z tych technik ma już dosyć długą historię (sieci neuronowe), to szereg ostatnich wyników (szczególnie metody unsupervised learning) przywróciły nadzieję, na dalszy rozwój tej dziedziny.

Proponowany kurs obejmuje wybrane metody analizy danych oraz podstawy symulacji Monte Carlo. Nacisk zostanie położony zarówno na zrozumienie idei statystycznych leżących u podstaw tych metod jak i na pokazanie implementacji uproszczonych wersji algorytmów. Dodatkowo, przedstawione zostaną także istniejące narzędzia typu scikit-learn. Zakłada się, że uczestnicy kursu powinni umieć programować na poziomie podstawowym, a

także posiadać wiedzę na poziomie typowego kursu uniwersyteckiego ze statystyki. Implementacja przykładowych algorytmów będzie wykonywana w języku Python.

Tentative list of topics (not in order of the presentation):

*Linear Regression
*Logistic Regression
*Neural Networks
*Models evaluations: bias-variance trade-off, ROC curve,...
*Support Vector Machines
*Decision Trees and Random Forests
*KNN
*Clustering K-means clustering
*Dimensionality reduction - Principal Component Analysis
*Bayesian classifiers
*Basic Monte Carlo techniques:
*naive MC
*Rejection sampling
*Importance sampling
*Variance reduction
*Re-sampling techniques
*Genetic algorithms