Foundations of Machine Learning

Presentation

Learning objectives

The proliferation of sensors along with data collection and recording systems is such that the amount of information available to users is too large to be processed without the help of high-performance IT tools and algorithms. Among the algorithms that are essential for data processing, classification algorithms are widely used, either to aggregate data into coherent groups (clustering or unsupervised classification), or to decide on the automatic assignment of new data to groups that have already been formed (supervised classification).

The learning objectives of this course are to :

- be familiar with and know how to implement the main supervised and unsupervised classification algorithms,

- be familiar with the conditions for implementing these algorithms and the prerequisites for any data pre-processing,

- be able to quantitatively assess the quality of these algorithms.

- Filtering and prediction of temporal sequences.

Description of the programme

Supervised classification: data management (creation of training, validation and test sets). Metrics in supervised classification (recall, precision, ROC curves and area under the curve, confusion matrices). Details of the main supervised classification algorithms: k-nearest neighbours (KNNs), wide margin separators (linear and kernel SVMs), random trees and forests, neural networks.
Unsupervised classification: data pre-processing (dimension reduction). Details of ascending hierarchical classification methods (study of dissimilarity criteria), k-means and Gaussian mixtures (EM algorithm).
Time series: statistical models, autoregressive models (statsmodels, ARIMA, ARIMAX, SARIMA, etc.)

Implementation and manipulation of these methods using the python library sklearn.

Generic central skills and knowledge targeted in the discipline

- Data pre-processing
- Choice and evaluation of a classification algorithm

- Presentation of classification results (presentation of metrics and/or graphical representation)
- Good knowledge of the sklearn library

How knowledge is tested

Implementation of classification algorithms on real data (iris, moon, mnist, telecom churn and cardiovascular disease prediction data) or simulated data ((non)linearly separable data, data from (multi)-normal distributions) and critical analysis of the results.

Bibliography

- *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow*, 2nd edition, Aurélien Géron, O' Reilly Media, 2019, 600 pp., ISBN: 978-1-492-03264-9
- Vapnik, V. Statistical Learning Theory. Wiley-Interscience, New York, (1998)

Dinov, ID. "Expectation Maximization and Mixture Modeling Tutorial". *California Digital Library*, Statistics Online Computational Resource, Paper EM_MM,

Teaching team

- Valeriya STRIZHKOVA

Total des heures		**22h**
CM	Master class	22h