Data Analysis

Person in charge: Jérémie Sublime
Prerequisite: Algorithm, statistics
Organization: 7 x 3h Lectures and 7 x 3h of lab practice
Evaluation: Regular reports on lab practice, Project, Final Exam
ECTS: 5 credits

Context

Data analysis is an exploratory task the aim of which is to process data with the goal of discovering interesting and relevant knowledge. This process includes data cleaning, data transformation and data modeling. The field of data analysis combines methods coming from several domains such as statistics, supervised learning, clustering, visualization and modelling. It is an increasingly active domain used in both academic research and the industrial world for a wide range of application such as decision systems, recommender systems, classification systems, etc.

Objectives

This course will provide the students with a broad range of knowledge on methods and concepts related to the analysis of multidimensional data, and to enable them to understand and interpret the results of these methods.

Knowledge

This module enables students to develop the following concepts and skills.

- **Concepts**
    - Univariate statistics (revisions), multivariate statistics
    - Principal component analysis
    - Linear and multilinear regression, logistic regression
    - Introduction to unsupervised and supervised learning
    - Clustering
    - Data visualization

- **Know-How**
  - Being able to identify different types of data : univariate, multivariate, binary, numerical, categorical, qualitative, quantitative, etc.
  - Being capable of using the different statistical methods seen in class and to identify when to use which: central tendency measures, dispersion measures, correlation, principal component analysis
  - Understanding and knowing when to use the different advanced data mining methods seen in class: regression, clustering, classification
  - Being able to interpret and explain the results of the aforementioned methods.
  - Implementing and using these methods with R

Pedagogical Approach

Each lecture will be followed by 3h of lab practice to apply and test the methods seen during the lecture. A project starting in the middle of the semester will allow the students to tackle real data and to re-use several methods seen in class.

References

- Ad Feelders, "*Advanced Data Mining*" (2011)
- Srinivasan Parthasarathy, "*Introduction to Data Mining*"
- Christopher M. Bishop, "*Pattern Recognition and Machine Learning*" (2006)
- R. O. Duda, P. E. Hart, D. Stork, Wiley and Sons, "*Pattern Classification*" (2000)