COURSE in Unsupervised Learning

INSTRUCTOR(S): Alessio Farcomeni (Ph.D.); Francesco Dotto (Ph.D)
EMAIL: alessio.farcomeni@uniroma2.it
francesco.dotto@uniroma3.it

COURSE DESCRIPTION

The course covers the main statistical techniques used to find groups in the data (i.e., identify discrete structures not directly observed) and, even when there is only one group, outliers. Furthermore, it discusses dimensionality reduction methods used to summarize data in few dimensions, create rankings/scores, compress information. Principles of robust estimation are also introduced.

As an example, consider a company and its customer database where for each customer (unit) a number of characteristics (variables) linked with customer behaviour are measured: number of visits, total amount spent, overall approval of services, etc. Unsupervised learning techniques can help us find answers to questions like: are there different types of customers? If yes, how many and what are their profiles? Are there few very unusual customers?

Dimensionality reduction techniques can help us find answers to questions like: how can we rank customers with respect to propensity of making business with us? How can we plot the data summarizing all of the information? What kind of information is available in the data available?

LEARNING OUTCOMES

- ✓ ability to use statistical learning techniques in the presence of unmeasured data labels
- ✓ ability to perform anomaly detection at basic level
- ✓ ability to evaluate and compare the resulting grouping structure
- ✓ ability to reduce data dimensionality for descriptive and scoring purposes

METHODOLOGY

Emphasis is on principles and specific models/techniques. Each method is introduced by examples and described in mathematical formulas. Some math is essential but very few derivations are made. Models and techniques are discussed from a theoretical and practical point of view, describing their definition/properties and their implementation by using statistical software *R*. Particular importance is given to the interpretation of results. Computer laboratory hours give to the students the possibility to practice what they learn.

ASSESSMENT

Written exam; weighting: 70%
Project; weighting: 30%

Group assignments aim at assessing the capabilities of analyzing data, as well as the ability to communicate the relevant findings. The students are expected to produce a technical report no longer than 6 pages. The final exam is a written test of 60 minutes containing 5 short questions and 10 multiple-choice questions.

OUTLINE

1. Introduction and overview
2. Non-Hierarchical Clustering Methods
    - 2.1 K-means
    - 2.2 PAM, Clara
    - 2.3. Robust methods
        - 2.3.1 Trimmed K-means
        - 2.3.2 Snipped K-means
        - 2.3.3 Anomaly detection
3. Hierarchical Clustering
    - 3.1 Single linkage
    - 3.2 Other linking methods

4 Dimension Reduction through Principal Component Analysis (PCA)

Other topics might be mentioned, if time permits.

TEACHING MATERIAL

The course material will be made available during the course, including slides, data sets, and supplementary materials (scripts in *R*, etc.).

TEXTBOOKS

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, Springer Series in Statistics. https://web.stanford.edu/~hastie/ElemStatLearn/

Farcomeni, A. and Greco, L. (2015) Robust Methods for Data Reduction, Chapman & Hall/CRC Press

ADDITIONAL SUGGESTED READING

Chatfield, C. and Collins, A. J. (1981) Introduction to Multivariate Analysis, Chapman & Hall/CRC Press

Witten J.D., Hastie T., Tibshirani R. (2014). An Introduction to Statistical Learning with Applications in R. Springer, Springer Series in Statistics.