COURSE: Unsupervised Learning

INSTRUCTORS: Roberto Rocci (Ph.D.); Monia Ranalli (Ph.D.)
EMAIL: roberto.rocci@uniroma2.it
monia.ranalli@uniroma2.it

COURSE DESCRIPTION

The course covers the main statistical techniques used to identify latent structures (i.e. structure not directly observable) in the data. Technically, these are formalized as the identification of one or more latent variables underlying the data. Depending on the type of the latent variables we obtain techniques able to reduce the number of variables/characteristics (principal component analysis and factor analysis) and/or the number of units/subjects (cluster analysis and finite mixture).

As an example, consider a company and its customer database where on each customer (unit) are recorded a number of characteristics (variables): age, sex, amount spent, types of products purchased, etc. Unsupervised learning techniques can help us find answers to questions like:

– [variables reduction] are there relationships (correlations etc.) between the observed variables? If so, which ones? Can these be summarized in one or more prototype/latent variables able to highlight the different purchasing behaviour of customers?

– [units reduction] are there different types of customers? If yes, how many and what are they?

LEARNING OUTCOMES

- study some statistical learning techniques and models for data reduction
- appreciate why and when these methods are required
- provide some understanding of techniques used in the literature
- promote use of useful techniques in your research
- learn a statistical language/software (R)

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 only 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 a statistical package. Some hours of computer laboratory give to the students the possibility to practice what they learn.

ASSESSMENT

The assessment consists of

30% group assignment

70% final written exam

The group assignments aim at assessing the capabilities of analysing data, as well as the ability to communicate the relevant findings. The students are expected to produce a technical report no longer than 8 pages.

The final exam is a written test of 90 minutes containing 10 short questions and 15 multiple-choice questions.

OUTLINE

1. Introduction to Unsupervised Learning
2. Non model based techniques
   - 2.1 Principal component analysis
   - 2.2 Cluster analysis:
     - 2.2.1 K-means
     - 2.2.2 Ward
     - 2.2.3 Link methods
3. Model based techniques
   - 3.1 The multivariate Gaussian distribution
   - 3.2 Factor analysis
   - 3.3 Finite mixture models
     - 3.3.1 Introduction
     - 3.3.2 EM algorithm
     - 3.3.3 Mixture of Gaussians
     - 3.3.4 Mixture of linear regressions

TEACHING MATERIAL

The course material will be made available during the course: slides, readings, datasets, supplementary materials (scripts in R etc).

SUGGESTED READING

Bishop C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.
Marden J.I. (2015). *Multivariate Statistics*. http://stat.istics.net/Multivariate/
McLachlan G.J., Peel D. (2000). *Finite Mixture Models*. Wiley, New York.
Duda R.O., Hart P.E., Stork D.G. (2001). *Pattern Classification*. Wiley, 2nd Edition.

ADDITIONAL SUGGESTED 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. http://www-stat.stanford.edu/ElemStatLearn/

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