Introduction to Statistical Learning

B. A. in Global Governance

University of Rome Tor Vergata

Rosario Barone*

*Email: rosario.barone@uniroma2.it

Statistical Learning

- *Statistical learning theory* is a framework for machine learning
- Mainly interested in prediction
- Applications in text mining, image processing, speech recognition, bioinformatics etc..

Prerequisites

- Introductory Statistics
- Mathematics (Probability theory)
- Statistical Inference

Textbook

- Witten J.D., Hastie T., Tibshirani R. (2014). An Introduction to Statistical Learning with Applications in R. Springer, Springer Series in Statistics
- Chatfield, C. and Collins, A. J. (1981) Introduction to Multivariate Analysis, Chapman & Hall/CRC Press
- Everitt, B. S. and Hothorn, T. (2006) A Handbook of Statistical Analyses Using R. CRC Press. Available for free at: `http://www.ecostat.unical.it/tarsitano/Didattica/LabStat2/Everitt.pdf`

Additional (more technical) reading:

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

Informations

- Statistical Learning, 9 ECTS:
  - 6 ECTS will be held by me
  - 3 ECTS will be held by Prof. Alessio Farcomeni
- Midterm exam: October 26, 2022.
- Written exam.
- An oral proof may be required from the students.

Main topics

- Introduction to R software
- Linear regression
- Logistic regression
- Logistic regression
- Principal component analysis
- Cluster analysis
- Machine learning methods for supervised learning
- Modern applications: text mining, image processing

First Week plan

- Introduction to the R software, practicums on the pre-requisites.

- You are encouraged to replicate my analyses on your laptops.

Statistical software: why?

Statistical software are specialized computer programs for analysis in statistics and econometrics.

- Huge amount of data are modelled in a simple way (descriptive or inferential analyses are easily manageable)
- Often the estimation of unknown model parameters requires numerical approximation, e.g. often there is no closed form solution to the maximization (or integration) of the likelihood function
- Programming languages allow to extend extisting models or develop model created *ad hoc* in particular circumstances

R and RStudio

1. R is a programming language for statistical computing and graphics
2. RStudio is an Integrated Development Environment (IDE) for R

Main features:

- Data processing
- Programming

*Several tutorials for the download and installation are online available: for example https://www.youtube.com/watch?v=cX532N_XLIs

R and RStudio

R packages (libraries) are extensions to the R statistical programming language.

- Libraries contain code, data, and documentation
- R is open source: packages are developed from users (in most of the cases researchers)
- Nice introduction to R: https://www.slideshare.net/bcbbslides/an-introduction-to-r