Laurea / B.A.
in Global Governance

Academic Year 2021-2022
Syllabus
Statistical Learning
9 CFU
Dott.ssa Barbara Guardabascio - Prof. Alessio Farcomeni

Course Description

The course covers some statistical techniques for supervised and unsupervised learning. The R software for statistical computing will be also introduced and used throughout.

Supervised learning techniques are used to predict a target variable (linear and logistic regression) based on predictors, and/or to assess interrelationships among predictors and a target variable (linear and logistic regression). As an example, suppose you want to predict the risk that a family will be materially deprived next year. This can be done by using data that can be measured at baseline (number of family members, disposable income, health status, etc.) and use these to predict material deprivation for a sample of families with known status. Incidentally, you will also understand how health status affects the risk of material deprivation.

Unsupervised learning techniques are used to find groups in data, that is, to predict target categorical variables that are not measured (cluster analysis). Additionally, they are used to summarize data (dimension reduction, done with principal component analysis in this course). As an example, suppose you want to assess an unmeasurable trait, like happiness. Suppose your target units are geographic regions. Happiness can be measured indirectly through a series of variables (questionnaires, indices, etc.). A general score is obtained through dimension reduction by finding the optimal weighted average of all measurements. Cluster analysis will separate regions in few (two, three, four) groups, with respect to levels of happiness. Different policies can then be scheduled for each group. The last 3 CFU will be dedicated to machine learning methods (classification and regression trees, random forests, shallow and deep neural networks) for supervised learning. Modern applications will be then introduced, where data is extracted from text corpora (natural language processing), images (computer vision), audio tracks.

The main objectives of this course are to provide students with the ability to select the statistical learning technique needed to answer specific questions (based on data), to perform data analysis appropriately, and to interpret the results correctly.

Prerequisites

Prerequisite is an introductory statistics and statistical inference course like “Statistical Tools for Decision Making” of the B. A. in Global Governance. Also some math is essential, but only few derivations are made.

Teaching Method

The course is carried out through lectures and practicums. Techniques will be introduced by examples and described in mathematical formulas. Focus will be on the practical implementation of each technique, and interpretation of results. In the final part of the lesson students will be able to practice the newly introduced topics.

Schedule of Topics

Topic 1	Introduction to R software
Topic 2	Linear regression
Topic 3	Logistic regression
Topic 4	Principal component analysis
Topic 5	Cluster analysis
Topic 6	Machine learning methods for supervised learning
Topic 7	Modern applications: text mining, image processing

Textbook and Materials
Reading material on each course topic (handouts, slides, data sets, R scripts), will be made available to the students by the course instructors during the course.

Suggested books are:

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/

Assessment
If in presence, assessment will be based on a written exam. This will include closed and open questions. If online, assessment will be carried out through a written test for the first five topics; while assessment for the last two topics will be carried out separately through a brief oral discussion, which might include an *impromptu* practicum.

Office hours
upon appointment by e-mail

E-mail
guardabascio@istat.it
alessio.farcomeni@uniroma2.it