Statistical Learning

Tommaso Proietti

DEF Tor Vergata

An introduction to Statistical Learning and Data Mining

Introduction I

This course is about *statistical learning* with (possibly big) data.

- Statistics has been defined as the art (or science) of learning from data. "Statistics concerns what can be learned from data" (A.C. Davison, Statistical Models, CUP, 2003).
- *Data Mining* as the "process of seeking interesting or valuable information within large data sets" (D. Hand et al., Statistical Science, 2000).
- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner". (Hand, Mannila & Smyth, Principles of Data Mining, MIT Press, 2001).
- Nowadays a more popular word is *Data Science*.

Introduction II

- The course offers an insight into the main statistical methodologies for the synthesis and visualisation and the analysis of business and market data, Emphasis will be given to empirical applications using modern software tools (R and Matlab).
- The aim of the course is to provide students with the most relevant analytical tools that are useful for addressing relevant research questions. This involves being able to extract information from complex data readily available from the economic environment, as well as being able to analyse this data in a way that leads to useful models and methods for the prediction of future outcomes.
- Lecturer:
  Tommaso Proietti
  Office: Room 47 Dipartimento di Economia e Finanza (B Building, 2nd Floor)
  Office hours: typically Monday 2-4 p.m. (tommaso.proietti@uniroma2.it)

Syllabus

1. Introduction to statistical learning. Tools for data analysis, visualisation and description.
2. The linear regression model.
3. Model selection and evaluation: bias-variance trade-off, model complexity and goodness of fit. Cross-validation. Selection using information criteria.
4. Regularization and shrinkage methods: rigde regression, lasso, forward stagewise regression. Methods Using Derived Input Directions: principal component regression.
5. Linear methods for classification: Bayes Classification Rule. Discriminant analysis. Canonical variates. Logistic regression.
6. Semiparametric regression: Regression splines and smoothing splines.
7. Kernel smoothing methods: Local polynomial regression. Density estimation. Nearest neighbour classification.
8. Additive Models, tree-based methods. GAM, Regression and classification trees. Boosting.

Assessment

- 30% Group Assignments
- 70% Final Exam

The final exam is a 2-hour written paper. The examination will contain direct questions on material from the course. This is a closed book exam. A pocket calculator may be needed. It is a written paper consisting of three main questions and a set of shorter questions. Students are expected to describe the statistical methods covered in the second part of the course and their applicability to real life situations. Both theoretical and practical (computational) aspects should be covered. The short questions aim at assessing the students' capability to interpret the statistical output and to diagnose the goodness of fit of a statistical method.

Where to study

The main references for the course are

- G James, D Witten, T Hastie, and R Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, Springer Series in Statistics, 2013. Dowloadable at www.statlearning.com
- T Hastie, R Tibshirani and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, Springer Series in Statistics, 2009.
  Website: http://www-stat.stanford.edu/ElemStatLearn/.

Slides, readings, datasets and supplemental material will be made available in the course website.

Statistical learning and data mining

We distinguish two main learning problems.

- Supervised learning. Data are grouped or ordered by some response. We wish to predict an outcome variable (output, dependent variable) from a set of features or characteristics (inputs).
  - Classification (prediction of a binary or multinomial outcome)
  - Regression (prediction of a quantitative outcome)
- Unsupervised learning: no outcome is available (cluster analysis, multidimensional scaling, principal components)

A typical dataset has at least two dimensions: individuals and variables (measurements). They lend themselves to the representation as **data matrices**. Three dimensional arrays (where the third dimension is time or space) are also common. It is important to distinguish variables types and their scale of measurement.

Measurement scales

1. Qualitative (Categorical)
   - Nominal (binary, multinomial). The measurement deals with the allocation of a case to a response category. Examples: sex, marital status, solvency. We can compare measurements according to the identity principle (equal or different).
   - Ordinal. Response categories are ordered. We can state which category is higher or lower.
2. Quantitative
   - Interval scale. The origin is arbitrary. ({Temperature in degrees Celsius}). We can compare two measurements using the algebraic sum (e.g. y_1 and y_2 differ by $y_1 - y_2$)
   - Ratio scale. The measurements have a natural zero (e.g. sales, n. of customers). We can compare measurements using their ratio.