Statistical Learning
RED–Rome Economics Doctorate
Spring 2022

Syllabus

Instructor

Professor Franco Peracchi (franco.peracchi@uniroma2.it)
Office hours: TBA.

Lectures

TBA

Goal

The aim of this course is to introduce students to a set of tools for modeling and prediction with complex (long and wide) datasets. This is a recently developed area in statistics and econometrics which blends with parallel developments in computer science, in particular machine learning. The course encompasses a variety of methods, both frequentist and Bayesian, including classical methods for regression and classification; asymptotic approximations vs. resampling methods; model uncertainty and model selection; model averaging; shrinkage estimators; principal components and partial least squares; linear regression smoothers; projection pursuit, generalized additive models, and neural networks; tree-based methods.

Software

This course relies on both R, a free software environment for statistical computing and graphics, and Stata, a commercial statistical package with excellent data management and graphics capabilities. Both run on MacOS, Unix and Windows. You can freely download the most recent version of R, version 4.0.4 ("Lost Library Book"), from your preferred CRAN mirror (http://cran.r-project.org/mirrors.html).

Grades

Homework 33%, Final exam 67%.

Homework

Spending a significant amount of time each week on the assigned homework is essential to learning the material covered. Homework must be handed in class, on the dates indicated below. There is no credit for late homework. Working in group (up to 3 people) is strongly encouraged but each student needs to hand in her/his own solution.

Homework due dates:

- Problem set 1: TBA

- Problem set 2: TBA
- Problem set 3: TBA

Final exam

It is a take-home exam to be completed in one week. The exam covers all the material discussed in the course and includes an empirical part to be carried out using either R or Stata.

Course Outline

1. Approaches to statistical learning: Classical, Bayesian, Fisherian.
2. Linear vs. nonlinear models for prediction and classification.
3. Asymptotic approximations vs. resampling methods.
4. Model uncertainty and model selection: Classical pre-test estimators, model selection criteria, cross-validation, post-selection estimators, model averaging.
5. Shrinkage: James-Stein estimators, ridge regression, LASSO and extensions, penalized M-estimation.
6. Dimensionality reduction: Principal component regression, partial least squares.
7. Smoothing: Polynomial regression, splines, kernel and nearest neighbor methods, local polynomial fitting.
8. High-dimensional data: Projection pursuit, generalized additive models, neural networks.
9. Tree-based methods: Decision trees, bagging, random forests, boosting.

References

The recommended references are:

- Hastie T., Tibshirani R., and Friedman J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer: New York [ESL]. Available at http://web.stanford.edu/~hastie/Papers/ESLII.pdf.
- James G., Witten D., Hastie T., and Tibshirani R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer: New York [ISL]. Available at http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf. Also see the book's website at http://www.StatLearning.com.

Additional references are:

- Breiman L., Friedman J. H., Olshen R. A., and Stone C. H. (1984). *Classification and Regression Trees*. Wadsworth and Brooks/Cole: Monterey, CA.

- Bühlmann P., and van de Geer S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer, Berlin.
- Efron B., and Hastie T. (2016). *Computer Age Statistical Inference. Algorithms, Evidence, and Data Science*. Cambridge University Press: New York [CASI]. Available at http://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf.
- Efron B., and Tibshirani R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall: New York.
- Hastie T., Tibshirani R., and Wainwright M. (2013). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall: New York [SLS]. Available at http://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS.pdf.
- Lancaster T. (2004). *An Introduction to Modern Bayesian Econometrics*. Blackwell: Malden, MA.
- Leamer E. E. (1978). *Specification Searches*. Wiley: New York. Available at https://www.anderson.ucla.edu/faculty_pages/edward.leamer/books/specification_searches.htm.
- Wasserman L. (2006). *All of Nonparametric Statistics*. Springer: New York.

Suggestions for further reading will be provided in class.