Statistical Learning
RED–Rome Economics Doctorate
Spring 2024

Syllabus

Instructor

Professor Franco Peracchi (peracchi@uniroma2.it)
Website: http://www.eief.it/eief/index.php/people/faculty-az?id=174.
Office hours: Thursday 4:00–5:30 pm, or by appointment.

Lectures

Monday, Tuesday and Wednesday, 11:00 am–1:00 pm, for three weeks, from March 11 to March 27, 2024.

Goal

The goal of this course is to introduce students to a set of tools for prediction, classification, and causal analysis 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 supervised learning methods, derived from both frequentist and Bayesian approaches, including "classical" methods for regression and classification; asymptotic approximations vs. the bootstrap and other resampling methods; model uncertainty, pre-testing and post-selection estimators; shrinkage estimators; principal components and partial least squares; linear smoothers; projection pursuit, generalized additive models, and neural networks; clustering; tree-based methods; and causal learning.

Software

This course relies on both R (https://www.r-project.org), a free software environment for statistical computing and graphics, and Stata (https://www.stata.com), a commercial statistical package with excellent data management and graphics capabilities, plus its own programming language (Mata). Both run on MacOS, Unix and Windows, and are integrated with Python (https://www.python.org). You can freely download the most recent version of R, version 4.3.2 ("Eye Holes"), 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: March 18.
- Problem set 2: March 25.
- Problem set 3: April 1.

Final exam

Following the exam/grading guidelines of RED–Rome Economics Doctorate, the file exam is a classroom exam, scheduled for TBD.

Grading is in decimals with a maximum grade of 31. The minimum grade for a Pass is 18.

The exam covers all the material discussed in the course. The questions will resemble those assigned in the homework.

Course outline

- Lecture 1 (March 11). Introduction. Approaches to statistical learning: Frequentist, Bayesian, Fisherian and the maximum likelihood method.
- Lecture 2 (March 12). Linear models for prediction and causal analysis.
- Lecture 3 (March 13). Nonlinear models for prediction and classification. Asymptotic approximations vs. resampling methods.
- Lecture 4 (March 18). Model uncertainty and model selection: Classical pre-test estimators, model selection criteria, cross-validation, post-selection estimators.
- Lecture 5 (March 19). Shrinkage: James-Stein estimators, ridge regression, LASSO and extensions, penalized M-estimation. Dimensionality reduction: principal component regression, partial least squares.
- Lecture 6 (March 20). Linear smoothers: Polynomial regression, splines, kernel and nearest neighbor methods, local polynomial fitting.
- Lecture 7 (March 25). Flexible learning methods with many covariates: Projection pursuit regression, additive and generalized additive models, neural networks. Clustering: K-means and hierarchical clustering (if time permits).
- Lecture 8 (March 26). Tree-based methods: Decision trees, bagging, random forests, boosting.
- Lecture 9 (March 27). Causal learning: Double-selection estimation, post double-selection inference, causal trees.

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 https://hastie.su.domains/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 [ISLR]. A 2023 version with applications in Python is also available. Both can be dowloaded from the book's website at http://www.statlearning.com.

Additional references include:

- Cerulli G. (2023), *Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata*. 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 https://hastie.su.domains/CASI_files/PDF/casi.pdf.

- Efron B., and Tibshirani R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall: New York.

- Hansen B.E. (2022) *Econometrics*. Princeton University Press: Princeton (NJ).

- Hastie T., Tibshirani R., and Wainwright M. (2013). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall: New York [SLS]. Available at https://hastie.su.domains/StatLearnSparsity_files/SLS.pdf.

- Huber M. (2023). *Causal Analysis. Impact Evaluation and Causal Machine Learning with Applications in R*. MIT Press: Cambridge (MA).

- 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.

- Wainwright M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press: Cambridge, UK.

- Wasserman L. (2006). *All of Nonparametric Statistics*. Springer: New York.

Suggestions for further reading will be provided in class.