Barcelona GSE Graduate School of Economics

14D004

Computing Lab

3 ECTS

Overview and Objectives

The computing lab introduces students to programming techniques and hacking skills required for data science, it provides a solid training in computational statistics algorithms related to linear and generalised linear models including techniques for learning with massive datasets, and provides a hands-on experience on some practical but important aspects of machine learning, such as feature engineering and validity measures. The computing lab focuses on the use of scientific scripting languages and special attention is devoted to the R and Python language and working in a unix environment.

Course Outline

The course covers the following list of topics:

A. Introduction to python for data mining

The jupyter notebook programming environment
Introduction to Python and iPython
Data manipulation using numpy and pandas
Plotting libraries (matplotlib, pyplot, seaborn)
Introduction to data mining with python (studying a use-case)

B. Regression modelling with large data sets

Linear and generalized linear regression models (model definition, estimation and inference)
Basic regression algorithms (least squares problem and solutions, complexity and memory, iterative re-weighted least squares)
Learning with large data sets (incremental bounded-memory and stochastic gradient descent algorithms for large-n estimation, inference, software and tools, analysis of real datasets, extensions and related large-p estimation problems)

C. Applied machine learning in R

Unsupervised learning (clustering algorithms, number of clusters, validity measures)
Supervised learning (classification algorithms, hyperparameter fine-tuning, training and testing)
Data transformations (feature engineering, attribute selection) and performance optimisation

14D004

Computing Lab

3 ECTS

Required Activities

Attendance at classes, and submission of homeworks

Evaluation

3 projects in total, one per module

- The projects will be individual
- Will be given at the last class of each part
- Students will be given 2 weeks to submit their project.

The grade will be the average of these 3 projects.

Materials

Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Elsevier Inc, 2011

Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

ỖSoftware for Data Analysis: Programming with Rỗ, John Chambers, Springer 2008

Chambers, J. M. (1971). Regression updating. Journal of the American Statistical Association 66(336), 744Ố748

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction (Second ed.). Berlin; New York: SpringerVerlag Inc

Golub, G. H. and C. F. Van Loan (2012). Matrix Computations (4th Ed.). Baltimore, MD, USA: Johns Hopkins University Press

Toulis, P. and E. Airoldi (2016). Asymptotic and finite-sample properties of estimators based on stochastic gradients. Annals of Statistics