RUHR
UNIVERSITÄT
BOCHUM RUB

Data Science (Master students only)

August 27-31 2018

Lecturer

Paola Cerchiello Ph.D. Professor of Statistics at the University of Pavia, she got BS in Economics at the University of Pavia in 2002 (degree with Honors), Degree IUSS (Scuola Universitaria Superiore Pavia) in 2002 and PhD in Statistics, University Milan-Bicocca, 2006.

She teaches Statistics and Big Data Analysis in Master Degree in International Business and Management (MIBE) at Unipv, she also teaches Data Science in II level master Madas of Collegio Carlo Alberto-University of Turin. Senior Data Scientist at Data Science Lab in Economics and Finance where she carries out research and consulting projects, for leading institutions such as the European Union,the Italian Ministry of Research, the Italian Banking Association, Intesa San Paolo, SAS Institute, Sky etc.

She is member of research groups on Big Data at the Bank of Italy and at the Deutsche Bundesbank.

Her research activity is mainly devoted to the study of statistical models for unstructured and complex data in economics and finance. From an applied viewpoint, she focuses on text data analysis, systemic risk and reputational risk, sentiment analysis and deep learning. She has published around 40 papers in scientific international journals.

Course objectives

The aim of this course is to provide students with the knowledge of the most important algorithm of statistical learning and machine learning. At the end of the class, students will be able to apply practically the proposed models by employing one of the most relevant open source software, R.

Course content

The aim of this course is to study and apply the most relevant statistical models in the analysis of large and heterogeneous data set.

The perspective in mainly applicative: choosing and applying the most suitable model to exploit the informative power of data set with a particular attention to the correct and contextualized interpretation of final results.

The course will be held with the interactive employment of open source software R, to learn practically the complete workflow of a statistical analysis.

The covered statistical models are:

- The principles of statistical learning;
- Multivariate regression;

- Regularized Regression: Ridge and Lasso;
- Logistic Regression
- Decision Tree;
- Bagging, Boosting, Random Forest;
- Cross Validation & Model Accuracy

Prerequisites

Knowledge of basic concepts of statistics like inference, confidence interval, statistical test is required.
For the coding part, some familiarity with R language, or similar software, would be useful, but not strictly required.

Instructional methods

Both theoretical and practical lectures within a highly interactive framework.

Reading list

- *Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014.* ***An Introduction to Statistical Learning: With Applications in R****. Springer Publishing Company, Incorporated.*
- *Hastie, T., Tibshirani, R.,, Friedman, J. (2001).* ***The Elements of Statistical Learning****. New York, NY, USA: Springer New York Inc..*

Time schedule

6 hours of teaching per day (an hour lasts 45 minutes).
Example:
Lecture from 09.00 to 10.30; break from 10.30 to 10.45, and another lecture 10.45-12.15.
Lunch break: 12.15-13.15.
Afternoon: Lecture from 13.15 to 14.45.

Assessment

The final grade is based on a list of multiple choice and open questions.
The questions will regard both the theoretical and practical arguments.