SYLLABUS DEL CORSO

Data Mining M

2021-2-F8204B014

Learning objectives

The course aims to provide an advanced understanding of the core principles and techniques for data mining and supervised learning, and their implementation in the R software environment for statistical computing. Special emphasis will be given to applied predictive modelling.

At the end of the course, students will be able to analyse complex data sets by exploring, transforming and modelling the data.

Contents

The course integrates theoretical principles with practicals of data analysis and programming in R.

- Theory: bias-variance tradeoff, regularized estimation, splines and generalized additive models, conformal prediction

- Practice: how to get your data into R, get it into the most useful structure, transform it, visualise it and model it for prediction of future observations

Detailed program

- Prediction error and the bias-variance trade-off
- The model versus the modeling process
- Regularized estimation: *ridge*, *lasso* and *best subsets*
- Splines and generalized additive models
- Conformal prediction

- Computational aspects

Prerequisites

Knowledge of topics covered in the courses *Probability and Statistics M* and *Advanced Statistics M* is highly recommended.

Teaching methods

Lessons are held both in classroom and in lab, integrating theoretical principles with practicals of data analysis and programming in R.

Assessment methods

The exams consists of two parts:

1. written exam
2. homework

The final grade is determined by a weighted average of 1. and 2. The oral exam is optional.

Textbooks and Reading Materials

- Course repository: https://github.com/aldosolari/DM
- Arnold (2019) ⎯⎯⎯
- Azzalini, Scarpa (2004). *Analisi dei dati e data mining*. Springer-Verlag Italia
- Gareth, Witten, Hastie, Tibshirani (2013). *Introduction to Statistical Learning with applications in R*. Springer
- Hastie, Tibshirani, Friedman (2009). *The Elements of Statistical Learning.* Springer
- Kuhn, Johnson (2013). *Applied Predictive Modelling*. Springer
- Kuhn, Johnson (2019). ⎯⎯⎯
- Wickham, Grolemund (2015) *R for Data Science*. O'Reilly Cookbooks

Semester

First semester, first cycle.

Teaching language

The lessons are held in Italian, but most of the textbooks are in English.