UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining

2627-3-E4101B026

Learning objectives

The course aims to provide a comprehensive view of data mining, from the pre-processing of the data to the selection of the best statistical model for analysing and understanding the problem. During the course, the main techniques for data processing will be addressed and supervised statistical methods will be presented. Furthermore, concepts related to text mining will be introduced.
At the end of the course, the student will be able to compare and select the best Data Mining method for the problem under analysis. He/she will be able to deal with the main data issues and, independently, deal with a real problem in the best way.

The course contributes to the achievement of the learning objectives in the learning area of the three-year degree course: ‘Statistics’.

Contents

How to deal with missing values.
Supervised classification/regression methods.
Trade-off bias variance.
Text mining.
Market basket analysis.

Detailed program

1. Introduction to data mining.
2. Pre-processing: treatment of missing values. Single and multiple imputation methods.

3. Introduction to classification with examples and introductory concepts. Classification methods: linear discriminant, quadratic discriminant, k-nn and decision trees.
4. Trade off bias variance. Definition of overfitting and related mitigation techniques.
5. Text mining with examples and basic concepts: pre-processing (e.g. elimination of stop words) and graphical representations for text mining.
6. Market Basket Analysis and aPriori algorithm.

Prerequisites

Multivariate Statistical Analysis and R language.

Teaching methods

Lessons will be held both in the classroom and in the laboratory, integrating theoretical and practical-application aspects of data analysis and programming in R.

The 42 hours of teaching will be divided as follows:

- 30 hours of lectures;
- 12 hours of laboratory activities.

Assessment methods

Written
(20 out of 32) Written test aimed at verifying the topics presented in the classroom.

Project
(12 out of 32) Application project to be carried out independently or in a group (max. 3 people) on a dataset assigned by the lecturer or chosen by the students. The project is carried out in R and must demonstrate the ability to tackle a real problem in all its aspects using what has been seen in class. The project consists of both the R code and a presentation report produced using Rmarkdown.

Textbooks and Reading Materials

Main book:
Gareth J., Witten D., Hastie T., Tibshirani R., *An Introduction to statistical learning with application in R*, springer (2013).

Useful reading materials for R:
W. N. Venables, D. M. Smith and the R Core Team, An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics.
https://cran.r-project.org/doc/manuals/R-intro.pdf
C. Agostinelli, Introduzione a R. https://cran.r-project.org/doc/contrib/manuale.0.3.pdf

Further readings:
http://www.feat.engineering

Materials will be also provided during the course.

Semester

II Semester - III period

Teaching language

Italian

Sustainable Development Goals

QUALITY EDUCATION