UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining

2122-3-E4102B085-E4102B086M

Learning objectives

Data mining

The course aims at introducing statistical models of DATA MINING both from the theoretical and from the applicative point of view.

The student at the end of the course should be able to understand, discern and propose complex models and algorithms, being able to assess the studied topics analyzing read dataset.

Contents

The course deals with complex/algorithmic modelling techniques and main problems and algorithm of Data Mining

Detailed program

Data mining

Principles of Data mining, robustness, over fitting and validation. Association rules, Statistical models: linear, discriminant analysis, logistic models, (binary and multinomial), Algorithms for the classification: (Naive Bayes, Nearest Neighbour, lasso regression, neural network, Classification TREE, PLS, Bagging, Boosting and Random forest)

Prerequisites

Students need to pass before the exam of Analisi statistica Multivariata

Teaching methods

During Covid-19, lessons will be taken by partial presence and streeming web platforms.

Assessment methods

<u>WRITTEN EXAM: PROJECT WORK</u>

Project work (also in group, to complete before the date of the oral exam) involving a data analysis (R or SAS) on a dataset chosen by the student to replicate arguments and analyses discussed during lab sessions.

Analyses of the Project work

Data mining (sas Entreprise Miner or R)

1 applied work with binary target (classification)

(To do: descriptive analysis, propose different classifiers and validation strategies, preprocessing, tuning of models, assessment, score of new data)

Web portals for the choice of the dataset:

https://archive.ics.uci.edu/ml/datasets

www. kaggle.com

<u>ORAL EXAM</u>

The outputs of the project work (completed during the period before the oral exam) must be printed and presented/discussed at the oral exam, IF EXAMS ARE HOLD IN PRESENCE. OTHERWISE, THE DISCUSSION OF THE PROJECT WORK via WEB platforms (during COVID19)

The oral exam deals with questions on statistical THEORY (see arguments) and on the comments of outputs of the project work to assess the comprehension of principal statistical tools and consequently the "modus operandi" of the conducted statistical analyses.

The student should demonstrate to understand, discern and explain the functioning of complex models and algorithms, being able to explain the studied topics and to analyze real dataset.

To resume, to pass the exam the student should complete two project works, one for statcomp, one for data mining.

Textbooks and Reading Materials

Data mining

Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R

http://www-bcf.usc.edu/~gareth/ISL/

Chapter 2-3-4-5- 8

Handouts on moodle

Semester

I semester cycle II

Teaching language

ITA