UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Mining and Machine Learning

2223-3-E4102B087

Learning objectives

Data Mining e Machine Learning

Data Mining Section
This section aims at introducing complex methotodologies for modelling statistical models both from the theoretical and from the applicative point of view

Machine learning section
The course aims at introducing statistical models of Machine learning 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 real dataset with R studio.

Contents

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

Detailed program

Data Mining section

- (1) SAS language and R (overiview)

- (2) Interpretation of complex linear Models (Anova, Ancova, GLM)
- (3) Robust methods (Bootstrap, Jacknife, Robust Regression, IRLS, WLS, nonparametric regression, loess smoothing and splines)
- (4) Step of robust model building
- (5) missing data mechanism, missing imputation, (y, X)-transformation, Influence, diagnostics, eteroskedasticity, model selection

Machine learning section

Problems with large dataset, robustness, overfitting and validation strategies. Association rules, Statistical models: linear, discriminant analysis, logistic models, (polytomic and ordinal), Algorithms for the classification: (Naive Bayes, Nearest Neighbour, Neural Network, Classification and Regression TREE, PLS, Bagging, Boosting and Random forest)

Prerequisites

Students need to pass before the exam of Analisi Statistica Multivariata

Teaching methods

Lesson in presence

Assessment methods

ORAL EXAM: discussion of a PROJECT WORK

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 of each section:

DATA MINING
Analysis of a quantitative target: descriptive analysis, construction of a robust model (trasformations, diagnostics, model selection, heteroskedasticity, robust inference)

MACHINE LEARNING
Analysis of a binary target (classification)
(Descriptive analysis, preprocessing, propose different classifiers, validation strategies, tuning of models, assessment, choice of best threshold, score of new data)

Web portals for the choice of the dataset:

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

www. kaggle.com

DISCUSSION ORAL EXAM

The outputs of the project work (completed during the period before the oral exam) must be printed and presented/discussed at the oral exam
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.

Textbooks and Reading Materials

Data Mining

Principles of Econometrics (chapters 2, 4 ,6 ,8 9, 12, 13) Carter Hill, William E. Griffiths, Guay C. Lim.

An Introduction to Statistical Learning with Applications in R (Chapter 3 (no section 3.5), Chapter 4, 6, until 6.1, chapter 7) Carter Hill, William E. Griffiths, Guay C. Lim.
http://www-bcf.usc.edu/~gareth/ISL/

Slides

Suggested texts

Principles of Econometrics associate R book https://bookdown.org/ccolonescu/RPoE4/

A Handbook of Statistical Analyses Using R (2nd Edition) Chapters 5,6,7,8,10

Machine Learning

Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R (Chapter 2-3-4-5- 8)
http://www-bcf.usc.edu/~gareth/ISL/

Handouts on moodle

Semester

I semester

Teaching language

ITA

Sustainable Development Goals

QUALITY EDUCATION