UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining e Machine Learning

2627-3-E4102B087

Learning objectives

Data Mining e Machine Learning (15 cfu, 120 hours: 8 hours per CFU)
Material: https://elearning.unimib.it/course/view.php?id=67389

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.

In particular, the following specific objectives are pursued:

1. knowledge and understanding (theory of data mining and machine learning);
2. applying knowledge and understanding (through laboratory lessons and group project work);
3. making judgements (through the project work to be presented at the exam, in-class quizzes, and independently conducted data analysis);
4. Communication skills (through the presentation of the project work during the oral exam);
5. learning skills (through supplementary materials provided on Moodle, which the student can study independently based on the knowledge acquired during the common part of the course).

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
Construction of a robust (estimator) model strating from a base model.

- (1) R and dplyr (overview)
- (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, misure di Influenza, diagnostiche, heteroschedaticità, model selection.
- (6) multilevel models for longitudinal data and hierachcal data (cenni)

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

ONLY lesson in presence
VIDEO-LESSONS (OLD) https://drive.google.com/drive/folders/10zCiGvEVf0xo18GwLU9W1NA6rj-g93aX?usp=sharing
(only with UNIMIB credentials).

Assessment methods

ORAL EXAM: discussion of a PROJECT WORK and theory arguments

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. The analyses of both Project works (one for each module-section) are detaied below.

DATA MINING
Analysis of a quantitative target:

1. construction of a robust model (trasformations, diagnostics, model selection, heteroskedasticity, robust inference)
   or 2) fit a multilevel model (time data or hierarchical data) or a sample seletion models in case of a truncated sample

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

Carter Hill, William E. Griffiths, Guay C. Lim.
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/section 4.1, 4.2, 4.3 , 6.1, 6.2, chapter 7)
https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html

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
https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.ehu.eus/ccwintco/uploads/9/93/A_Handbook_of_Statistical_Analyses_Using_R_Second_Edition.pdf&ved=2ahUKEwjLzbfZq-WGAxWvgv0HHRx7AzAQFnoECBEQAQ&usg=AOvVaw0P4Jf6CnMmRwFth4y5zQsh

Machine Learning

Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R (Chapter 2-4-5-6-8-10.1, 10.2, 12(parte PCA))
https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html

Handouts on moodle

Semester

I semester

Teaching language

ITA

Sustainable Development Goals

QUALITY EDUCATION