UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining

2021-3-E4102B085-E4102B086M

Learning objectives

Data mining and computational statistics (divided in two modules)

Computational statistics

The course aims at introducing complex methotodologies for modelling statistical models both from the theoretical and from the applicative point of view

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

Computational statistics

- **(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**
- **(6) Time series regression**

Data mining

Principles of Data mining, robustness, over fitting and validation. Association rules, Statistical models: linear, discriminant analysis, logistic models, (polytomic and ordinal), Algorithms for the classification: (Naive Bayes, Nearest Neighbour, 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 of each module:

Computational statisitics (sas base or R)

1 applied Complete work with

A) quantitative target

(descriptive analysis, trasformations, diagnostics, model selection, heteroskedasticity checks, strategies to construct a robust model)

and finally a logistic regression with binary target (binarize the previous target) using covariates of interest,

checking collinearity and separation, fit a model

B) analysis with a time series (ARIMA, stazionarity) and regressione with other covariates

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:

Cross section data:

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

www. kaggle.com

Data for time series:

https://bookdown.org/ccolonescu/RPoE4/

https://otexts.com/fpp2/

https://www.econometrics-with-r.org/ittsraf.html#apatadlm

https://www.economodel.com/time-series-analysis

https://online.stat.psu.edu/stat510/

https://data.world/datasets/time-series

data(package = "fpp2") univariate

data(package = "AER") multivariate

data(package = "urca") multivariate

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

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.

Textbooks and Reading Materials

Computazional statistics

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 (chapters 6, 7) Carter Hill, William E. Griffiths, Guay C. Lim.

Chapter 3 (no section 3.5), Chapter 4, 6, until 6.1, 7

Slides

Suggested texts

Principles of Econometrics associate R book https://bookdown.org/ccolonescu/RPoE4/ **(consigliato)**

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

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 cycles I and II

Teaching language

ITA