UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining M

2425-1-F8204B014

Learning objectives

The course aims to provide data analysis and data mining tecniques and to improve predictive modelling skills by using the R software environment for statistical computing.

The course contributes to the achievement of the training objectives in the learning area of: "**Statistics**".

Contents

The detailed program of the course is available at course web page. The main topics are:

- A-B-C: linear models and computational aspects
- Overfitting, bias and variance tradeoff, optimism
- Model selection and penalized methods for linear models (best subset regression, ridge regression, lasso, elastic-net)
- Nonparametric estimation (local linear regression, regression and smoothing splines)
- Additive models (GAM and MARS)

Detailed program

- **A-B-C**
  - Linear models and the modelling process
  - Cholesky factorization
  - Orthogonalization and QR decomposition
  - Iterative methods

    - Generalized linear models
- **Optimism, conflicts, and trade-offs**
    - Polynomial regression
    - Training and test set
    - Bias-variance trade-off, optimism
    - Cross-validation and generalized cross-validation
    - Information criteria (AIC, BIC, etc.)
- **Shrinkage and variable selection**
    - Best subset selection
    - Principal components regression
    - Ridge regression
    - LARS and Lasso
    - Elastic-net
- **Nonparametric regression**
    - Local linear regression
    - Regression and smoothing splines
    - Nonparametric regression: bivariate case
    - The curse of dimensionality
- **Additive models**
    - Generalized Additive Models (GAM)
    - Multivariate Adaptive Regression Splines (MARS)

Prerequisites

Knowledge of the topics (i) linear algebra, (ii) linear models, (iii) generalized linear models (GLMs), (iv) inferential statistics, and (v) probability theory, is required. Moreover, it is required a solid knowledge of the R software.

Knowledge of topics covered in the courses *Probability and Statistics M* and *Advanced Statistics M*, i.e. advanced probability and inferential statistics, is also highly recommended.

Teaching methods

Lessons are held both in classroom and in lab, integrating theoretical principles with practicals aspects of data analysis and programming in R.

The 47 hours of teaching are organized as follows:

- 35 hours of lectures, in person;
- 12 hours of laboratory activities conducted interactively and remotely.

Assessment methods

The exam is made of two parts:

- (20/30) Written examination (open questions): a pen-and-paper exam about the theoretical aspects of the course.

- (10/30) Individual assignment: a data challenge.

The final grade is obtained as the sum of the above scores.

You will be given a prediction task, and you will need to submit your predictions about the assigned case study and produce a report of about 4-5 pages. The data challenge will be announced in the second half of the course. Both parts are mandatory and you need to submit the assignment before attempting the written part. The report expires after one year from the moment the competition has been announced.

Textbooks and Reading Materials

Required

- Azzalini, A. and Scarpa, B. (2011), *Data Analysis and Data Mining*, Oxford University Press.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, Second Edition, Springer.

Optional

- Efron, B. and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press.
- Lewis, Kane, Arnold (2019) *A Computational Approach to Statistical Learning*. Chapman And Hall/Crc.

Additional teaching material will be made available in the course website.

Semester

Second semester

Teaching language

English

Sustainable Development Goals

QUALITY EDUCATION