UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici II

2425-2-F8203B042-F8203B013M

Learning objectives

The course falls within the learning areas of statistical sciences, computer science, and social sciences. The course aims to provide students with preparation regarding analytical and inferential procedures concerning: non-parametric bootstrap, multivariate Gaussian distribution, generalized linear models for count data, and univariate and multivariate Gaussian mixture models, as well as predictive models. The course aims to develop a critical understanding of the model assumptions underlying the theory through empirical applications with real and simulated data.

Knowledge and understanding

Teaching enables the student to:

• Analyze data with advanced univariate and multivariate statistical models developed for both categorical and continuous response variables.
• Develop simulation methods.
• Use the semantics of the R software, also through the RMarkdown environment, to learn a replicable and reproducible research method. The generated documents include code, results, and comments on the code and performed analyses.
• Rigorously interpret the results of the empirical analyses and provide a description also for dissemination purposes to non-academic audience.

Ability to apply knowledge and understanding

The course allows the student to:

• Develop statistical inference using modern bootstrap techniques.
• Estimate, select, and interpret the statistical model-based clustering techniques especially considering finite mixtures of Gaussian distributions for heterogeneous populations.
• Apply theoretical knowledge to the analysis of data collected in various fields, including epidemiology, medicine,

biology, genetics, and public health.
• Conceptualize and estimate through the maximum likelihood method models with latent variables.
• Implement code with the open source software R.

The course enables students to acquire solid theoretical foundations and to develop applications through a "problem-solving" approach. The course pertains to data science, which is now essential for the target job contexts (biostatistics/statistics/demography and related) of graduates in Biostatistics.

Contents

In the first part of the course, the main probability distributions used to simulate realizations from random variables are reviewed. The resampling procedure known as bootstrap is presented to obtain precision measures in a non-parametric context for some estimators of interest.
In the second part of the course, the Expectation-Maximization (EM) algorithm is introduced as a method for imputing missing data using maximum likelihood estimates of the parameters of a generalized linear model. After introducing Gaussian mixture models, the steps of the EM algorithm for the maximum likelihood estimation of the parameters of these models and latent variable models with discrete distribution are described. Theoretical lessons are accompanied by practical exercises conducted with many different data. The course provides skills in using the semantics of the R software, also utilizing the RMarkdown library through the knitr package to integrate code, analysis results, and comments.

Detailed program

The first part of the course deals with simulation methods and linear congruential methods to generate pseudo-random numbers. Graphical tools for testing the series and statistical tests such as Kolmogorov-Smirnov and Chi-Squared tests are illustrated. Simulation of random numbers from specific distributions is considered. Some theoretical features of the exponential, binomial, and Gaussian distributions and convolution of random variables are exposed.
Resampling methods, such as jackknife and bootstrap, are introduced in the second part of the course. The bootstrap is applied for bias adjustment and the estimation of dispersion. Bootstrap confidence intervals based on the percentile method and the bias-corrected accelerated bootstrap method are explained.
Among the optimization methods, the Expectation-Maximization algorithm is considered and explained first as a tool to impute missing values through a generalized linear model and then as a tool to maximize the log-likelihood function for incomplete data problems. Finite mixture models are introduced both for continuous and for categorical data, and a particular focus is given to mixture of Gaussian distributions and latent variable models for categorical data.
Some time is devoted to explaining the theory by imparting the flavor of the empirical applications using data collected from different fields arising in epidemiology, pharmacoepidemiology, medicine, biology, ecology, and environmental sciences. They are developed within the statistical software R, RStudio with the RMarkdown interface. The main R packages used are bootstrap, dplyr, MASS, MultiLCIRT, tscount, mclust e skimr.
The student is encouraged to develop reproducible documents in which he/she critically comments on the code and the analysis results, also through cooperative learning.

Weekly exercises are assigned, and students are encouraged to write reports in which they comment on the code and provide the reader with an explanation of the analysis procedure performed, along with a critical description of the results obtained.
Students are invited to work on the assigned exercises in groups to promote cooperative learning. During the course, the solutions to the assigned exercises are discussed.

Prerequisites

For an easier understanding of the course content, it is recommended to know Probability and Statistical Inference notions. The student should also know the basic semantics of the programming language in the R environment.

Teaching methods

Classroom lectures cover the theoretical aspects of some advanced statistical models, theory is complemented by practical exercises that enable students to learn both the theory and data analysis techniques. Weekly summarizing exercises are assigned, which involve applying the proposed models to real or simulated data. During the course, with the help of R in the RStudio environment and the RMarkdown interface, students learn to create reproducible documents. They are encouraged to tackle application problems with the additional goal of developing cooperative learning. The scheduled hours of traditional teaching are 30, and those of interactive teaching are 24, including lesson concerning exercises.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. The examination is written with open questions and an optional oral part is possible; there are no intermediate tests. The written exam lasts around an hour and a half and takes place in the computer lab. During the exam, open theory questions must be answered, and exercises must be solved based on the topics covered during the course. The theory questions assess the understanding of the theoretical concepts taught during the course. The empirical analyses are conducted using the R environment, Rstudio, and RMarkdown allowing verification of the ability to understand the problem and resolve it by applying advanced statistical models to real or simulated data. Students must also elaborate on reports in which the procedure is described, and the results are illustrated. The examination is open book, and students can consult all the material as well as the R code provided during the lectures. The student passes the test with a mark of at least 18/30.

Textbooks and Reading Materials

The teaching material consists mainly of handouts prepared by the teacher. These cover theory, applications, exercise and solutions developed with R software. All the files are available on the course page of the university's e-learning platform. In addition, the teacher publishes the following material at the end of each lesson: slides, R and SAS code, exercises, datasets, and solutions to some of the exercises. Previous exam texts are also published on the same page.

The main references are listed in the bibliography of the handouts, some of which are as follows and are available in the university library, also in ebook format:

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). Latent Markov Models for longitudinal data, Chapman and Hall/CRC, Boca Raton.

Bishop, Y. M., Fienberg, S. E., Holland, P. W. (2007). Discrete multivariate analysis: theory and practice. Springer Science & Business Media, New York.

Blitzstein, J. K., Hwang, J. (2014). Introduction to probability, Chapman & Hall/CRC.

Gentle, J. E., Hardle W., Mori Y. (2004). Handbook of computational statistics. Springer-Berlin.

Lange, K. (2010). Numerical analysis for statisticians, 2nd Edition, Springer, New York.

Pennoni, F. (2023). Dispensa di Modelli Statistici II, Teoria e Applicazioni con R. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Semester

Semester I, cycle I, October-November 2024

Teaching language

The course is provided in Italian. Erasmus students can use the handouts material in English and ask the teacher to carry out the exam in English.

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING | REDUCED INEQUALITIES | CLIMATE ACTION