SYLLABUS DEL CORSO

Statistical Modelling

2425-1-FDS01Q040

Aims

The course falls within the learning areas of statistics, computer science and social sciences and it aims to provide students with methodological and applied background on advanced statistical models: multiple linear regression and some extensions, some generalized linear models and some model-based approaches to cluster analysis concerning univariate and multivariate finite mixture models of Gaussian distributions, along with predictive models.

Knowledge and understanding

The student is introduced to advanced statistical models for analysing data with different types of response variables. The relevant assumptions underlying the theory are also illustrated by considering the maximum likelihood and least squares estimation methods for model parameters. Data analysis is conducted using R software and the RMarkdown environment, which allows for the creation of reproducible documents containing code, results and comments. Applications cover real and simulated data from various fields such as economics, finance, and social sciences. The student is also encouraged to provide a critical evaluation of the results obtained from the empirical analyses. The course enables students to acquire solid elements of theory and applications. It concerns data science, and this knowledge is essential nowadays in every working environment, and it is compulsory for the next course of student studies.

Ability to apply knowledge and understanding

The course provides skills in using the semantics of the open-source software R for the descriptive analysis of multivariate data and parameter estimation of univariate and multivariate models. Through R and RStudio, students learn how to systematically set up statistical reasoning by analyzing data and writing reports that illustrate code, analysis and results. Theory is complemented by practical applications also developed during tutoring lectures. The course enables students to acquire a solid theoretical foundation and the ability to apply modern statistical method for data analysis, as well as developing the ability to conduct reproducible and replicable research.

Contents

In the first part of the course, following a brief introduction to the conceptual framework of statistical inference and causality issues, the resampling procedure known as bootstrap is illustrated to obtain measures of accuracy for estimators of interest. Next, the multiple linear regression model is presented along with its assumptions. The methods of ordinary least squares and maximum likelihood estimation are introduced, as well as their statistical properties. Measures of fit, regression diagnostics and prediction are also covered. Generalised linear models are discussed, including the multiple logistic regression and multinomial logit models. The expectation-maximisation algorithm is introduced as a tool for maximum likelihood estimation of classification model parameters. Probabilistic classification models for supervised learning are also discussed. The course provides skills in the use of R software semantics, utilizing the RMarkdown libraries via the knitr package to integrate code, analysis of the results of applications using real and simulated data, and add comments on the code and the obtained results.

Detailed program

The course starts with an introduction to the picture of statistical inference and some related concepts in causal inference.

- The first part of the course introduces the resampling method known as bootstrap for determining the standard error as a measure of accuracy. This method is applied to various estimators using relevant data deriving from different sources such as psychology, environment and many other fields.

- The second part of the course covers the multiple linear regression model, considering least-squares and maximum likelihood as estimation methods. The properties of the least-squares estimators as well as of the maximum likelihood estimators are discussed on the basis of the model assumptions.

- During the course, the student's knowledge on univariate distributions is extended to include the bivariate and multivariate Gaussian distributions. Random realizations are drawn from these distributions, which are also represented graphically with contour lines.

- Various diagnostic tools for model evaluation based on regression residuals are considered, with particular emphasis related to the outliers, influential points, and leverage points. The problem of selecting the most relevant explanatory variables using criteria such as the Akaike information criterion is addressed. Additionally, students learn how to evaluate model predictions.

- Generalised linear models for the analysis of categorical response variables with two or more categories are introduced. The multiple logistic regression model and the multinomial model are illustrated, with particular emphasis on the interpretation of regression coefficients.

- Gaussian mixture models for supervised learning are introduced in order to provide hits to students into statistical pattern recognition (discriminant analysis) through probabilistic classification following a mixture-based approach. Results of the generative models based on estimated posterior probabilities are evaluated using training and validation sets. Diagnostic tools based on classification error, Brier score, receiving operating characteristic curve (ROC), as well as area under the curve (AUC), are discussed and presented in the applied contexts.

Some time is dedicated to explaining theory by providing empirical applications using data from different fields such as economics, finance, biology, ecology, and environmental sciences. They are developed within the statistical software R, using the RStudio environment and many different libraries along with the RMarkdown interface and the knitr library. This approach aims to familiarize students with the principles of reproducible research. Students are expected to write reproducible reports where they critically comment on the code and the results of the empirical analyses. Cooperative learning is encouraged through assigned homeworks.

Prerequisites

For an easier understanding of the course content, it is recommended to know the contents of the course Foundations of Probability and Statistics. The course assumes prior knowledge of the following topics: probability of an event, probability distribution function, density, cumulative distribution functions, the law of total probability, independence of events, Bayes theorem, expectation and variance of a random variable, standardization and percentiles of a random variable, continuous and discrete random variables such as Bernoulli, binomial, Poisson, geometric, uniform, exponential, Gaussian, Student-t, and chi-squared, graphs and numerical measures to describe data, statistical inference, maximum likelihood inference and basic knowledge of multivariate data analysis and linear algebra. Students should also know the basic semantics of the programming language in the R environment.

Teaching form

All the lectures are delivered in presence. They cover theoretical aspects and are complemented by practical exercises that enable students to learn theory and apply models to analyze real and simulated data. Lessons take place in the computer lab. Weekly summarizing exercises are assigned as homework to reinforce the learning of the theory and its applications.

During the course, with the help of R in the RStudio environment and the RMarkdown interface, students also learn to create reproducible documents. They are encouraged to tackle application problem with the additional goal of developing cooperative learning. Tutoring sessions are also scheduled to help students develop exercises and compare solutions. The scheduled hours of traditional teaching are 30, and those of interactive teaching are 27, including tutoring lessons.

Textbook and teaching resource

The teaching material consists mainly of handouts prepared by the teacher, covering both theory topics and the applications developed with R software. All the files are available on the course's page on the university's e-learning platform. Additionally, the teacher publishes the slides, calculation programs, and datasets at the end of each lesson. Weekly exercises are assigned, and some solutions are provided and discussed. Examples of examination texts are also published on the same page.
The primary references will be listed in the bibliography of the handouts. Among others, the following are noted, and some of these are also available in the library and as eBooks:

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.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). Model-based clustering and classification for data science: With applications in R. Cambridge University Press.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. D. (2021). Regression: Models, methods and applications. Springer Berlin, Heidelberg.

Faraway, J. J. (2014). Extending the Linear models with R, 2nd Edition, Chapman & Hall, CRC Press. Hastie, T., D. and Tibshirani, R. (2013). An introduction to statistical learning, New York, Springer.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, 2nd Edition. Chapman and Hall/CRC, London.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ Xie, Y., Dervieux, C. and Riederer E. (2020). R Markdown Cookbook. Chapman & Hall, CRC.

Semester

Semester II, March-May 2025

Assessment method

The following methods of verifying learning apply to both students attending and non-attending lectures held in the lab. The examination consists of a written test with open questions, and an optional oral exam. The written exam has a maximum total duration of an hour and a half and takes place in the computer lab. During the examination, students are required to answer open theory questions and solve exercises and practical exercises with data as those assigned weekly 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. These analyses allow students to demonstrate their ability to understand and solve problems by applying advanced statistical models to real or simulated data, and to produce reproducible reports that describe the code, and illustrate the results. During the examination, the use of study materials and R code implemented during the course is permitted. Each question will be marked approximately 3 points. To pass the test, a student must achieve a mark of at least 18 out of 30.

Office hours

Weekly, annunced on the elearning page during the course.

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING | GENDER EQUALITY | SUSTAINABLE CITIES AND COMMUNITIES | CLIMATE ACTION