UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Bayesian Inference

2122-2-F8203B042-F8203B042M

Learning objectives

The course provides knowledge on the basic and advanced statistical principles under the Bayesian paradigm. The methods are illustrated according to an integrated approach with classical statistical inference.

Knowledge and understanding

The course allows the student to learn:

- the Bayes' rule and the use of probability to update beliefs from the data;

- Bayesian inferential methods: prior,

computation of the likelihood and posterior distributions;

- Monte Carlo methods to simulate the posterior distributions;

- calculus of the predictive distribution for features predictions and model checking;

- Markov Chain Monte Carlo algorithms: Metropolis-Hastings and Gibbs sampler;

- Bayesian inference and prediction for the multiple linear and logistic regression models;

- latent Markov models for the analysis of longitudinal data.

Ability to apply knowledge and understanding

The course allows the student:

- to apply the Bayesian methodology by using data available in the field of biostatistics, epidemiology, medicine, biology, environmental impact, genetics, and public health;

- to apply suitable statistical models when repeated observations for the same unit are available over time;

- to apply model-based classification methods;

- to implement suitable R and SAS code to carry out the analyses;

- to provide reproducible documents with code, results, and comments.

The theory is supported with practical applications based on real and simulated data. R is used within the Rstudio interface, and Rmarkdown, SAS is also proposed so that the student gets a deep knowledge of two program languages.

This course provides the main concepts of Bayesian inference, an essential statistical method in the theoretical and data analysis fields for the job contexts (biostatistics/statistics/demography and related) of students in the Biostatistics degree program. The course is found to be essential for the subsequent courses.

Contents

Introduction to Bayesian inference and Bayes' rule. Methods of model specification and a priori distributions.

Determination of the posterior distribution by exact methods, conjugate families: Gaussian, Poisson-Gamma, Beta-Binomial, Multinomial-Dirichelet.

Introduction to Bayesian non-parametric inference.

Methods to summarize the posterior distribution: credibility intervals and intervals with the highest posterior density.

Introduction to stochastic Markov processes, random walk.

Markov chain models for longitudinal data.

Introduction to the latent Markov models for panel data with covariates.

Introduction to the Markov Chain Monte Carlo Methods: Metropolis-Hastings algorithm and Gibbs sampler.

R environment and RStudio interface with the RMarkdown to integrate code and output within the knitr library. The main R libraries are the following: probBayes, learnBayes, LMest. SAS software with proc MCMC.

Detailed program

The Bayesian paradigm is introduced and compared with the frequentist approach and the Bayes ‘rule, and the total probability rule. A short introduction to the Bayesian non-parametric methods is provided, the notion of exchangeability and De Finetti’s theorem are explained. The Bayes’billiard example is presented to introduce the Beta-Binomial model. Choice and specification of the prior distribution. Conjugate families: Gaussian, Poisson-Gamma, Beta-Binomial, and Multinomial-Dirichlet distributions and non-informative priors. Methods to draw

conclusions from the posterior distribution: Bayesian interval estimation, credible intervals, and intervals with the highest posterior density. The prediction context is also considered along with the empirical Bayes estimation. The theory is supported by several examples of the application of Bayesian models in biostatistics through real and simulated data concerning epidemiology, drug epidemiology, medicine and biology, and ecology and environmental sciences.

An introduction to the stochastic processes within the Markov random field is proposed. Properties and features of the Markov chains are illustrated and explained with the use of simulations. The random walk process is also described.

Markov chain models for longitudinal data are explained, and the Latent Markov models for panel data with covariates are introduced from a theoretical and applied perspective.

Markov Chain Monte Carlo (MCMC) algorithms are provided with a focus on Metropolis-Hastings and Gibbs sampling algorithms. Diagnostic evaluations of the convergence are considered.

Some time is devoted to explaining the theory by imparting the flavor of the applications on real data. The examples are developed within the statistical environment R, RStudio, RMarkdown to make reproducible documents. The SAS software is proposed to perform the analyses to estimate Bayesian linear and logistic models with PROC MCMC.

Prerequisites

The student is encouraged to know the content of the following courses: Statistics, Probability, and Statistical Inference and Statistical Models II.

Teaching methods

The lectures are held in the lab since the theoretical part is placed side-by-side with the computer's applications using R and SAS software. Many practical examples based on real and simulated data referred to different contexts are proposed to the students to be solved with R through the RMarkdown interface and SAS software. The student is also encouraged to develop cooperative learning to interact with each other and finalize the required steps of the analysis. Exercises are carried out to report in a written form the results by adding critical comments and create reproducible documents.

During the Covid-19 emergency period, classes will be held remotely (videotaped lectures) with periodic meetings in videoconference via Webex platform and/or in-person according to the schedules provided by the university, and that will be announced on the course page.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. The examination is in written form with optional oral; there are no intermediate tests. The written exam has a maximum total duration of two hours and takes place in the computer lab. During the test, it is necessary to solve the exercises applied in the light of the theoretical arguments developed during the course and answer some

theory questions. The analyses are conducted using the R environment, Rstudio, RMarkdown, and SAS. The exercises allow verifying the ability to understand the problem and its resolution by applying advanced statistical models to real or simulated data and the elaboration of reports in which the procedure is described, and the results are illustrated. In addition, the theory questions allow verifying the learning of the theoretical concepts taught during the course.

During the emergency period due to Covid-19 depending on the university arrangements will take place in the computer lab or via video conferencing via the Webex platform.

Textbooks and Reading Materials

The teaching material consists mainly of handouts prepared by the teacher. They cover both the theory topics and the applications developed with R or SAS software. All the files are available on the page of the e-learning platform of the university dedicated to the course. In addition, the teacher publishes at the end of each lesson: the slides, the calculation programs, the exercises, the datasets, and the solutions of the exercises. On the same page are also published some previous exam texts.

During the Covid-19 emergency period, video recordings of lectures are also posted on the course page.

The primary reference texts are listed in the bibliography of the handouts; among others, the following are noted. Some of these also available in ebook the following:

Albert, J. (2009). *Bayesian computation with R*. Springer Science & Business Media.

Albet, J., Hu, J. (2019). *Probability and Bayesian modeling*. Chapman and Hall/CRC.

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

Migon, H. S., Gamerman, D., Louzada, F. (2014). *Statistical inference: an integrated approach*. Chapman & Hall.

Pennoni, F. (2021). *Dispensa di Inferenza Bayesiana -parte di teoria e applicazioni con R e SAS*. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Robert, C., Casella, G. (2004). *Monte Carlo Statistical Methods* (Second edition). Springer–Verlag, New York.

Dipak, D. K., Ghosh, S. K., Mallick, B. K. (2000). *Generalized linear models: A Bayesian perspective*.CRC press.

SAS/STAT PROC MCMC, *User's guide*, SAS Institute, 2012.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Semester

Semester I, cycle II, November 2020-Janusry 2021

Teaching language

The course is delivered in Italian. Erasmus students can use the didactic material in English and ask the teacher for the exam in English.