UNIVERSITÀ CATTOLICA del Sacro Cuore

Data analysis for insurance

I MOD.: PROF. DIEGO ZAPPA; II MOD.: PROF. GABRIELE CANTALUPPI

Course aims and intended learning outcomes

Introduction to linear and not linear models, basic methods of statistical learning and categorical data analysis with a special emphasis on generalized linear models. Most of lectures will be provided in the computer lab to offer an introduction to the R language and whereas necessary to other ad hoc statistical software.

On successful completion of the course participants are expected to possess:

1. knowledge of concepts, terms and methods of the most used statistical learning techniques and grasp of their strengths and weaknesses (DD1- Knowledge and understanding);
2. ability to correctly apply statistical learning methods to real insurance, economics and management problems (DD2- Applying knowledge and understanding);
3. quantitative thinking addressed to make independent judgements, driven by application of statistical learning methods (DD3- Making judgements);
4. ability to present statistical learning arguments and the conclusions from them, by means of the extraction of qualitative information from quantitative data, with clarity and accuracy and in forms that are suitable for the audiences being addressed, both orally and in writing (DD4-Communication);
5. mastery of statistical learning methods, rigorous reasoning and data-driven decision-making, useful for quantitative analyses in other courses of the curriculum, as well as for analyses required in careers in insurance and in all fields involving management of data (DD5- Lifelong learning skills).

Course content

MODULE I: *Prerequisites, Multiple Linear and Nonlinear Regression. Principal component analysis.*

1. Some remarks from mathematical statistics useful for the master degree in Actuarial sciences.
2. Some remarks on matrix algebra. Singular Value decomposition. Covariance and correlation matrices. The Bivariate and the multivariate normal distribution. Cochran's theorems. Examples.
3. Linear and Nonlinear models. The Gauss-Newton algorithm. Identification. Inference. Residuals. Examples.
4. Principal component analysis. Hotelling's and Pearson's method. Relationships among components, variables, scores. Geometric properties of principal components. Examples.
5. From linear Gaussian models to GLMs. The Exponential Dispersion Family Distributions for a GLM. Likelihood. Fitting. Inference.
6. GLMs for count, continuous and categorical responses.
7. Logistic regression for binary response data.

MODULE II: *Statistical learning*

1. Introduction to Statistical Learning. Supervised and unsupervised learning. Assessing model accuracy.
2. Resampling methods. Cross-validation and bootstrap.
3. Variable selection in regression problems. Best subset selection and stepwise selection.
4. Shrinkage methods. The ridge regression and the lasso.
5. Cluster analysis. Hierarchical and non-hierarchical methods. The K-Nearest Neighbors algorithm.
6. Tree based methods. Regression trees and classification trees.

Reading list

Lecturer’s notes / slides

UNIVERSITÀ CATTOLICA del Sacro Cuore

I module

L. FAHRMEIR-TH. KNEIB-S. LANG-B. MARX, *Regression. Models, Methods and Applications,* Springer, New York, 2013.

A. AGRESTI, *An Introduction to Categorical Data Analysis,* John Wiley, New York, 2018.

II module

G. JAMES-D. WITTEN-T. HASTIE-R. TIBSHIRANI, *An Introduction to Statistical Learning,* Springer, New York, 2017, http://www-bcf.usc.edu/~gareth/ISL.

L. FAHRMEIR-TH. KNEIB-S. LANG-B. MARX, *Regression. Models, Methods and Applications,* Springer, New York, 2013.

(Details about which part of the books are suggested for reading will be given at the beginning of each module)

Teaching method

Lectures with examples from actuarial, business, economic, financial, health case studies.

Assessment method and criteria

Module I: for those attending regularly the lectures, the exam will consist of a case study to be solved in the PC lab (time 150minutes). For those that cannot attend regularly the lectures, written exam with questions on the methods presented during the course (Time: 45 minutes).

Module II: Written exam. Time: 45 minutes.

The final mark will be the weighted average of the marks in the 2 modules: I module=60%, II module=40%. If the final score does not correspont to an integer, the mark will be rounded up to the nearest larger integer.

Aim of the exam is to assess reasoning analytic abilities on the course subjects. Language properties and communication abilities are also assessed.

Notes and prerequisites

Students enrolling in this course should have a basic understanding of mathematical and statistical techniques at the level of the bachelor degree (undergraduate programme) in economic studies (see for an example the profile “Quantitative methods for finance and insurance” taught at the faculty of “Banking, Finance and Insurance Sciences”).