Universidad Zaragoza

Academic Year/course: 2022/23

62236 - Advanced statistical data analysis

Syllabus Information

Academic Year: 2022/23
Subject: 62236 - Advanced statistical data analysis
Faculty / School: 110 - Escuela de Ingeniería y Arquitectura
Degree: 534 - Master's Degree in Informatics Engineering
ECTS: 3.0
Year: 2
Semester: First semester
Subject Type: Optional
Module:

1. General information

2. Learning goals

3. Assessment (1st and 2nd call)

4. Methodology, learning tasks, syllabus and resources

4.1. Methodological overview

The methodology followed in this course is oriented towards achievement of the learning objectives. It is based on an active methodology that promotes student participation. A wide range of teaching and learning tasks are implemented, such as

- The general contents of the course are presented in sessions where the formal explanation is complemented with appropriate examples.
- Classes in the computer lab deal with both data analysis and modelling of real tasks. In these seasons the students learn to use statistical free software R.
- Each student develops an individual task concerning the use of statistical procedures in big data cases. Students can choose data-bases they are particularly interested in or, alternatively, data-bases provided by the instructor. In both cases a written report is mandatory.

4.2. Learning tasks

The course (3 ECTS: 75 hours) includes the following learning tasks:

- **Classroom sessions** (30 hours organized in two-hour sessions per week). These sessions involve theoretical aspects, problem sets and data analysis. Regular working sessions take place in the computer lab. In these sessions real situations that promote interest in a wide range of statistical techniques are presented. The associated concepts and statistical procedures are shown from a practical viewpoint. The students are encouraged to model and to solve real problems by means of free software. The R language is used and, in this regard, functions, standard libraries available in R Project are introduced to address different techniques.
- **Project** (20 hours). Each student has to elaborate a project for the statistical analysis of a collection of data, with high-dimension, using appropriate techniques to draw conclusions.
- **Autonomous work and study** (20 hours). Study of general principles or ideas and devoted to practical tasks.
- **Tutorials** (5 hours).

4.3. Syllabus

The course will address the following topics:

1. Introduction
   - Statistical learning.
   - Exploratory data analysis.
   - Sampling and statistical inference: point and interval estimation, hypothesis testing.
   - Likelihood: Estimation by maximum likelihood, likelihood ratio test.
   - Statistical decision theory. Bayesian methods.
   - The EM algorithm. The MCMC method.
   - Statistical simulation
3. Recognition of explicit relationships: Regression Models
   - Simple linear regression, review and validation of the model Box-Cox transformation, prediction.
   - General linear model, covariates and factor analysis of variance.
   - Automatic modeling procedures: best subset, stepwise.
   - Validation, cross validation, bootstrap methods.
   - Regression with high dimensionality.
   - Models with non-Gaussian response: GLM and GAM.
5. Supervised pattern recognition: Logistic Regression.
   - Binary logistic regression models.
   - Multinomial logistic regression models.
   - Crosstabulation, log-linear models.
7. Unsupervised pattern recognition.
   - Cluster analysis, k-means method.
   - Hierarchical cluster.

4.4. Course planning and calendar

The course is organized in 2 hours of class per week.

Further information concerning the timetable, classroom, office hours, assessment dates and other details regarding this course, will be provided on the first day of class or please refer to the EINA website and the course website.

4.5. Bibliography and recommended resources

http://psfunizar10.unizar.es/br13/egAsignaturas.php?codigo=62236