

## 62236 - Advanced statistical data analysis

### Syllabus Information

**Academic year:** 2024/25

**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

The subject aims to train the student in the construction of stochastic models that explain the relationships between variables or individuals in observational studies. This requires knowledge of tools in maximum likelihood and Bayesian estimation, EM algorithm, MCMC methods and resampling.

Statistical techniques are introduced and students learn to select the most suitable one to obtain solutions in the analysis of large databases. Using regression techniques, predictions of the response and bounds of its error are calculated in supervised learning situations. For situations where there is no response variable, but a set of variables that represent reality, unsupervised learning methods recognize patterns of variables and cases, with the aim of characterizing them and reducing the dimension.

### 2. Learning results

1. To interpret observational or experimental data, extract the information they contain, identify the relationships between them and evaluate hypotheses in the presence of uncertainty and variability, interpreting their results appropriately.
2. To understand the estimation methods, by maximum likelihood and Bayesian, know the tools and algorithms for estimation in large databases.
3. To apply statistical procedures for the construction and validation of empirical models that express the relationship between a response variable and other variables whose value can be known.
4. To use the most relevant multivariate analysis techniques that help explain the relationships between data and identify patterns when there is no response variable.
5. To apply learning techniques, data mining and process mining for knowledge extraction in environments that handle large amounts of data through the Web.
6. To apply mathematical techniques for the analysis of large amounts of data on the Web.

### 3. Syllabus

- Introduction
  - Exploratory data analysis and statistical learning.
  - Sampling and statistical inference: point and interval estimation, hypothesis testing.
  - Likelihood: Maximum likelihood estimation, likelihood ratio test.
  - Statistical decision theory. Bayesian methods.
  - EM algorithm and MCMC method.
  - Statistical simulation.
- Recognition of explicit relationships: regression models
  - General linear model, covariates and factors, analysis of variance.
  - Automatic model construction procedures: best subset, stepwise.
  - Validation, cross validation, bootstrap methods.
  - Models with non-Gaussian response: GLM and GAM.
- Assisted pattern recognition
  - Binary and multinomial logistic regression models.
  - Multinomial logistic regression models.
  - Log-linear models.
- Unsupervised pattern recognition
  - Cluster analysis, k-means.
  - Hierarchical cluster.

#### 4. Academic activities

The subject consists of 3 ECTS which represent an estimated student's dedication of 75 hours (35 hours of teacher's explanations and 40 non-face-to face hours) distributed as follows:

- Approximately 30 hours of educational activities with the teacher (master classes including professional seminars, problem solving, case studies and laboratory practices).
- 20 hours of group work.
- 20 hours of effective work and individual study.
- 5 hours dedicated to different evaluation tests.

#### 5. Assessment system

The student must demonstrate achievement of the intended learning results through the following assessment activities:

**Monitoring of the work carried out in the practical sessions.** Work developed in the practical sessions related to data analysis [20%]. Learning results: 1, 2, 3 and 4.

**Directed work.** An individual or group project in which the knowledge and skills acquired in the subject will be put into practice. In the evaluation of the tutored work proposed throughout the four-month-period, both the report presented, and the suitability and originality of the proposed solution will be taken into account. [70%]. Learning results: 1 to 6.

**Oral presentation and debate** on the development and results of the academic work (10%). Learning results: 1 to 6.

The student who does not opt for the evaluation procedure described above, does not pass these tests during the teaching period or who would like to improve their grade will be entitled to a global test that will be scheduled within the exam period corresponding to the first or second call.

#### 6. Sustainable Development Goals

7 - Affordable and Clean Energy

15 - Life on Land