IF.2301 – Data Science and Processing

General information

Module Title: Data Science and Processing
Module ID: IF.2301
Head of the module: Patricia CONDE-CESPEDES / Hélène URIEN
ECTS: 4
Average amount of work per student: 75 hours, including 48 hours supervised
Teamwork: a data science project to be done in groups of 2 or 3 people.
Keywords: probability, statistics and data science

Presentation

Nowadays, we can easily access a huge amount of data. Data science is a field of study in artificial intelligence that combines tools from computer science, probability, and statistics to extract meaningful insights from raw data. Indeed, Probability and Statistics are a keystone for building models in Data Science. Probability theory is a branch of mathematics that studies the degree of uncertainty in a random process described by random variables. Whereas statistics consists of using data sampling, mainly for two main purposes: to describe certain phenomena (descriptive statistics) and to infer properties about the probability distribution of the random variables describing the population of the sample (statistical inference). Most statistical methods depend on probability theory. In data science, probability and statistics are mainly used for the estimation and prediction of a random phenomenon.

Educational objectives

Link with the Isep competency framework

The knowledge and skills developed in this module are in the field of probability and statistics, in contexts of use relating to data analysis, signal processing, and learning the scientific method. Most of the application examples will be contextualized.

The main objective of this course is to provide students with the foundations of probability theory and statistical analysis commonly used in data science problems. The course is at the same time theoretical and practical. By the end the students will analyze real datasets using recent methods with languages Python and R.

Prerequisite

- *Notions of probability, notions of linear algebra*

Content/Program

- Probabilities
  - Notion of event and probability
  - Conditional Probabilities & Independence
  - Real random variable
  - Typical values of a real random variable
  - Characteristic function of a real random variable
  - Transformation of a real random variable
  - Two-dimensional real random variables
  - Expectancy, characteristic function and moments for 2 random variables
  - Concept of convergence, the law of large numbers and the central limit theorem
- Statistics
  - Descriptive statistics

    - Statistical Theory of Estimation
    - Hypothesis testing
- Data Science
    - Introduction to Data Science
    - Linear Algebra Reminders for Data Science
    - Single and multiple linear regression
    - Principal Component Analysis and Applications

Tools used

- R (for the Statistics part)
- Python (for the data science part)

Subsequent mobilizations at ISEP

Achieving good results in this module is a must to join the Data Intelligence track, and recommended for the Digital and Healthcare, and Embedded Systems tracks.

Pedagogical methods

Learning methods

This module is based on a problem-based approach, through the systematic use of contextualized problems. Each component of the theoretical course is followed/accompanied by tutorials and practical work on machines with R software and Python (for data science).

Course of the module (Hours of face-to-face teaching):

- Classes (12 sessions of 1h30 and 1 session on a machine of 3h)
- Practical work (12 sessions of 2 hours)
- Tutorials on a machine (1 session of 3 hours)

Evaluation methods

- 1 probability exam around the middle of the semester.
- 1 review of statistics towards the end of the semester.
- 1 data science project to be done in pairs or in 3 people.
- Class participation is counted for additional points.

Language of work

- English.

Bibliography, Webography, Other sources

- MIT-OPEN-Courseware: "Probabilities and applied statistics"
- Gilbert Saporta (2011) Probabilities, data analysis and statistics. 3rd edition.
- Handout of the course.