IF.1205 – Data Science Fundamentals

General information

Module Title: Data Science Fundamentals
Module ID: IF.1105 / IF.1205
Person in charge: 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

Specialized skills

- Solving Constraints, Multidisciplinary Scientific and Technical Problems in the Field of ICT
    - Problem modeling and formal treatment
        - Implementation of a problem decomposition heuristic.
        - Precision of resources useful for the resolution.
        - Planning the resolution and successive refinements.
        - Search for suitable solutions.
- Design software or hardware technological objects with safe and standardized operation
    - Ensuring the quality and safety of a system (availability, reliability, maintainability, security, confidentiality – integrity):
        - Analyze the way the system works as well as its malfunctions.
        - Model the mode of operation and failures of a system and its components.
        - Apply appropriate quantitative approaches that provide indicators or measures that characterize operational safety and manage risks.
- Understanding research methods and how to apply them in ICT
    - Experiment with methods specific to the subject at hand.
    - Produce results that add value to the intelligence of the subject at hand.

Transversal skills

- Acting as a dynamic and effective player in a group
    - Work in a team, in a network, and in a culturally diverse environment.
    - Lead a team, motivate it and help it evolve.
    - Managing conflict, diversity and differences.
    - To be a force of proposal.

- Acting as a good communicator in a scientific and technical environment open to the international world
    - Listen and be heard.
    - Conduct a dialogue, argue and convince.
    - Communicate in multiple languages.
    - Document in an efficient and easily usable manner, regardless of the intended audience, the activities performed or the products produced.
    - Have a communicative approach adapted to the situations envisaged, transparent and effective for its employees.
- Acting as a responsible professional concerned with strategic issues
    - Demonstrate rigour, act with professional probity and intellectual honesty.
    - Demonstrate autonomy.
    - Demonstrate critical thinking.
    - To be concerned with the dissemination of technical and scientific knowledge.

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)

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

- Gilbert Saporta (2011) Probabilities, data analysis and statistics. 3rd edition.
- MIT-OPEN-Courseware: "Probabilities and applied statistics"
- Poly for both parts of the course with some useful references.