IF.2301 Data science fundamentals

Subject code : IF.2301
Person in charge : Patricia Conde-Cespedes
ECTS : 3
Organization: = Lectures (12 sessions of 1h30 and 1 session using a machine of 3h) + tutorial courses (12 sessions of 2h + 1 session of 3h).

Evaluation : 2 exams + short project in data science+ participation during the sessions.

Prerequisite : combinatorics, notions of probabilities, linear algebra.

Context

We live in an era where we can easily have access to huge amount of data. Data science is a field of study that combines tools coming from computer science, probability and statistics to extract meaningful insights from raw. Indeed, Probability and Statistics constitute a keystone to build models in Data Science. The Probability theory is a branch of mathematics that studies the degree of uncertainty in a random process described by random variables. Whereas statistics consists in using data sampling, mainly for two main purposes: describing some phenomena (descriptive statistics) and inferring properties about the probability distribution of the random variables describing the population of the sample (Statistical inference). Most Statistical methods depend on the theory of probability. In data science, probability and statistics are mainly used for estimation and predictions of a random phenomenon.

Objective

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.

Content

The subject is divided in 3 main parts.

Part I : Probabilities :

- Introduction to theory of probability
- Conditional probabilities and independence
- Random variables
- Moments of a real-valued random variable
- Function characteristic of a real-valued random variable
- Transformation of a real-valued random variable
- Real-valued bi-dimensional random variables
- Characteristic function and moments for a couple of random variables
- Concept of convergence, law of large numbers and central limit theorem

Part II Statistics:

- Descriptive statistics
- Point and interval estimation

- Hypothesis testing

Part III Data Science

- Introduction to data science
- The basics of linear algebra for data science
- Principal component analysis (PCA) and applications
- Linear regression

For the 3 components, probability, statistics and data science, the theoretical course is accompanied by tutorials and practical courses with R software and Python. So that the students can assimilate the theoretical knowledge experimentally and with practical examples from everyday life.

Pedagogical Approach

14 sessions (one per week) composed by :

- 12 sessions of 1,5h lecture and 2h of tutorial courses
- 1 session of 3h lecture
- 1 session of 3h tutorial course

5 sessions are devoted to Part I, 1 session for the transition between parts I and II, 4 sessions for part II and 3 sessions for part III

4 sessions include practical courses in Python and R where students will have the opportunity to analyse real datasets.

Language

All the lectures, tutorial courses and practical courses are given in English. All the material is given in English.

Evaluation

- 1 exam of probabilities (Part I) by the middle of the term
- 1 exam of statistics (Part II) by the end of the term.
- Short project in data science with a real dataset (Part III)
- Participation during the sessions.

Bibliography

- MIT-OPEN-Courseware: « Probabilities and applied statistics »
- Gilbert Saporta (2011) Probabilités, analyse des données et statistique. 3ème édition.
- Handout of the course.