Course guide
205099 - 205099 - Advanced Engineering Data Analysis

Last modified: 02/04/2024

Unit in charge: Terrassa School of Industrial, Aerospace and Audiovisual Engineering
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.

Degree: MASTER'S DEGREE IN INDUSTRIAL ENGINEERING (Syllabus 2013). (Optional subject).
MASTER'S DEGREE IN AERONAUTICAL ENGINEERING (Syllabus 2014). (Optional subject).
MASTER'S DEGREE IN SPACE AND AERONAUTICAL ENGINEERING (Syllabus 2016). (Optional subject).

Academic year: 2024 **ECTS Credits:** 3.0 **Languages:** English

LECTURER

Coordinating lecturer: Fernández Martínez, Daniel

Others:

TEACHING METHODOLOGY

The course adopts a hands-on approach. However, the learning process is developed through a combination of theoretical explanation and its application to a real case.
The theoretical sessions will be imparted with the aid of PowerPoint presentations and will include application to real data sets.
The lab classes will be imparted with R and RStudio. The implementation of practices fosters generic skills related to teamwork and presentation of results and serves to integrate different knowledge of the subject.

LEARNING OBJECTIVES OF THE SUBJECT

The main objective of this course is to provide the students with the knowledge of advanced data analysis showing their most basic methodologies and techniques.
The course adopts a hands-on approach introducing basic programming in R and providing guidance on data analysis strategies for Engineering data sets.
The students will be able to think critically about data, use graphical and numerical summaries, apply standard statistical multivariate methods, and draw contextualized and critical conclusions from such analyses.

STUDY LOAD

Type	Hours	Percentage
Hours large group	15,0	20.00
Self study	48,0	64.00
Hours small group	12,0	16.00

Total learning time: 75 h

CONTENTS

Module 1: Introduction, R and RStudio, Basic Statistics & Exploratory Data Analysis

Description:
Introduction to the course.
Introduction to the basics of the R language and the statistical software RStudio.
Learning how to run basic statistics with this platform via the application of basic exploratory data analysis and pre-processing of the data (i.e., detecting missing values and outliers, and checking variable distribution)

Full-or-part-time: 12h
Theory classes: 5h
Self study : 7h

Module 2: Principal Components Analysis

Description:
Introducing the basics of supervised and unsupervised learning focusing on dimension reduction, classification, and clustering.
Explaining the details of Principal Component Analysis and how to run it in R. All illustrated with examples.

Full-or-part-time: 12h
Theory classes: 5h
Self study : 7h

Module 3: Linear Discriminant Analysis

Description:
Explaining the details of Linear Discriminant Analysis and its extensions (QDA, FDA, MDA, among others).
Learning how to run all those methods in R.
All illustrated with examples.

Full-or-part-time: 12h
Theory classes: 5h
Self study : 7h

Module 4: Classification

Description:
Explaining decision trees via CART: Classification and Regression Trees.
All illustrated with examples.

Full-or-part-time: 12h
Theory classes: 5h
Self study : 7h

Module 5: Clustering

Description:
Explaining several clustering methods: hierarchical, partition (k-means), and probabilistic (model-based clustering).
All illustrated with examples.

Full-or-part-time: 12h
Theory classes: 5h
Self study : 7h

Module 6: Quiz & Final Project

Description:
A written test will be held on the last week of class and evaluate the assimilation of the basic concepts of the subject.
Additionally, the final project would be also presented orally in class on the last week of class.

Full-or-part-time: 15h
Theory classes: 2h
Self study : 13h

GRADING SYSTEM

The course assessment will be based on three main tasks: Exercises (activity 1), Exam (activity 2), and a Final Project (activity 3), which key points are:
- The exercises conducted throughout the course aim to consolidate the learning of the techniques shown in the theoretical classes.
- The written test will be held on the last week of class and evaluate the assimilation of the basic concepts of the subject.
- The final project will be presented orally on the last week of class and it is where the students must show their maturity to solve a real problem using pre-processing of the data, applying dimension reduction, and supervised and unsupervised learning methods, and contextualizing the results. Some characteristics of this project are:
o Students will choose between different alternatives to solve the problem.
o This project will be presented and publicly defended, in which the student must answer any questions about the theoretical models and methods used in the solution.
o The presentation of the project will be done during the last week of class.
Each exercise and the final project will be conducted using the statistical software R and will lead to the drafting of the relevant report writing and may be made jointly, up to a maximum of three students per group. The weight of each task in the final grade is:
- Exercises (30%)
- Exam/Quiz (30%)
- Final Project (40%)
Anyone that does not attend to any of the evaluative activities will be graded with a 0.
This course is conceived as a continuity and all three tasks are interrelated. In this way, all students who have failed the subject after the three tasks can take a reconduction. This reconduction will be schedule during the final exam period. The exam will contain theoretical questions, exercises, and questions related to the final project.
The mark from the reconduction exam will substitute for all the other marks and will be the final mark for the course. If the final grade after the reconduction is greater than or equal to 5, the final grade for the course will be 5.

BIBLIOGRAPHY

Basic:
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. The elements of statistical learning: data mining, inference, and prediction [on line]. 2nd ed. New York [etc.]: Springer, cop. 2009 [Consultation: 07/10/2022]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-0-387-84858-7. ISBN 9780387848570.
- Johnson, Richard A; Wichern, Dean W. Applied multivariate statistical analysis [on line]. Sixth edition. Harlow, Essex: Pearson, [2014] [Consultation: 23/01/2023]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5174865. ISBN 9781292024943.
- Husson, François; Lê, Sébastien; Pagès, Jérôme. Exploratory multivariate analysis by example using R [on line]. Second edition. Boca Raton: CRC Press, Taylor & Francis Group, 2017 [Consultation: 23/01/2023]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=4856173. ISBN 9781315301860.

Complementary:
- Manly, Bryan F. J; Navarro Alberto, Jorge A. Multivariate statistical methods : a primer. Fourth edition. Boca Raton: CRC Press, Taylor & Francis Group, [2017]. ISBN 9781498728966.
- Peña, Daniel. Análisis de datos multivariantes [on line]. Primera edición. Madrid: McGraw-Hill/Interamericana de España, S.L, [2013] [Consultation: 23/01/2023]. Available on: https://www-ingebook-com.recursos.biblioteca.upc.edu/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=4203. ISBN

9788448136109.
- Larose, Daniel T.; Larose, Chantal D. Discovering knowledge in data: an introduction to data mining [on line]. 2nd ed. Hoboken, N.J.: John Wiley & Sons, 2014 [Consultation: 23/01/2023]. Available on: https://onlinelibrary-wiley-com.recursos.biblioteca.upc.edu/doi/book/10.1002/9781118874059. ISBN 9781118874059.