Course guide
240615 - 240615 - An Introduction to Data Science

Last modified: 16/05/2023

Unit in charge: Barcelona School of Industrial Engineering
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.

Degree: BACHELOR'S DEGREE IN INDUSTRIAL TECHNOLOGY ENGINEERING (Syllabus 2010). (Optional subject).

Academic year: 2023 **ECTS Credits:** 4.5 **Languages:** English

LECTURER

Coordinating lecturer: JOSEP GINEBRA

Others: JOSEP GINEBRA

PRIOR SKILLS

To have passed Estadística.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
1. Basic knowledge on the use and programming of computers, operative systems, data bases and computer software with an engineering application.
2. Knowledge and capacities to organise and manage projects. Knowing the organisational structure and functions of a project office.

Transversal:
3. EFFICIENT ORAL AND WRITTEN COMMUNICATION. Communicating verbally and in writing about learning outcomes, thought-building and decision-making. Taking part in debates about issues related to the own field of specialization.
4. TEAMWORK. Being able to work as a team player, either as a member or as a leader. Contributing to projects pragmatically and responsibly, by reaching commitments in accordance to the resources that are available.

TEACHING METHODOLOGY

All classes will be taught in a computer lab. The data analysis will be done with MINITAB and with R. Every week there will be small data analysis assignements to be done at home. Students will have to do a final project. On the Q1 of 2020/21 the classes will be online, but students will also have the option to attend the class in person if they want to.

LEARNING OBJECTIVES OF THE SUBJECT

At the end of the course the student should be able to identify situations where it is useful to analize data, to identify the model and/or method of analysis that is best for his data, to build a model that summarizes the information in the data and allows to make predictions, to reduce the dimensionality and visualize multivariate data, to implement supervised and unsupervised classification algorithms, and to evaluate the quality of the results obtained.

STUDY LOAD

Type	Hours	Percentage
Self study	67,5	60.00
Hours medium group	45,0	40.00

Total learning time: 112.5 h

CONTENTS

Chapter 1: Introduction

Description:
(ENG) 1.- Statistical modeling. 2.- Multivariate analysis.

Full-or-part-time: 3h 30m
Theory classes: 1h 30m
Guided activities: 1h
Self study : 1h

Chapter 2: Linear models

Description:
(ENG) 1.- Normal linear model. 2.- Estimation by least squares and other criteria. 3.- ANOVA table and goodness of fit. 4.- Confidence intervals and tests for the coefficients. 5.- Prediction. 6.- Model checking through residual analysis. 7.- Model selection and cross validation. 8.- Model interpretation; Colinearity, bias and causality. 9.- Use of categorical explanatory variables.

Full-or-part-time: 30h
Theory classes: 6h
Laboratory classes: 6h
Guided activities: 6h
Self study : 12h

Chapter 3: Non-linear models

Description:
1.- Normal non-linear model. 2.- Model fit. 3.- Confidence intervals and tests. 4.- Model checking.

Full-or-part-time: 6h
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Self study : 3h

Chapter 4: Categorical and count response models

Description:
(ENG) 1.- Binary response model. 2.- Model fit. 3.- Confidence intervals and tests for the coefficients. 4.- Model checking. 5.- Model selection and cross validation. 6.- Model interpretation. 7.- Nominal logistic model. 8.- Contingency tables and logistic model. 9.- Count response model.

Full-or-part-time: 22h 30m
Theory classes: 4h 30m
Laboratory classes: 4h 30m
Guided activities: 4h 30m
Self study : 9h

Chapter 5: Time series models

Description:
1.- Description of a time series; Stationarity and seasonality. 2.- AR models. 3.- MA models. 4.- ARIMA models. 5.- Seasonal ARIMA models.

Full-or-part-time: 13h
Theory classes: 3h
Laboratory classes: 3h
Guided activities: 3h
Self study : 4h

Chapter 6: Visualization of multivariate data (Dimensionality reduction)

Description:
(ENG) 1.- Principal components analysis. 2.- Correspondence analysis.

Full-or-part-time: 6h
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Self study : 3h

Chapter 7: Cluster analysis (Unsupervised classification)

Description:
1.- Hierarchical methods. 2.- Partition methods (k-means algorithm). 3.- Variable cluster analysis.

Full-or-part-time: 6h 30m
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Guided activities: 1h 30m
Self study : 2h

Chapter 8: Discriminant analysis (Supervised classification)

Description:
1.- Linear discriminant analysis. 2.- Assessment of the performance of a classifier. 3.- Quadratic discriminant analysis. 4.- Logistic discriminant analysis. 5.- Classification tree methods. 6.- Nearest neighbor classifier. 7.- Sensitivity, specificity and the ROC curve.

Full-or-part-time: 13h
Theory classes: 3h
Laboratory classes: 3h
Guided activities: 3h
Self study : 4h

GRADING SYSTEM

There will be a take home midterm exam and an in class final exam.

Grade = 0,1 Assignments + 0,3 Final Project + 0,1 Midterm + 0,5 Final Exam

During the 2019-20 spring semester, as a consequence of the covid19 crisis, the qualification method will be the same one, with the only difference that the final exam will not be in class, but a take home exam.

BIBLIOGRAPHY

Basic:

- Gareth, James [et al.]. An Introduction to statistical learning : with applications in R [on line]. 2nd ed. New York: Springer Verlag, 2021 [Consultation: 11/01/2022]. Available on: https://ebookcentral.proquest.com/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=6686746. ISBN 9781071614174.
- Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning : data mining, inference and prediction [on line]. 2nd ed. New York [etc.]: Springer, cop 2009 [Consultation: 25/08/2022]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-0-387-84858-7. ISBN 0387952845.
- Peña, Daniel. Regresión y diseño de experimentos. Madrid: Alianza, 2002. ISBN 9788420693897.
- Venables, William N; Ripley, B.D. Modern Applied Statistics with S. 4th ed. New York: Springer Verlag, 2003. ISBN 0387954570.
- Peña, Daniel. Análisis de datos multivariantes [on line]. 1a ed. Madrid: McGraw-Hill/Interamericana de España, S.L., 2008 [Consultation: 05/04/2023]. Available on: http://www.ingebook.com/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=4203. ISBN 9788448191849.
- Weisberg, Sanford. Applied Linear Regression. 3rd ed. New York: Wiley, 2005. ISBN 0471663794.
- Everitt, B.S.; Dunn, G. Applied Multivariate Data Analysis [on line]. 2nd. New York: Wiley, 2001 [Consultation: 20/04/2023]. Available on: https://onlinelibrary-wiley-com.recursos.biblioteca.upc.edu/doi/book/10.1002/9781118887486.
- Greenacre, Michael J. Correspondence Analysis in Practice [on line]. 3rd ed. Boca Raton: CRC Press, 2017 [Consultation: 07/10/2020]. Available on: https://www.taylorfrancis.com/books/9781315369983. ISBN 9781498731782.

Complementary:

- Wakefield, Jon. Bayesian and frequentist regression methods. New York: Springer Verlag, 2013. ISBN 9781441909244.
- Clarke, B.; Fokoue, E.; Zhang, H.H. Principles and theory for data mining and machine learning [on line]. 1st ed. New York: Springer, 2009 [Consultation: 26/08/2022]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-0-387-98135-2. ISBN 9780387981345.
- Dobson, Annette J. An Introduction to Generalized Linear Models. 4th ed. Boca Raton: Chapman Hall, 2018. ISBN 9781138741515.
- Johnson, Richard; Wichern, Dean. Applied multivariate statistical analysis. 6th ed. Englewood Cliffs, N.J: Pearson, 2007. ISBN 9780131877153.
- Peña, Daniel. Análisis de series temporales. Madrid: Alianza, 2005. ISBN 8420691283.