Last update: 19-07-2019

270215 - AD - Data Analysis

Coordinating unit: 270 - FIB - Barcelona School of Informatics

Teaching unit: 715 - EIO - Department of Statistics and Operations Research

Academic year: 2019

Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Teaching unit Compulsory)

ECTS credits: 6 Teaching languages: Catalan

Prior skills

Knowledge of basic statistical concepts, descriptive statistics, hypothesis testing. Familiarity with the statistical software R.

Degree competences to which the subject contributes

Basic:

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

Specific:

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.

CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.

Generical:

CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

Transversal:

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

Last update: 19-07-2019

270215 - AD - Data Analysis

Teaching methodology

The learning process is a combination of theoretical explanation and practical application. The theory classes are used to explain the basic scientific contents of the course, whereas the laboratory sessions work on their application to solve real-life problems.

Practicals and project form the basis for working out the transversal competences of the students, related to team-work and public presentation of results. Practicals and project also serve to integrate the different pieces of knowledge of the course.

For hands-on computer training we use the R statistical environment.

Learning objectives of the subject

1.Exploratory Data Analysis
2.Discriminant Analysis with probabilistic hypothesis
3.Multivariate modeling
4.Time series

Study load

Total learning time: 150h	Hours large group:	30h	20.00%
	Hours small group:	30h	20.00%
	Guided activities:	0h	0.00%
	Self study:	90h	60.00%

Last update: 19-07-2019

270215 - AD - Data Analysis

Content

Data preprocessing

Degree competences to which the content contributes:

Description:
Outliers, missing data and transformations

Principal component analysis

Degree competences to which the content contributes:

Description:
Multivariate description of a table of continous variables. Regression with principal components.

Factor analysis

Degree competences to which the content contributes:

Description:
The singular value decomposition, biplots, factor analysis

Multidimensional scaling (MDS)

Degree competences to which the content contributes:

Description:
Distance measures. Metric multidimensional scaling. Algorithms.

Cluster analysis

Degree competences to which the content contributes:

Description:
Hierarchical clustering techniques. Agglomeration methods. Ward's criterion. Dendrogram.

Correspondence analysis

Degree competences to which the content contributes:

Description:
Contingency tables. Row and column profiles. Independence and chi-square statistics. Simple correspondence analysis. Biplot.

Last update: 19-07-2019

270215 - AD - Data Analysis

Discriminant analysis

Degree competences to which the content contributes:

Description:
Multivariate normal distribution. Fisher's linear discriminant analysis.

Univariate time series models

Degree competences to which the content contributes:

Description:
Exponential smoothing, ARIMA models

Intervention analysis

Degree competences to which the content contributes:

Description:
Outliers, seasonal effects, intervention analysis.

Last update: 19-07-2019

270215 - AD - Data Analysis

Planning of activities

Data preprocessing

Hours: 12h
Theory classes: 4h
Practical classes: 0h
Laboratory classes: 4h
Guided activities: 0h
Self study: 4h

Description:
Practical on data preprocessing

Specific objectives:
1

Principal component analysis

Hours: 14h
Theory classes: 4h
Practical classes: 0h
Laboratory classes: 4h
Guided activities: 0h
Self study: 6h

Description:
Application of principal component analysis in practical data analysis

Specific objectives:
1

Factor analysis

Hours: 9h
Theory classes: 2h
Practical classes: 0h
Laboratory classes: 3h
Guided activities: 0h
Self study: 4h

Description:
Practical data analysis using the method

Specific objectives:
1

Multidimensional scaling

Hours: 8h
Theory classes: 2h
Practical classes: 0h
Laboratory classes: 2h
Guided activities: 0h
Self study: 4h

270215 - AD - Data Analysis

Description:
Analysis of distance matrices with this method

Specific objectives:
1

Clustering	Hours: 12h Theory classes: 4h Practical classes: 0h Laboratory classes: 4h Guided activities: 0h Self study: 4h

Description:
Application of the method to quantitative data matrices.

Correspondence Analysis	Hours: 8h Theory classes: 2h Practical classes: 0h Laboratory classes: 2h Guided activities: 0h Self study: 4h

Description:
Application of the method with cross tables.

Specific objectives:
2

Discriminant Analysis	Hours: 12h Theory classes: 4h Practical classes: 0h Laboratory classes: 4h Guided activities: 0h Self study: 4h

Description:
Application of the method to empirical data sets

Specific objectives:
2

Last update: 19-07-2019

270215 - AD - Data Analysis

Univariate time series models	Hours: 14h Theory classes: 4h Practical classes: 0h Laboratory classes: 4h Guided activities: 0h Self study: 6h

Description:
Fitting time series models to data sets on the computer

Specific objectives:
4

Intervention analysis	Hours: 9h Theory classes: 2h Practical classes: 0h Laboratory classes: 3h Guided activities: 0h Self study: 4h

Description:
Application of intervention analysis to real data sets

Specific objectives:
4

Practical on exploratory data analysis	Hours: 18h Guided activities: 3h Self study: 15h

Description:
Student do an exploratory analysis of a data set and hand in a questionnaire about it.

Specific objectives:
1, 2, 3, 4

Project	Hours: 16h Guided activities: 3h Self study: 13h

Description:
Students realize, in couples, a complete multivariate study of a certain dataset using the techniques they studied during the course, and hand in a written report about it.

Specific objectives:
1, 2, 3, 4

Last update: 19-07-2019

270215 - AD - Data Analysis

Exam concering basic concepts	Hours: 16h 30m Guided activities: 2h Self study: 14h 30m
Description: There are two exams related to the theoretical concepts of the course. Specific objectives: 1, 2, 3, 4	

Qualification system

The student's final grade for the course is based on grades obtained for weekly homework assignments (25%), a partial exam half-way the course (25%), a final exam covering the second half of the course (25%) and a project (25%).

Each weekly assignments consists of resolving a questionnaire. These assigments aim at consolidating knowledge of the techniques exposed in the theoretical sessions. The assignments require analysis of datasets in the statistical environment R.

A project is carried out by a group of two students, and students have to show they can resolve problems with the techniques they have learned during the course. Each group hands in a written report about their project at the end of the course.

The two exams will be programmed according to the calendar of the faculty, and evaluate if students have assimilated the basic concepts of the material of the course.

For the resit exam, the student can choose to do a re-examination of only the first partial (25%), or of only the second partial (25%), or of both partials (50%). The re-evaluation thus represents at most 50% of the final course grade.

Last update: 19-07-2019

270215 - AD - Data Analysis

Bibliography

Basic:

Manly, B.F.J.; Navarro, J.A. Multivariate statistical methods: a primer. 4th ed. Boca Raton: CRC Press, Taylor & Francis Group, 2017. ISBN 9781498728966.

Johnson, R.A.; Wichern, D.W. Applied multivariate statistical analysis [on line]. 6th ed. Harlow, Essex: Prentice-Hall, 2014 [Consultation: 18/09/2019]. Available on: <https://ebookcentral.proquest.com/lib/upcatalunya-ebooks/detail.action?docID=5174865>. ISBN 9781292037578.

Peña, D. Análisis de datos multivariantes [on line]. Primera edición. McGraw-Hill/Interamericana de España, S.L, 2013Available on: <http://www.ingebook.com/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=4203>. ISBN 9788448191849.

Cuadras, C.M. Nuevos métodos de análisis multivariante [on line]. CMC Ediciones, 2012Available on: <http://www.ub.edu/stat/personal/cuadras/metodos.pdf>.

Shumway, R.H.; Stoffer, D.S. Time series analysis and its applications: with R examples. 4th ed. Springer, 2017. ISBN 9783319524511.

Complementary:

Mardia, K.V; Kent, J.T; Bibby, J.M. Multivariate analysis. Academic Press, 1979. ISBN 0124712509.

Anderson, T.W. An introduction to multivariate statistical analysis. 3rd ed. Wiley, 2003. ISBN 0471360910.

Aluja, T.; Morineau, A. Aprender de los datos: el análisis de componentes principales: una aproximación desde el Data Mining. EUB, 1999. ISBN 8483120224.

Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. Time series analysis: forecasting and control. 5th ed. Wiley, 2016. ISBN 9781118675021.

Peña, D. Análisis de series temporales. 2a ed. Madrid: Alianza, 2010. ISBN 9788420669458.

Brockwell, P.J.; Davis, R.A. Time series: theory and methods. 2nd ed. Springer-Verlag, 1991. ISBN 9781441903198.