VNIVERSITAT DE VALÈNCIA

Course Guide
44653 Exploratory data analysis

COURSE DATA

Data Subject	
Code	44653
Name	Exploratory data analysis
Cycle	Master's degree
ECTS Credits	4.5
Academic year	2023 - 2024

Study (s)

Degree	Center	Acad. year	Period
2221 - M.U. en Ciencia de Datos	School of Engineering	1	First term

Subject-matter

Degree	Subject-matter	Character
2221 - M.U. en Ciencia de Datos	5 - Exploratory data analysis	Obligatory

Coordination

Name	Department
GOMEZ SANCHIS, JUAN	242 - Electronic Engineering
IÑIGUEZ HERNANDEZ, MARIA DEL CARMEN	130 - Statistics and Operational Research
MARTINEZ SOBER, MARCELINO	242 - Electronic Engineering

SUMMARY

This course includes the first stages associated to a data analysis problem as well as the first linear statistical models related to regression and classification methods.

Data scientists usually face a set of data from different sources, format, organization, code, etc. The correct data acquisition, organization, outlier treatment, missing data imputation, transformation, feature selection, removal of redundant information, etc. is one of the most relevant and difficult stages of data analysis. This is a key stage in order to perform an appropriate processing of the data and ensure reliable and robust results (model selection, classifiers, assembling, estimation, hypothesis tests, prediction, etc).

This block also focuses on the subsequent stages relatives to data preparation and the statistical analysis of the data through regression and classification lineal models. These models are the basis ones for learning about *output* variables from a set of known *input* variables and a statistical model which connect them by mean of probabilistic tools.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

There are no specified enrollment restrictions with other subjects of the curriculum.

Other requirements

Introduction to Data Science

OUTCOMES

2221 - M.U. en Ciencia de Datos

- Students should demonstrate self-directed learning skills for continued academic growth.
- Be able to assess the need to complete their technical, scientific, language, computer, literary, ethical, social and human education, and to organise their own learning with a high degree of autonomy.
- Capacidad de análisis y síntesis, en la elaboración de informes, en la exposición, comunicación y defensa de ideas.
- Ability to access and manage information in different formats for subsequent analysis in order to obtain knowledge from data.
- Capacidad de organización y planificación de actividades de investigación, desarrollo y consultoría en el área de ciencia de datos.
- Ser capaces de acceder a herramientas de información (bibliográficas y de empleo) y utilizarlas apropiadamente.
- Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.
- Extraer conocimiento de conjuntos de datos en diferentes formatos.
- Entender la utilidad de la ciencia de datos y sus elementos asociados, así como su aplicación en la resolución de problemas, eligiendo las técnicas más adecuadas a cada problema, aplicando de forma correcta las técnicas de evaluación y, finalmente,
  interpretando los modelos y resultados.

LEARNING OUTCOMES

To know the techniques and algorithms in order to pre-process and extract the more important characteristics of a set of data.
To select the more appropriate transformations for the problem to be solved.
To know the basic principles of the statistical learning.
To know and understand the basic statistical methodology and the linear models for regression and classification settings. To apply them in real problems.

To implement linear models by means of the R software.

DESCRIPTION OF CONTENTS

1. Introduction to exploratory data analysis

In this block an introduction showing the main aspects of data visualization will be done in order to get a correct data visualization.

2. Getting and cleaning data

In this block the different data types (continuous , discrete) , importing data stored in the most common formats , data conversion , detection of anomalous data will be presented.

3. Statistical data analysis

In this block, a first approach to statistical and visual data analysis is presented. This task is a fundamental part in the understanding of the available data and in the detection of wrong values (univariate , bivariate and multivariate analysis , correlation , covariance , etc.)

4. Data transformations

This block presents methods of data transformation. In this processing step, the data are transformed or consolidate so that the resulting mining process may be more efficient, and the patters found may be easier to understand.

5. Linear models: Regression models and Analysis of Variance models

Input and output variables. Simple linear regression and multiple linear regression. Distribution normal multivariate. Regression, correlation and causality. Estimation and hypothesis testing. ANOVA table. Prediction.

6. Structured Regression Models

Variable selection. Shrinkage methods: Ridge Regression, the Lasso, and the elastic net.

7. Generalized linear models

Exponential family. Linear regression of an Indicator Matrix. Logistic Regression.

WORKLOAD

ACTIVITY	Hours	% To be attended
Theoretical and practical classes	45,00	100
Development of individual work	10,00	0
Study and independent work	6,00	0
Readings supplementary material	1,50	0
Preparation of evaluation activities	6,00	0
Preparing lectures	10,00	0
Preparation of practical classes and problem	6,50	0
Resolution of case studies	5,00	0
TOTAL	**90,00**	

TEACHING METHODOLOGY

The course will combine the theoretical and the practical part, without separating sessions devoted to theory from those devoted to practice. The lessons will be taught in a computer equipped classroom.

In the theoretical part of the classes, the teacher will introduce the concepts and methods with examples and exercises to be solved by the students.

The practical sessions will be synchronized with the theory. In these sessions, the students will learn by solving problems, exercises and case studies, in order to acquire the skills of this course.

EVALUATION

The educational evaluation of knowledge and skills achieved by the students will be made continuously throughout the course, and will consist in the following blocks of evaluation:

1. Exercises and class works submitted during the course and/or partial exams: 60% of the final grade.
2. Final exam: 40% of the final grade.

Grades obtained in paragraph 1 shall only be kept in the two examination sittings of the academic year in which they were made, since their evaluation is only possible in the teaching period.

REFERENCES

Basic

- L. Han, M. Kamber, and J. Pei. (2012) Data Mining Concepts and Techniques (third Edition). Morgan Kaufman, Elsevier
- N. Zumel and J. Mount (2014). Practical Data Science with R. Manning Publications Co
- D. Pyle (1999). Data preparation for data mining. Academic Press
- G. J. Myatt and W. P. Johnson. (2014). Making Sense of Data I. Wiley.
- Y. Zao and J. Cen (2013) Data mining Applications with R. Academic Press
- R. D. Peng (2016) Exploratory Data Analysis with R. Lean Publishing (https://leanpub.com/exdata)
- G. James, E. Witten, T. Hastie, and R. Tibshirani. (2015). An Introduction to Statistical Learning with applications in R. Corrected 6th printing. Springer http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf
- https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R
- https://en.wikibooks.org/wiki/R_Programming
- T. Hastie, R. Tibshirani, and J. Friedman (2008). The Elements of Statistical Learning. Second Edition. Springer

Additional

- Cirillo (2016) RStudio for R Statistical Computing. Cookbook Paperback
- J .Albert and M. Rizzo. (2012) R by example. Springer