Course guide
270107 - MD - Data Mining

Last modified: 30/01/2024

Unit in charge: Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science.
715 - EIO - Department of Statistics and Operations Research.

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

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer: CARINA GIBERT OLIVERAS - MARIO MARTÍN MUÑOZ - SERGI RAMIREZ MITJANS

Others: Primer quadrimestre:
XAVIER ANGERRI TORREDEFLOT - 11
SONIA GARCIA ESTEBAN - 12
CARINA GIBERT OLIVERAS - 11, 12
MANUEL GIJON AGUDO - 11, 12
CAROLINE LEONORE KÖNIG - 11, 12

Segon quadrimestre:
MANUEL GIJON AGUDO - 11, 12
MARIO MARTÍN MUÑOZ - 11, 12

PRIOR SKILLS

Foundations of probability and statistics. Basic Programming in R

REQUIREMENTS

- Prerequisite PE
- Prerequisite PRO2

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CSI2.2. To conceive, deploy, organize and manage computer systems and services, in business or institutional contexts, to improve the business processes; to take responsibility and lead the start-up and the continuous improvement; to evaluate its economic and social impact.
CSI2.3. To demonstrate knowledge and application capacity of extraction and knowledge management systems .
CSI2.6. To demonstrate knowledge and capacity to apply decision support and business intelligence systems.

Generical:
G3. THIRD LANGUAGE: to know the English language in a correct oral and written level, and accordingly to the needs of the graduates in Informatics Engineering. Capacity to work in a multidisciplinary group and in a multi-language environment and to communicate, orally and in a written way, knowledge, procedures, results and ideas related to the technical informatics engineer profession.
G9. PROPER THINKING HABITS: capacity of critical, logical and mathematical reasoning. Capacity to solve problems in her study area. Abstraction capacity: capacity to create and use models that reflect real situations. Capacity to design and perform simple experiments and analyse and interpret its results. Analysis, synthesis and evaluation capacity.

TEACHING METHODOLOGY

The learning methodology will consist in the analysis of case studies concerning complex data sets from real problems. From these problems the body of necessary scientific knowledge will be introduced. The theoretical and practical lessons are interleaved such that programming and/or integration of data mining functions enhance the assimilation of the various concepts explained. The open programming environment R will be used in the laboratory.

The laboratory classes will be devoted to solving problems related to the knowledge provided in the theory classes and to the resolution by the students of a similar problem. This problem may include the resolution of very brief conceptual questions and will be delivered for its evaluation. Finally, the students must complete two full practical works, a statistical modeling problem and a modelling problem of the "scientific", "transaction" or "marketing" kind (only one of them must be chosen by the student). This last practical work will be presented orally to the whole class.

LEARNING OBJECTIVES OF THE SUBJECT

1.Knowing the types of the main problems of Data Mining
2.Data quality assesment and preprocessing
3.Problem solving: identify the statistical and/or machine learning techniques more appropriate to solve the problem
5.Implement simple learning algorithms
6.Validation of results
7.Presentation of results in a professional environment for decision making

STUDY LOAD

Type	Hours	Percentage
Hours large group	30,0	20.00
Guided activities	6,0	4.00
Self study	84,0	56.00
Hours small group	30,0	20.00

Total learning time: 150 h

CONTENTS

Introduction to Data Mining.

Description:
Statistical modeling and types of problems: analysis of binary data ("transactions"), analysis of scientific data and analysis of data from enterprises

Visualization and dimensionality reduction

Description:
Feature selection and extraction. Visualization of multivariate data.

Clustering

Description:
Direct partitioning methods, hierarchical methods and expectation maximization

Predictive Methods

Description:
Regressió lineal múltiple i generalitzada. Regressió Logística. Xarxes Neuronals

Decision Trees

Description:
Classification and regression trees (CART).

Validation protocols and data resampling

Description:
Holdout, cross-validation and the bootstrap

Generation of association rules

Description:
A-priori and Eclat algorithms.

Discriminant Analysis

Description:
Bayesian decision theory. LDA and QDA Discriminant Analysis and Naïve Bayes

Non parametric discrimination

Description:
Nearest neighbours

Regression Shrinkage and Variable Selection

Description:
Regularized linear regression. LASSO and the Elastic Net methods.

Formal concept analysis

Description:
Formal method for pattern finding

Preprocessing

Description:
a

Bagging i ensemble methods

Description:
Bagging i ensemble methods

ACTIVITIES

Development Unit 1

Specific objectives:
1

Full-or-part-time: 2h
Theory classes: 2h

A review of R language

Full-or-part-time: 6h
Laboratory classes: 6h

Development of item 2

Specific objectives:
2

Full-or-part-time: 16h
Theory classes: 4h
Laboratory classes: 4h
Self study: 8h

Development of item 3

Specific objectives:
2

Full-or-part-time: 9h
Theory classes: 3h
Laboratory classes: 2h
Self study: 4h

Development of Item 4

Specific objectives:
2

Full-or-part-time: 11h
Theory classes: 3h
Laboratory classes: 4h
Self study: 4h

Development of item 5

Specific objectives:
2

Full-or-part-time: 9h
Theory classes: 3h
Laboratory classes: 2h
Self study: 4h

Development of Item 6

Specific objectives:
2

Full-or-part-time: 7h
Theory classes: 3h
Self study: 4h

Development of Item 7

Specific objectives:
2

Full-or-part-time: 9h
Theory classes: 3h
Laboratory classes: 2h
Self study: 4h

Development of Item 8

Specific objectives:
2

Full-or-part-time: 11h
Theory classes: 3h
Laboratory classes: 4h
Self study: 4h

Development of Item 9

Specific objectives:
2

Full-or-part-time: 11h
Theory classes: 3h
Laboratory classes: 2h
Self study: 6h

Development of Item 10

Specific objectives:
6

Related competencies :
G9. PROPER THINKING HABITS: capacity of critical, logical and mathematical reasoning. Capacity to solve problems in her study area. Abstraction capacity: capacity to create and use models that reflect real situations. Capacity to design and perform simple experiments and analyse and interpret its results. Analysis, synthesis and evaluation capacity.

Full-or-part-time: 13h
Theory classes: 3h
Laboratory classes: 4h
Self study: 6h

Practice 1

Specific objectives:
2, 3, 5, 6

Related competencies :
G9. PROPER THINKING HABITS: capacity of critical, logical and mathematical reasoning. Capacity to solve problems in her study area. Abstraction capacity: capacity to create and use models that reflect real situations. Capacity to design and perform simple experiments and analyse and interpret its results. Analysis, synthesis and evaluation capacity.

Full-or-part-time: 23h
Guided activities: 3h
Self study: 20h

Practice 2

Specific objectives:
3, 5, 6, 7

Related competencies :
G3. THIRD LANGUAGE: to know the English language in a correct oral and written level, and accordingly to the needs of the graduates in Informatics Engineering. Capacity to work in a multidisciplinary group and in a multi-language environment and to communicate, orally and in a written way, knowledge, procedures, results and ideas related to the technical informatics engineer profession.
G9. PROPER THINKING HABITS: capacity of critical, logical and mathematical reasoning. Capacity to solve problems in her study area. Abstraction capacity: capacity to create and use models that reflect real situations. Capacity to design and perform simple experiments and analyse and interpret its results. Analysis, synthesis and evaluation capacity.

Full-or-part-time: 23h
Guided activities: 3h
Self study: 20h

GRADING SYSTEM

The evaluation of the course will be based on the grade obtained in the exercises developed during the lab sessions. On the other hand there will be two practical works. For each practical work, the student will deliver the corresponding written report. Finally, at the end of the course, the students must present orally the second practical work.

The student will be required to show the necessary reasoning as well as English skills. These skills will be are evaluated using the corresponding rubrics.

The overall laboratory grade is the average of the grades obtained for the exercises developed out of the laboratory sessions.

The final mark will be obtained as follows:

Lab = overall laboratory grade
PR1 = grade for the first practical work
PR2 = grade for the second practical work

Final grade = 0.2*Labo + 0.4*Pr1 + 0.4*Pr2

In both practical works (counting 40% each), 35% corresponds to the technical correction and 5% corresponds to the 'reasoning' generic competence, so that this competence gets an overall weight of 10% of the final grade.

BIBLIOGRAPHY

Basic:
- Hand, D.J. Construction and assessment of classification rules. Wiley, 1997. ISBN 978-0-471-96583-1.
- Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer, 2009. ISBN 9780387848570.
- Hernández Orallo, J.; Ramírez Quintana, M.J.; Ferri Ramírez, C. Introducción a la minería de datos. Pearson, 2004. ISBN 9788420540917.
- Maindonald, J.H.; Braun, J. Data analysis and graphics using R: an example-based approach. 3rd ed. Cambridge University, 2010. ISBN 9780521762939.
- Duda, R.O.; Hart, P.E.; Stork, D.G. Pattern classification. 2nd ed. John Wiley & Sons, 2001. ISBN 0-471-05669-3.

Complementary:
- Aluja Banet, T.; Morineau, A. Aprender de los datos: el análisis de componentes principales: una aproximación desde el Data Mining. EUB, 1999. ISBN 9788483120224.

RESOURCES

Hyperlink:
- http://www.cran.es.r-project.org- http://www.cs.waikako.ac.nz- http://www.kdnuggets.com/