

Subject

Data Mining I

Year: 1

Credits: 5 ECTS

Language: Spanish

Competencies

Core competences:

CB1. Students have demonstrated knowledge and understanding in an area of study that builds on the foundation of general secondary education and is usually at a level that, while relying on advanced textbooks, also includes certain elements involving cutting-edge knowledge in their field of study

CB2. Students are able to apply their knowledge to their work or vocation in a professional manner and possess the skills that are typically demonstrated through the development and defence of arguments and problem solving within their field of study.

General competences:

CG1. Analytical thinking and ability to respond to complex challenges

Specific competences:

CE2. Have the programming skills required to address the entire data life-cycle

CE3. Have statistical and mathematical skills applied to data science

CE4. Be able to formulate a problem related to the world of data in terms of modeling

Learning outcomes

RA1. Be able to analyze, logically and from different perspectives and disciplines, the complex challenges they face. To this end, identify which are the key and important aspects of the problem, analyze them and, after a critical and reasoned consideration, propose, in a reasoned manner, different creative and transformative alternatives/solutions.

RA9. Have the programming skills required to address the entire data life-cycle

RA10. Have statistical and mathematical skills applied to data science

RA11. Be able to formulate a problem related to the world of data in terms of modeling

Syllabus

Introduction to data mining

- Extracting knowledge from data
- Applications
- Supervised and unsupervised learning

Data for the data mining task

- Types of variables
- Data preparation: cleaning, missing values
- Reduction in the number of attributes
- UCI data repository

Linear and logistic regression

- Numerical versus categorical type
- Equation of the line obtained by means of least squares
- Equation using logistic regression
- Comparison

Decision trees: regression and classification (N8, J48, C4.5)

- Terminology
- The algorithm
- Using entropy for selecting attributes
- Maximizing information gain
- Gini Index
- Estimating prediction accuracy by using train and test
- Confusion matrix: false positives and true positives
- Overfitting

Introduction to classification: Naive Bayes and KNN classifier

- What is a classifier

- Naive Bayes: a posteriori probability
- K-Nearest Neighbours (KNN): distance measurements

Clustering

- Introduction
- Types of algorithms: k-means and hierarchical clustering
- K-means
- Hierarchical clustering

Rules of association

- Shopping basket example
- Transactions and itemsets
- Support of an itemset
- Generating rules of association
- The a priori algorithm

Recommendation Systems

- Objectives of a recommender system
- Basic models for a recommender algorithm
- Evaluation of a recommender system
- The cold start problem
- Comparison of recommenders
- Other recommendation algorithms

Training activities

The training activities planned for this module are the following:

- Challenge-based learning (2 ECTS)
- Workshops (0.5 ECTS)
- Online resources (0.5 ECTS)
- Learning communities (1 ECTS)
- Individual work (1 ECTS)

Assessment system

Assessment will be by means of the continuous assessment system, providing constant feedback to both teachers and students on the learning process throughout the academic period:

- Learning activities involving the presentation of knowledge and individual study may be assessed by means of oral and/or written tests, which will account for a maximum of 60% of the final mark.
- The training activities aimed at acquiring the practical skills of the subjects will be

assessed through the completion of various activities (assignments, case studies, challenges, etc.) accounting for at least 40% of the final mark.

Details of the assessment and marking will be made explicit in the annual academic planning of the subjects, in accordance with the teachers responsible and the determining factors of each course.

Bibliography

- Bramer, M (2016). Principles of Data Mining. 3rd edition. United Kingdom. Springer
- Aggarwal, C (2016). Recommender Systems. Switzerland. Springer