COURSE DATA

Data Subject

Code	36427
Name	Grouping and varieties
Cycle	Grade
ECTS Credits	6.0
Academic year	2023 - 2024

Study (s)

Degree	Center	Acad. year	Period
1400 - Degree in Computer Engineering	School of Engineering	4	First term
1406 - Degree in Data Science	School of Engineering	3	First term

Subject-matter

Degree	Subject-matter	Character
1400 - Degree in Computer Engineering	16 - Optional subject	Optional
1406 - Degree in Data Science	9 - Machine Learning and Data Mining	Obligatory

Coordination

Name	Department
MARTINEZ GIL, FRANCISCO	240 - Computer Science

SUMMARY

The subject 'Agrupamiento y Variedades' is the natural complement of the subject 'Machine Learning' that is also taught in the first semester of the third year.

The most important techniques for finding structures and patterns in unlabelled data sets are reviewed. The course is divided into two parts. The first covers the most widespread unsupervised learning techniques, also known as clustering techniques. Hierarchical grouping, partitioning grouping in its variants 'Hard' (K-Means) and 'Soft' (Fuzzy-C-Means), and non-compact clustering models such as density-based clustering (DBSCAN) and graph-based clustering (spectral clustering) are reviewed.

In the second part, the knowledge about dimensionality reduction already introduced in the second course subject 'Linear Models' is expanded. While in this subject PCA is introduced as the most important linear model for dimensional reduction, in 'Agrupamiento y Variedades" we will extend this problem to nonlinear models, that is, to spatial data organizations that cannot be modeled using hyperplanes. This second part begins with a simple model, Self-Organizing Maps (SOM). Later, some techniques will be reviewed within the group known as 'Manifold Learning', specifically ISOMAP and Locally Linear Embedding. Finally, the technique for data visualization known as t-SNE will be reviewed, which connects us with the second year 'Data Visualization' course.

Apart from the different techniques presented, the student will acquire / review knowledge about essential concepts in machine learning, such as the concepts of similarity, metrics, feature space, partition, or the problem of the curse of dimensionality.

Due to the numerous topics that this quarterly subject tries to tackle, the presentation of the topics will necessarily be superficial, presenting the basic ideas of each technique, the types of problems it intends to solve and exposing the basic algorithms.
The knowledge taught in this subject is the basis for understanding the learning techniques presented in the subjects 'Natural Language Processing' and the different options for analysis of geographic data, audio and voice, health and the Web and social networks.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

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

Other requirements

It is necessary to bear in mind the contents of probability reviewed in the subject 'Probability and Simulation', as well as the concept of matrix diagonalization and spectrum of a matrix given in the subject 'Algebra', as well as the content on optimization techniques of the subject 'Optimisation' as well as the contents of the subject 'Linear Models' referring to dimensionality reduction techniques.

OUTCOMES

1400 - Degree in Computer Engineering

- SI3 - Ability to actively participate in the specification, design, implementation and maintenance of information and communication systems.
- C1 - Ability to know the fundamentals, paradigms and techniques in the field of intelligent systems, and to analyse, design and build computer systems, services and applications that use these techniques in any field of application.
- C2 - Ability to acquire, obtain, formalise and represent human knowledge in a computable form for solving problems through a computer system in any field, particularly in those related to aspects of computing, perception and action in intelligent environments.
- C3 - Ability to recognise and develop computational learning techniques and to design and implement applications and systems that use them, including those for the automatic retrieval of information and knowledge from large volumes of data.

1406 - Degree in Data Science

- (CG02) Ability to solve problems with initiative and creativity and to communicate and transmit knowledge, abilities and skills, which should include the ethical and professional responsibility of the activity of a data scientist.

- (CG03) Capability to elaborate models, calculations, reports, to plan tasks and other works analogous to the specific field of data science.
- (CT03) Ability to defend your own work with rigor and arguments and to expose it in an adequate and accurate way with the use of the necessary means.
- (CT05) Ability to evaluate the advantages and disadvantages of different methodological and / or technological alternatives in different fields of application.
- (CE03) Ability to solve classification, modelling, segmentation and prediction problems from a set of data.
- (CE06) Ability to represent and visualise data sets for the extraction of knowledge.
- (CE07) Ability to model dependency between a response variable and several explanatory variables, in complex data sets, using machine learning techniques, interpreting the results obtained.
- (CE13) To know how to design, apply and evaluate data science algorithms for the resolution of complex problems.
- (CB3) Students must have the ability to gather and interpret relevant data (usually in their field of study) to make judgements that take relevant social, scientific or ethical issues into consideration.
- (CB4) Students must be able to communicate information, ideas, problems and solutions to both expert and lay audiences.

LEARNING OUTCOMES

Know the concept of clustering and main algorithms (hierarchical, partitioned) (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)
Know the main maniflod -based algorithms that exist. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)
Know the problems of the proposed algorithms and possible solutions. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

DESCRIPTION OF CONTENTS

1. Introduction

1.1 Unsupervised learning
1.2 Notion of similarity. Metric concept. Metric types
1.3 Basic concepts: feature space, feature vector, partition, proximity matrix.
1.4 Concept of clustering. Taxonomy

2. Hierarchical clustering

2.1 Basic ideas. Aglomerative and divisive clustering
2.2 Linkage Types
2.3 Basic algorithms
2.4 Dendrograms. Interpretation
2.5 Properties. Quality measures

3. Partitional clustering

3.1 Expectation-Maximization model (EM).
3.2 Basic ideas. Soft and Hard clustering
3.3 'Hard' clustering. K-Means algorithm
3.4 Associated problems. Initialization. Choice of the number of clusters.
3.5 'Soft' clustering Fuzzy-C-Means algorithm
3.6 Associated problems.

4. Graph-based clustering

4.1 Clustering based on graphs. Spectral clustering
4.2 Basic ideas. Graph representation of data.
4.3 The Laplacian of the graph. Spectral decomposition.
4.4 Algorithms. Implementations with libraries.

5. Ndensity-based clustering

5.1 Basic ideas. Density concepts
5.2 DBSCAN algorithm
5.3 Implementations in libraries and examples

6. Evaluation and validation of clustering

6.1 Techniques for cluster evaluation
6.1.1 Purity. Rand Index. Measure F
6.2 Techniques for cluster validation
6.2.1 Regularity. Compactness and isolation. Hopkins statistics. SSE

7. Introduction to dimensionality reduction techniques

7.1 The problem of the curse of dimensionality
7.2 Manifold concept and problems of nonlinear groupings.
7.3 Concept of intrinsic dimensionality

8. Self-Organized Maps (SOM)

8.1 Concept of competitive learning. SOM concept.
8.2 Algorithm. Strengths and weaknesses of the approach. Implementations in libraries.
8.3 Examples.

9. Manifold Learning

9.1 Introduction. Basic ideas about Manifold Learning techniques
9.2 ISOMAP algorithm. Strong and weak points. Parameters.
9.3 Locally Linear Embedding (LLE) algorithm. Strong and weak points. Parameters.
9.4 Examples

10. Dimensionality reduction for visualization

10.1 Introduction to the problem.
10.2 t-SNE algorithm. Strong and weak points. Behavior
10.3 Implementations in libraries.
10.4 Examples.

WORKLOAD

ACTIVITY	Hours	% To be attended
Theory classes	32,00	100
Laboratory practices	20,00	100
Classroom practices	8,00	100
Development of group work	10,00	0
Development of individual work	10,00	0
Study and independent work	20,00	0
Readings supplementary material	10,00	0
Preparation of evaluation activities	10,00	0
Preparing lectures	10,00	0
Preparation of practical classes and problem	10,00	0
Resolution of case studies	5,00	0
Resolution of online questionnaires	5,00	0
TOTAL	**150,00**	

TEACHING METHODOLOGY

In the theoretical activities the topics will be developed exposing them to the teacher using audiovisual means, providing a global and integrating vision of the contents. Student participation in class will be encouraged through questions during the presentation and simple questions to fix the concepts presented. CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

In the problem solving classes, the problems raised one week in advance will preferably be solved by students, so that the student has enough time to work it at home. Discussion of problems in the classroom will be encouraged. CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

Tasks of greater complexity and scope than those proposed in the classroom activities will be planned in the laboratories. Group work in pairs of the proposed practices will be encouraged. Also the student will become familiar with the scientific calculation libraries of Python as well as with tools for (code presentation such as Jupyter Notebook. CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

EVALUATION

First call:

At least a partial objective test will be carried out during the semester of course delivery.

A final objective test will be carried out in the First Call.

The value of the partial tests may reach up to 50% of the theory grade (SE1). The rest of the percentage will be assigned to the final test.

The percentage over the final grade for this part (SE1) will be 50%. (CB5, CT03, CT05, CE03, CE13)

The value of laboratory practices (SE2) will represent 35% of the total grade for the course (CB5, CT03, CT05, CE03, CE13)

The value of the continuous evaluation (SE3) will represent 15% of the total grade (CB5, CT03, CT05, CE03, CE13)

It is necessary to obtain a minimum grade of 4.5 in each of the previous parts (SE1, SE2, SE3) in order to pass the course.

Second call

The continuous assessment mark, as it involves face-to-face activities, is considered non-recoverable in the second call. Although there is no restriction on the minimum grade of the first call.

For the lab mark, the student will carry out the defense of several practices that summarize the contents or a practice exam will be carried out at the end of the theory exam. The professor must decide between both methods. This mark will be the 100% of the (SE2) mark for the second call.

Instead, the minimum grade restriction remains for the theory. The theory grade will only be the grade for the second exam final exam (without considering the partial exams taken in the course).

The final grade will be obtained as 50% SE1, 35% SE2, 15% SE3.

REFERENCES

Basic

- Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Pearson (2006)
- An introduction to Statistical Learning . Gureth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Springer (2013)
- Pattern Recognition. Sergios Theodoridis. AP (2009) Hay versión electrónica
- Data Mining. Concepts and Techniques. Jiawei Han, Micheline Kamber, Jian Pei. Morgan Kaufmann.(2012)

Additional

- Scikit-Learn Users Guide   (Hay versión electrónica)
- Python Data science handbook. Jacob Vanderplas. O'Reilly. (2016)