COURSE DATA

Data Subject

Code	36426
Name	Machine learning
Cycle	Grade
ECTS Credits	6.0
Academic year	2023 - 2024

Study (s)

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

Subject-matter

Degree	Subject-matter	Character
1406 - Degree in Data Science	9 - Machine Learning and Data Mining	Obligatory

Coordination

Name	Department
MARTIN GUERRERO, JOSE DAVID	242 - Electronic Engineering
MARTINEZ SOBER, MARCELINO	242 - Electronic Engineering
VIVES GILABERT, YOLANDA	242 - Electronic Engineering

SUMMARY

The subject "Machine Learning" represents the first contact of the student of the Degree in Data Science with non-linear mathematical models and the corresponding learning algorithms that allow the extraction of information stored in databases to fundamentally solve problems of the following types:

- Classification
- Grouping
- Regression
- Prediction
- Planning

Each of these problems requires an approach that may be different from the point of view of learning. Therefore, the course begins by reviewing basic concepts and definitions to set the framework that will allow introducing the different types of machine learning to be studied: supervised, unsupervised, semi-supervised, active and reinforcement learning. Next, we will study how to evaluate the results in these problems, the different existing metrics, the need to split the data sets to guarantee an acceptable performance and possible improvements that may arise, such as boosting or bagging techniques to generate ensembles.

Once the framework is established, different machine learning models are studied to solve the problems described above. In this way, “Machine Learning” goes one step further than subjects from previous years where students have focused mostly on more descriptive data analyses or on models based on linear approximations.

Regarding the models, the methods most widely used and most popular today are studied, with the exception of the neural networks described in the subject “Connectionist Models”, in the second semester. In particular, the characteristics, operation, adaptation to different problems and the interpretation of models based on Support Vector Machines (SVM) and decision trees will be described in detail; In the case of trees, generalization with ensembles will have special relevance since it gives rise to Random Forest (RF) models, which are one of the most powerful approaches that can be used to solve regression and classification problems.

Theory lessons will be taught in Spanish and practical and laboratory lessons as according to the information sheet available on the web page of the degree.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

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

Other requirements

The course is part of the 3rd year of the degree, when the student is supposed to have a basic knowledge of the processes involved in intelligent data analysis, such as dealing with noise, outliers or missing data, as well as representation and interpretation of the data in a multidimensional space and the generation of visualizations for a useful information extraction about the problem. There are no additional requirements to be able to follow the course. With respect to the contents taught in the first

OUTCOMES

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.
- (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

The most important learning outcomes of the subject are, as follows:

- Know and implement data-based trees.
- Know the base of the kernel methods and the different kernels that can be proposed.
- Obtain association rules from databases (basket analysis).
- Know the different ways of associating expert systems.

Each of these four results enables the student to acquire all the skills described above to a greater or lesser extent. However, specifying the results associated with the different competences, regarding basic and general competences, the most relevant learning results will be:

- Carrying out works and reports with different levels of guidance so that the student achieves a high degree of autonomy, both in terms of deciding on the most appropriate approaches to solve a problem, obtaining the best possible results and interpreting them in a multidisciplinary environment (CG02, CB3, CB4).
- Treatment of data sets in different formats, providing the student with the ability to carry out a correct data treatment regardless of their format (CG02, CG03).

The transversal competences will have the following results associated:

- Making formal presentations for an expert and non-expert audience (CT03, CT05).
- Defense of the arguments presented in the presentation against any doubts that may arise (CT03, CT05).
- Ability to critically evaluate the own work and that of other colleagues and colleagues (CT05).

Finally, the specific competences will be reflected in the following learning outcomes:

- Development of classification, modeling, grouping, prediction, regression and planning models, understanding the internal processing performed by models and algorithms and having the ability to propose variants that can improve the results achieved (CE03, CE07, CE13).
- Ability to adapt the models to the peculiar characteristics of each problem (CE03, CE07, CE13).
- Ability to choose the most appropriate models for each problem and evaluate them with objective metrics (CE13).
- Interpretation of the results of the modeling and usefulness of the solutions provided in a multidisciplinary environment (CE07).
- Understand the different types of learning algorithms and express their capacity both individually and in combination, if appropriate (CE07, CE13).

DESCRIPTION OF CONTENTS

1. Preliminary topics

This first thematic unit introduces necessary concepts necessary prepare the data sets so that they can be analyzed by machine learning methods. In particular, the following contents will be studied:

1. Definitions: sample, pattern, feature, algorithm, dimensionality,...
2. Standardization and coding
3. Feature selection

3.1. Filter methods

3.2. Wrapper methods

4. Feature extraction(*): Decomposition into singular values, Principal Component Analysis, Independent Component Analysis, ...

(*)These methods and other more sophisticated ones will be studied in much more detail in the subject Grouping and Varieties

2. Approaches to learning and related problems

The second thematic unit describes the different types of learning that define the algorithms that do the task of extracting information in machine learning models. Depending on the learning scheme used, different problems can be addressed as long as they are defined by a data set. The different sections that will be studied in thematic unit II are:

1. Supervised learning
1.1. Classification problems
1.2. Regression and prediction problems
2. Unsupervised Learning: Clustering and Segmentation Problems(*)
3. Semi-supervised approaches
3.1. Semi-supervised learning
3.2. Active learning
4. Reinforcement learning: optimization problems

(*)These methods will be studied in depth in the subject Grouping and Varieties

3. Model assessment

The third thematic unit focuses on the different metrics that can be used to assess the performance of machine learning models. Depending on the type of problem addressed and therefore the learning scheme, the metrics to consider are different. In addition, the suitability of one or another metric will be analyzed based on the final objective pursued so that the models can be biased to minimize or maximize certain criteria. Some problems to be taken into account and different techniques that improve the performance of the models and make them more useful for their use in a real problem will also be described. The table of contents of thematic unit III is as follows:

1. Overtraining and overfitting
2. Data set splitting
2.1. Hold-out
2.2. V-fold
2.3. Leave-one-out
3. Assessing the performance of machine learning models
3.1. Classification problems
3.2. Regression problems
4. Model improvement
4.1. Boosting
4.2. Expert Committees: Ensembles
4.3. Bagging

4. Support Vector Machines

Support Vector Machines (SVMs) represent a learning model initially proposed for classification tasks but that can also be applied for regression after carrying out some variations. They are especially indicated in sparse spaces with low data density, worsening their performance considerably as the number of samples in the data set increases. In this thematic unit its operation and characteristics will be reviewed, broken down as follows:

1. Introduction
2. Optimum separation hyperplane
3. The kernel trick
4. Supported Vector Regressor (SVR)

5. Decision trees

This last theoretical thematic unit describes machine learning models that are based on tree structures, beginning with single-tree models and ending up with Random Forests (RFs) that using bagging generate a set of trees; this model represents the state of the art in classification and regression problems. The table of contents of the thematic unit is:

1. Representation
2. Entropy and information gain
3. Pruning
4. Main algorithms

4.1. ID3
4.2. C4.5
4.3. CART: Classification and Regression Trees
4.4. CHAID: Chi-square Automatic Interaction Detection

5. Tree ensembles

5.1. Random Forest (RF)
5.2. Extremely Randomized Trees (ERTs)

6. Current trends in Machine Learning

After studying the most known methods in Machine Learning, the goal of this unit is to end up the course with the description of some topics that are currently proposed and researched, so that the student may know the latest developments of the field. The next table of contents is hence flexible and adaptable to the appearance of new interesting proposals:

1. Deep Learning.
2. Quantum Machine Learning.
3. New applications of Machine Learning.

7. Laboratory practice

Finally, due to its importance in the subject, it has been considered convenient to include as an independent thematic unit the practices to be carried out in the laboratory (computer room), where the student will learn to implement the models described in the theory classes. Many of the methods analyzed in the subject only make sense when they are developed in a laboratory environment in which their potential can be observed, since it can be relatively difficult to understand all their operating characteristics solely on the basis of theoretical studies, exercises and simple problems. Six laboratory practices are proposed, corresponding to the theoretical contents previously described in the previous thematic units:

1 Preprocessing of data sets
1.1. Standardization
1.2. Coding
1.3. Feature selection
1.3.1 Filter methods
1.3.2 Wrapper methods
2 Feature extraction
2.1. Set separation: Hold-out, V-fold, Leave-one-out
2.2. Principal Component Analysis (PCA)
3 SVMs in classification problems
3.1. Examples
3.2. Active learning with SVMs
4 SVRs in regression problems
5 Decision Tree-Based Models for Classification and Regression Problems
6 Models based on tree ensembles
6.1. Bagging
6.2. Random Forest

WORKLOAD

ACTIVITY	Hours	% To be attended
Theory classes	32,00	100
Laboratory practices	20,00	100
Classroom practices	8,00	100
Attendance at events and external activities	2,00	0
Development of group work	6,00	0
Development of individual work	5,00	0
Study and independent work	35,00	0
Readings supplementary material	4,00	0
Preparation of evaluation activities	20,00	0
Preparing lectures	4,00	0
Preparation of practical classes and problem	4,00	0

Resolution of case studies	7,00	0
Resolution of online questionnaires	3,00	0
TOTAL	150,00	

TEACHING METHODOLOGY

The teaching methodologies used in this subject are:

MD1 - Theoretical activities (CG03, CB4, CT03, CT05, CE03, CE07, CE13): Expository development of the subject with the participation of the student in the resolution of specific questions. Completion of individual evaluation questionnaires.

MD2 - Practical activities (CG02, CB3, CT05, CE03, CE07, CE13): Learning by means of problem solving, exercises and case studies through which skills are acquired on different aspects of the subject.

MD4 - Laboratory and / or computer classroom work (CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13): Learning by carrying out activities carried out individually or in small groups and carried out in laboratories and / or computer classrooms.

Next, the teaching-learning method to be used in the subject is described in more detail. The teaching methodology will have two different approaches, one for the theoretical and problem classes and the other for the practical laboratory classes. The UV e-learning tool *Aula Virtual* will be used, especially in terms of material availability, automatic evaluations and in remote classes, if applicable.

Regarding the theoretical classes, the learning will be done in both directions, from the teacher to the student, and from the student to the teacher. In the part that comes from the teacher, there will be two sources of knowledge generation. On the one hand, the master class in which the teacher will introduce the new concepts that appear, relating them to the students' prior knowledge to facilitate their understanding. On the other hand, before the class, the student will have material to prepare the theoretical classes minimally and thus streamline the teaching-learning process. This same material will also contain information so that the student can complement and review the information received in class.

In the process that flows from the student there will also be two approaches that correspond to the sources of knowledge generation previously discussed. The teacher's master class will be reinforced with the resolution of practical exercises and increasingly complex problems as the thematic unit progresses. Although the teacher will carry out some exercises as way of illustration, most of these problems will have to be solved properly by the students to guarantee a total understanding of the contents of the subject, which despite having a theoretical component, its strong point is the practical application to different problems and these exercises and problems enable the student to find out the particularities of the different algorithms in their application to problems of various kinds. In addition, the student will carry out works with an individual research component in contents that could be considered as a sophisticated version of those described in the subject; in particular, it is proposed that the student can carry out, present and defend works on current methods of machine learning, which at the same time will improve their foundations in the most basic contents that are described in detail in the subject. It is contemplated that some of these works may be voluntary and be carried out individually or in pairs.

Regarding practical classes, three teaching methodologies can be distinguished. Firstly, and prior to the lab sessions, the students will have to prepare independently the practical exercises to be carried out, consulting the doubts that may arise with the teachers, preferably before the class. This aspect will be supported by completing a short and simple questionnaire at the beginning of the practice to check that the preparation has been carried out correctly.

The second teaching methodology that appears in the practical classes is the work to be done during the practice session, which will basically consist of programming in Python to obtain the correct solution to the proposed problems. This work will be done individually or in pairs and will have the supervision of the teacher at all times; firstly, because the exercises will be previously explained and secondly because the student will be able to consult their doubts at all times in order to make the correct and adequate progresses.

The third teaching approach in practice again gives prominence to the student, who at the end of the practice will have to be able to make a critical discussion of the results achieved and correctly answer the questions asked by the teacher and carry out the exercises proposed in the session. Depending on the circumstances (classroom occupation, face-to-face class or not, etc.), this discussion could be carried out automatically using the tools available in the *Aula Virtual*.

EVALUATION

The final grade for the course will be obtained as a result of the weighted average between the theory and practical parts. According to the credits assigned to each part, the theory will have a representation of 2/3 in the final grade and the remaining third will correspond to the practical part.

The theory grade corresponding to the first call will come out as a result of:

- SE1 (60%; CG02, CG03, CB3, CB4, CT05, CE03, CE07, CE13): Objective tests, consisting of one or more examinations of theoretical questions, synthetic problems and real practical problems. A minimum grade of 5 (out of 10) will be required in this part to pass the course.

- SE2 (30%; CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13): Works, memories and oral presentations.

- SE3 (10%; CG02, CB4, CT03, CE03, CE07, CE13): Continuous assessment of each student, based on the student's participation and degree of involvement in the teaching-learning process, taking into account regular attendance at planned face-to-face activities and the resolution of regularly proposed questions and problems.

Regarding the qualification of the practical part, 40% of the grade will correspond with SE2 (CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13) and 60% to the grade obtained in the final practice (SE1; CG02, CG03, CB3, CB4, CT05, CE03, CE07, CE13), which will take place in the last lab session. The final practice will be an objective test that will be evaluated individually and that will consist of different exercises related to one or more previous practices. A minimum grade of 5 (out of 10) will be required in the final practice to pass the course. Of the 40% corresponding to the continuous evaluation, 70% will correspond to the completion of the exercises proposed in the practice session, which may be evaluated by the teacher at the end of the practice. The remaining 30% will come from the preparation prior to the practice session that will be quickly evaluated at the beginning of each practice session. Practices can be done individually or in pairs, with the exception of the final practice, which will necessarily be individual. In addition, the teacher may choose to individually evaluate the regular practice sessions even if they have been carried out by groups of two students.

The second call will be evaluated as the first one with two exceptions. First, in the theory part, SE1 will have a weight of 70% and SE3 of 0%. Second, in the practice part, the 100% of the grade will correspond with SE1, and will be needed a minimum grade of 5 (out of 10) to average out with the theory part.

In any case, the evaluation system will follow that established in the Regulations for Evaluation of the University of Valencia, for undergraduate and postgraduate studies

(https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639).

REFERENCES

Basic

- E. Alpayidin, F. Bach (2014). Introduction to Machine Learning, Third Edition, The MIT Press (disponible com a eBook per a la Universitat de València)
- S. Theodoridis (2015). Machine Learning: a Bayesian and Optimization perspective, Elsevier (disponible com a eBook per a la Universitat de València)
- D. Haroon (2017). Python Machine Learning Case Studies: Five Case Studies for the Data Scientist, Apress (disponible com a eBook per a la Universitat de València)

Additional

- C. M. Bishop (2016). Pattern Recognition and Machine Learning, Springer
- K. P. Murphy (2020). Machine Learning: a probabilistic perspective, Second Edition, The MIT Press
- R. O. Duda, P. E. Hart, D. G. Stark (2016). Pattern classification, Third Edition, John Wiley & Sons Inc.
- T. Hastie, R. Tibshirani, J. Friedman (2011) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer (Series in Statistics)
- S. Raschka, V. Mirjalili (2019). Python Machine Learning. Packt Publishing
- David V. (2017). Machine Learning with Python: The Basics. CreateSpace Independent Publishing Platform