



- 1. Code:** 14009 **Name:** Scalable Machine Learning techniques
- 2. Credits:** 6,00 **--Lecture:** 4,50 **--Practice:** 1,50 **Type of Course:** Compulsory
- Degree:** 189-Bachelor's Degree in Data Science
- Module:** 2-Compulsory subjects **Subject:** 6-Automatic Learning
- University Center:** SCHOOL OF INFORMATICS
- 3. Coordinator:** Gómez Adrian, Jon Ander
- Departament:** COMPUTER SYSTEMS AND COMPUTATION

4. References

Pattern classification
Pattern recognition and machine learning
Machine learning : a probabilistic perspective
Apache Spark 2.x Machine Learning Cookbook.

Duda, Richard O.
Bishop, Christopher M.
Murphy, Kevin P.
Amirghodsi, Siamak.

5. Course Outline

Course objectives

This subject is devoted to scalable Machine Learning (ML) techniques. In particular, it is focused on those techniques that can deal with huge datasets thanks to the fact their training algorithms can be parallelized in order to distribute the workload of the algorithms among several working nodes within a cluster of computers.

This subject introduces new ML techniques and reviews some techniques already studied in other subjects students have already seen in previous academic years.

The objectives of this subject are:

- (1) Showing students how to distribute the workload of the training algorithms of the Machine Learning techniques studied in this subject, i.e., scalable techniques. In some cases, even the inference will be distributed too.
- (2) Familiarize students with software tools in the field of Big Data for the parallelization/distribution of computations necessary for training models based on Machine Learning techniques.
- (3) Delve into some of the ML techniques that, although they have already been studied in previous courses, are relevant for the application of Artificial Intelligence to data analysis, both in classification and prediction tasks.

The teaching methodology used in this subject is "Project-Based Learning" where "Active Learning Methodologies" will be used in order to foster "Cooperative Learning".

In this context, students will work in groups to solve the problems and carry out the tasks posed during the semester in the academic works assigned to students. A total of four academic works per work group will be assigned during the semester. Every academic work includes the compilation of a written report and an oral presentation. Submissions and oral presentations would gather more than one task.

Written reports must be delivered according to the calendar that will be proposed at the beginning of the semester. The calendar will also include the oral presentations to be held in the seminars.
Students will present the results of the work done to solve the task developed by the working group for each one of the assignments.

The lab practices will be organized to help students to carry out the work to be done in the tasks of the academic works

Document signat electrònicament per Documento firmado electrónicamente por Electronically signed document by	UNIVERSITAT POLITÈCNICA DE VALÈNCIA	Data/Fecha/Date 30/05/2024	1 / 4	
Autenticitat verificable mitjançant Codi Segur Verificació Autenticidad verificable mediante Código Seguro Verificación Original document can be verified by Secure Verification Code		ALUW8QQ0F2Q https://sede.upv.es/eVerificador		



NOTE 1: the bibliography contains books on Machine Learning that will be used as the references where students can expand their knowledge of ML techniques. Additionally, the documentation of software tools will be used in order to see distributed/parallelized versions of the training algorithms of all studied ML techniques but Neural Networks, that will take advantage of hardware accelerators like NVIDIA GPUs.

NOTE 3: this subject uses English as vehicular communication language, trying to maximize students-teacher interaction in such language, with documentation and material in English too. However, evaluation of and questions from students can be in either English, Spanish or Valencian.

Contextualization of the course

Where is this subject located/planned within the study plan of this Bachelor's degree?

This subject is planned to be taught in the first semester of the last academic year of the Bachelor's degree, because it has many dependencies on other subjects where the necessary previous knowledge for this one has been studied.

What does this subject bring to the student's training?

A different perspective from the one studied in previous subjects of previous academic years, where the different Machine Learning techniques have been studied without exploring the possibility of leveraging the clusters of computers to distribute the workload of training models and, in some special cases too, for reducing the inference time.

Why is this subject relevant at the professional level?

Because Big-Data middle-ware, e.g., Apache Spark, has not been used in the previous academic years, and it is important all the students of this Bachelor's degree know the possibilities offered by the Big-Data software tools.

6. Recommended Prior Knowledge

- (14005) Statistical models for decision making I
- (14006) Statistical models for decision-making II
- (14010) Descriptive and predictive models I
- (14011) Predictive and descriptive models II
- (14021) Project III, data analysis

In addition to the previous knowledge required that has been described in terms of other subjects from previous academic years. It is assumed that students are familiarized with computer programming in Python as well as in the use of databases and load data from text and binary files with different formats. The assumed previous knowledge will be the starting point to work with Pandas and Scikit Learn, and other software tools.

7. Results

Fundamental results

CB02(GE) That the students know how to apply their knowledge to their work or vocation in a professional way and possess the competences that are usually demonstrated through the elaboration and defense of arguments and the resolution of problems within their field of study

CB05(GE) That students have developed the necessary learning skills to undertake further studies with a high degree of autonomy

CE04(ES) Apply statistical and machine learning methodologies and techniques for data processing and analysis.

CG03(GE) Plan the necessary processes for the extraction of knowledge from different data sources

CG01(GE) Conceive, design and use applications to extract knowledge from different data sources.

CG02(GE) Develop, manage and lead projects that use the techniques, tools and methodologies of data science in different application areas

CE16(ES) Represent knowledge in an actionable and interchangeable way to facilitate its automatic manipulation using the appropriate analysis tools.





8. Syllabus

1. Introduction
 1. Presentation of the subject
 2. Global overview of the contents
 3. Software tools to distribute workload on clusters of computers
2. Clustering and density estimation at large scale
 1. K-Means using distributed computing
 2. Gaussian Mixture Models (GMM) using distributed computing
3. Ensembles based on Decision Trees to work with huge datasets
 1. Review of basic concepts
 2. Random Forests using distributed computing
 3. Extremely Randomized Trees using distributed computing
 4. Gradient boosted decision trees using distributed computing
4. Memory based techniques to work with large datasets
 1. Kernel Density Estimation using distributed computing
 2. K-Nearest Neighbors using distributed computing
5. Distributed training of Neural Networks
 1. Fully connected NNs
 2. Convolutional NNs
 3. Recurrent NNs
 4. Cascade of classifiers
6. Active and on-line learning
 1. Active learning
 2. On-line learning
 3. Incremental learning
 4. Reinforcement Learning

9. Teaching and Learning Methodologies

<u>UN</u>	<u>LE</u>	<u>SE</u>	<u>PS</u>	<u>LS</u>	<u>FW</u>	<u>CP</u>	<u>AA</u>	<u>CH</u>	<u>NCH</u>	<u>TOTAL HOURS</u>
1	2,00	4,00	--	1,50	--	--	1,00	8,50	10,00	18,50
2	3,00	6,00	--	3,00	--	--	1,00	13,00	20,00	33,00
3	3,00	6,00	--	3,00	--	--	1,00	13,00	20,00	33,00
4	2,00	4,00	--	3,00	--	--	1,00	10,00	20,00	30,00
5	3,00	6,00	--	4,50	--	--	1,00	14,50	20,00	34,50
6	2,00	4,00	--	0,00	--	--	1,00	7,00	10,00	17,00
TOTAL HOURS	15,00	30,00	--	15,00	--	--	6,00	66,00	100,00	166,00

UN: Unit. LE: Lecture. SE: Seminar. PS: Practical session. LS: Lab sessions. FW: Field work. CP: Computer-mediated practice. AA: Assessment activities. CH: Contact hours. NCH: Non contact hours.

10. Assessment

Outline

	<u>Num. Acts</u>	<u>Weight (%)</u>
(05) Academic work	2	20
(15) Practical laboratory/field/computing/classroom test	4	20
(14) Written test	2	50
(11) Observation	1	10

EVALUATION FOR STUDENTS WHO WILL REGULARLY ATTEND CLASSES:

The final grade will be the sum of the following three parts:

(A) 50% written exams performed individually, each will be evaluated on the contents of theory and how to apply the concepts from theory to draw the solution to typical ML problems

(B) 40% academic works carried out collectively (working groups of 4 people) from use cases proposed by the teacher in the laboratory, including the oral presentation of each one of the scheduled academic works, evaluation of the laboratory work is included here





10. Assessment

(C) 10% observation of the implication of students in the resolution of problems and discussions during seminars and lab practices, in particular, the questions posed to other groups during oral presentations of the academic works

EVALUATION FOR STUDENTS WITH EXEMPTION FROM ATTENDING CLASSES:

Evaluating those students who do not regularly attend classes is not possible with the previously explained plan to get the final grade as the sum of three parts. This plan takes into account the work done by students as well as their active participation during the seminars and lab sessions.

This way, students that will not attend classes and lab sessions because they obtained the official exemption issued by ETSINF will be evaluated by means of a written exam (50% of the final grade) at the end of the semester plus one academic work (50% of the final grade) they have to submit before the written exam. The academic work will be the same the other students will carry out, but without the requirement of an oral presentation and it will be a unique document gathering all the required contents, instead of several ones.

RECOVERY:

For those students that did not pass the subject by following either the proposed progressive evaluation method or the evaluation plan for those with exemption of attending classes, a written exam corresponding to all the contents of the subject will be scheduled at the end of the semester.

This written exam will contain problems and questions related to both theory and lab practices.

11. Absence threshold

<u>Activity</u>	<u>Percentage</u>	<u>Observations</u>
Lecture Theory	20	
Seminar Theory	20	
Lecture Practice	20	
Laboratory Practical	20	
Computer Practice	20	
Field Practice	20	