uc3m | Universidad Carlos III de Madrid

Big Data Intelligence: methods and technologies

Academic Year: (2021 / 2022)   **Review date:** 07-06-2021

Department assigned to the subject: Computer Science and Engineering Department

Coordinating teacher: ALER MUR, RICARDO

Type: Compulsory ECTS Credits : 3.0

Year : 1 Semester : 1

REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

Basic programming knowledge.

OBJECTIVES

Basic Skills

Knowledge that provides a basis or opportunity for originality in developing and / or applying ideas, often in a research context

To be able to apply the broader (or multidisciplinary) acquired knowledge and ability to solve problems in new or unfamiliar environments within contexts related to their field of study

To be able to integrate knowledge and handle complexity and formulate judgments based on information that was incomplete or limited, including social and ethical responsabilities linked to the application of their knowledge and judgments

To be able to learn skills that enable them to continue studying in a way that will be largely self-directed or autonomous.

General Competencies:

To apply the theoretical underpinnings of the techniques for collecting, storing, processing and reporting, especially for large volumes of data as a basis for the development and adaptation of such techniques to specific problems

To be able to identify different techniques for storing, replicating and distributing large amounts of data, and differentiate them according to their theoretical and practical features

To identify analysis techniques most suitable for each problem and to know how to apply data for analysis, design and finding solutions

To obtain practical and efficient solutions to problems of processing large volumes of data, both individually and in teams

To be able to synthesize the findings from these analyses and to do clear and convincing presentations in a bilingual environment (English and Spanish) both in writing and orally

To be able to generate new ideas (creativity) and to anticipate new situations, in the context of data analysis and decision making

To use skills for teamwork and work with others in an autonomous way

Specific skills:

To identify and select software tools suitable for the treatment of large amounts of data

To design systems for processing data, from the collection and initial filtering, statistical analysis, and the submission of final results

To use techniques and operation research tools in procedures with massive data for analysing or

displaying results in decision support systems

To apply the basic and fundamental principles of machine learning to design procedures and improving them

To interpret functional specifications aimed at developing applications based on machine learning

To identify the opportunity to use machine learning to solve real problems

To perform detailed analysis and design of applications based on machine learning

Learning outcomes:

- Basic and fundamental knowledge of machine learning
- Understanding of basic machine learning techniques
- Practical application of basic machine learning techniques in real problems
- Capacity for analyzing the most appropriate tasks for each technique
- To understand when to use machine learning techniques for solving real problems

DESCRIPTION OF CONTENTS: PROGRAMME
1. Introduction / basic concepts
2. Methods for training classification and regression models
   2.1. Nearest neighbours
   2.2. Decision / regression trees and rules
3. Methodology
   3.1. Basic pipeline
   3.2. Hyper-parameter optimization
   3.3. Model evaluation
4. Methods for preprocessing and attribute selection (filter and wrapper):
5. Methods based on ensembles of models:
   5.1. Bagging / Random Forests
   5.2. Boosting / Gradient Boosting
6. Streaming (Spark)
7. Methods for imbalanced datasets
8. Software technologies for Machine Learning and Big Data:
   8.1. Python / Scikit-learn
   8.2. Mapreduce
   8.3. Spark / MLLIB / ML / Pyspark

LEARNING ACTIVITIES AND METHODOLOGY
Theory: Lectures will be focused on teaching all concepts related to machine learning. They will be carried out live (in-class).

Practical computer Sessions (reduced LIVE (in-class) sessions with student's own laptops): The practical classes will be developed so that, in a supervised way, students learn to solve practical cases. The practices will be carried out in groups of 2 students. There are several assignments related to topics in the course.

There will be tutorials to help the understanding both of theory and practice.

===

Training activities
Lectures
Laboratory practice
Team work
Individual student work

Teaching methodology:

Lectures with support of computer and audiovisual media, in which the main concepts of the subject are developed and basic literature is provided to supplement student learning.

ASSESSMENT SYSTEM
A) Assignments during the course (70%)
B) Final exam (30%)

In order to pass the course, A+B>=5. The final exam (B) has no cutoff grade.

% end-of-term-examination:	30
% of continuous assessment (assigments, laboratory, practicals…):	70

BASIC BIBLIOGRAPHY
 - Aurélien Géron Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019
 - Hastie, Tibshirani, Friedman The Elements of Statistical Learning, Springer, 2016

 - Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia Learning Spark, O'Reilly, 2015

 - Raul Garreta, Guillermo Moncecchi Learning scikit-learn: Machine Learning in Python, Packt Publishing, 2013

BASIC ELECTRONIC RESOURCES
 - . Spark: https://spark.apache.org/

 - . Scikit learn: http://scikit-learn.org/stable/index.html

 - Aurélien Géron . Github for Hands-on Machine Learning: https://github.com/ageron/handson-ml2