Machine learning applications

Academic Year: (2022 / 2023) **Review date: 20/05/2022 00:02:23**

Department assigned to the subject: Signal and Communications Theory Department

Coordinating teacher: ARENAS GARCIA, JERONIMO

Type: Compulsory ECTS Credits : 6.0

Year : 3 Semester : 2

REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

It is recommended to have completed the subjects about mathematical foundations from the first year (Calculus I and II, Linear Algebra, Probability and Data Analysis), the subjects related to programming and algorithms (Programming and Data Structures and Algorithms), as well as subject Statistical Learning. It is also advised that the students have already taken the Machine Learning (I and II) courses.

OBJECTIVES

- Design a data model suitable for an analysis task.
- Correctly and efficiently choose and use one or more data analysis methods including statistical or algorithmic techniques.
- Evaluate the results of the analysis and propose modifications to the analysis process.
- Know how to design and apply unsupervised inference methods for models with latent variables.
- Know how to design and apply data adaptation and curation techniques.
- Know how to design and apply natural language processing methods.
- Know how to design and apply recommendation systems.

DESCRIPTION OF CONTENTS: PROGRAMME

This course is divided into 3 thematic blocks. The first concerns the problem of adapting and cleaning a database, a critical preprocessing step that is addressed prior to any machine learning application. The next two blocks address two industry-relevant applications where machine learning techniques have achieved a great success. The understanding of how the different machine learning techniques have to be adapted to solve specific problems of interest to industry and society will provide students with a practical and general vision of applied Machine Learning.

The course ends with a final block where two visualization tools will be presented to the students, that will use them for the final project assignment.

PART I: TECHNIQUES DATA CURATION AND CLEANING

1. Problem Introduction. Data representation and visualization.
2. Organization and integration of databases from different sources.
3. Feature extraction and selection. Multivariate Analysis and Mutual Information Methods.
4. Data cleaning: data characterization, detection and imputation of corrupt data. Outlier detection.

PART II: NATURAL LANGUAGE PROCESSING
5. Text processing pipelines. Vector representation of texts.
6. Topic Modeling: Latent Semantic Indexing, Latent Dirichlet Allocation.
7. Text Vector representation and models for automatic translation using neural networks.

PART III: RECOMMENDATION SYSTEMS
8. Content-based recommendation systems.
9. Collaborative filtering recommendation systems. ALS and Prod2Vec.

BONUS TRACK: ADVANCED DATA VISUALIZATION TOOLS

- Visualization of Graph Data with Gephi
- Business Intelligence Tools

LEARNING ACTIVITIES AND METHODOLOGY

AF1: THEORETICAL-PRACTICAL CLASSES. They will present the knowledge that students should acquire. They will receive the class notes and will have basic texts of reference to facilitate the follow-up of the classes and the development of the subsequent work. Exercises, practical problems on the part of the student will be solved and workshops and evaluation test will be held to acquire the necessary skills.
AF2: Updated to allegation
AF3: INDIVIDUAL OR GROUP WORK OF THE STUDENT.
AF9: FINAL EXAM. In which the knowledge, skills and abilities acquired throughout the course will be assessed globally.
MD1: CLASS THEORY. Exhibitions in the teacher's class with support of computer and audiovisual media, in which the main concepts of the subject are developed and the materials and bibliography are provided to complement the students' learning.
MD2: PRACTICES. Resolution of practical cases, problems, etc. raised by the teacher individually or in groups.
MD3: TUTORIALS. Individualized assistance (individual tutorials) or group (collective tutorials) to students by the teacher.

ASSESSMENT SYSTEM

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

The Continuous Assessment is 70% of the student's grade and will consist of the following elements:

* 3 tests about laboratory exercises (30%): resolution of exercises similar to those proposed in the course notebooks using python.
* Final project (40%)

The final exam (30%) will consist of a written test about the theoretical and practical contents of the course

For the extraordinary call, students will be able to take a final exam for a value of 6 points (written test + laboratory) and, additionally, they will be offered a new final project for a value of 4 points.

BASIC BIBLIOGRAPHY

- null Data Visualization with Python for Beginners: Visualize Your Data using Pandas, Matplotlib and Seaborn, AI Publishing LLC, 2020

- C.C. Aggarwal Recommender Systems: The Textbook, Springer, 2016

- D. Juravsky, J.H. Martin Speech and Language Processing, Prentice Hall; 2nd edition, 2008

- J. Eisenstein Introduction to Natural Language Processing, MIT Press, 2019

- J. Ham, M. Kamber Data Mining: Concepts and Techniques (3rd. ed), Morgan Kaupfman, 2011

- S. Bird, E. Klein, E. Loper Natural Language Processing with Python, O'Reilly Media, 2009

ADDITIONAL BIBLIOGRAPHY

- C. Manning, H. Schütze Foundations of Statistical Natural Language Processing, MIT Press, 1999

- K. Murphy Machine Learning: A probabilistic Perspective, The MIT Press, 2012

- M. W. Berry Survey of Text Mining Clustering, Classification, and Retrieval, Springer, 2004

BASIC ELECTRONIC RESOURCES

- . Pandas Tutorials: https://pandas.pydata.org/pandas-docs/stable/getting_started/intro_tutorials/

- J. Arenas-García, J. Cid-Sueiro, V. Gómez-Verdejo . Introductory Notebooks on Machine Learning topics.: https://github.com/ML4DS/ML4all