Data Mining, course 2023-2024

R. Béjar, C. Mateu

Universitat de Lleida

Section 1

Contents

Data mining, focusing on Big Data

Goals of the course:

- To know the basics of fundamental data mining algorithms, but focusing on ones suitable for big data applications
- To understand how different machine learning algorithms work in the map-reduce framework of Apache Spark
- To know the basics of deep machine learning models

(1) Data mining algorithms

- We will cover basic data mining problems (like mining frequent item sets) and model-based learning problems, like clustering or regression problems
- We will discuss their computational complexity, relevant to understand their scaling behaviour as the data size increases.

(2) Machine Learning with Deep Learning

- Introduction to building and evaluating deep machine learning models
- Present typical domain applications and how to manipulate and build different Deep ML models: fully-connected models, convolutional models and models with hidden state and feedback

Section 2

Development

Working environment - 1st Part

- We will work with Apache Spark, using python 3.6+
- You can use your own Apache Spark installation, a regular installation, or working with virtual machines, google colab, docker...
- Make sure it is compatible with Spark version 3.x.x, and work always with python 3.6+

You already started using Spark in the previous subject of this BigData block !

Working environment - 2nd part

- We will be using machine learning libraries like pytorch and fast.ai
- You will be also provided with specific virtual machines and in many occassions, you will be using google colab

Class Sessions

- Class sessions will be on-site (on the Alcatel Lab)
- We will be using Jupyter notebooks to present the material and code examples. The notebooks will be available in the virtual campus.
- Use always your own laptop.
- For questions to the teachers outside of class hours, we can use the virtual campus videconf tool

Section 3

Evaluation

Your Assignments

Big data application (end of the semester) that can be developed in groups of two:

- Code (with good enough documentation) -> 40%
- Oral presentation -> 10%

Programming exercises (to present at different times of the semester):

- First part : 1 or 2 exercises -> 25%
- Second part : 1 or 2 exercises -> 25%