DATA MINING 2 - Introduction

Riccardo Guidotti

a.a. 2021/2022
Classes

• Classes
  • Monday, 11-13 (academic?), Room Fib C and MS Teams
  • Thursday, 11-13 (sharp?), Room Fib A and MS Teams

• Office Hours
  • Thursday, 15-17, MS Teams ???
  • Appointment [DM2 Meeting] at riccardo.guidotti@unipi.it

• Teaching Assistant
  • Francesco Spinnato [DM2 Meeting] at francesco.spinnato@sns.it
• Module 1: Imbalanced Learning, Dimensionality Reduction and Anomaly Detection
  • CRISP
  • Dimensionality Reduction
  • Imbalanced Learning
  • Anomaly Detection

• Module 2: Advanced Classification Methods
  • Naive Bayes Classifier
  • Linear and Logistic Regression
  • Support Vector Machines
  • Neural Networks
  • Ensemble
  • Gradient Boosting
  • Rule-based Classifiers

• Module 3: Time Series
  • Similarity
  • Approximation
  • Motif, Shapelets
  • Classification, Clustering

• Module 4: Sequential Patterns and Advanced Clustering
  • Sequential Pattern Mining
  • X-Means, OPTICS
  • Transactional Clustering

• Module 5: Ethic Principles
  • Explaianbility
Laboratory

- Python
- Jupyter Notebook

NumPy
pandas

\[ y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it} \]
Material

- Web Site: http://didawiki.cli.di.unipi.it/doku.php/dm/start
- Laura Igual et al. Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications.
- Slides, Exercises and Notebook
Exam

• Project
  • Topics proposed during the classes
  • A single report to be sent periodically and one week before the oral exam
  • Groups composed of up to 3 people

• Oral
  • Short discussion of the project (group presentation, where possible), plus
  • Questions on all topics presented during the classes
  • Exercises and questions about all topics

\[
\text{DM2 Mark} = 0.6 \times \text{Oral} + 0.4 \times \text{Project}
\]
\[
\text{DM Mark} = \frac{\text{DM1} + \text{DM2}}{2}
\]
**Dataset**

**HAR: Human Activity Recognition Using Smartphones**

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The dataset for the project can be found at: [https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+s martphones#](https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+s martphones#)

- Detailed guidelines on the course webpage
Homework and Suggestions

Homework
• Declare Project Groups by next Thursday 24th February adding your information at https://docs.google.com/spreadsheets/d/1SuU8YLHKQcGvg4itG7xkpYKpyTJ77_bHQIVtsRN4_Hg/edit#gid=251564882

Suggestions
• Download and start to play with the dataset and perform data understanding.
• Use a Github repository for python and ipython files.
• Use a shared Overleaf project (LaTeX) for the report.
Questions?

riccardo.guidotti@unipi.it
francesco.spinnato@sns.it
Let’s start!