DATA MINING 2- Introduction

Riccardo Guidotti

a.a. 2020/2021

Classes

- Classes
  - Monday, 14-16 (academic?), MS Teams
  - Wednesday, 16-18 (sharp?), MS Teams

- Office Hours
  - Thursday, 15-17, Room 296 Dept. Computer Science
  - Appointment [DM2 Meeting] at riccardo.guidotti@unipi.it

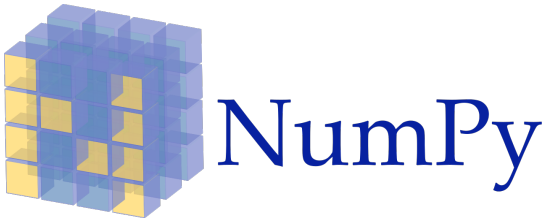
- Teaching Assistant
  - Salvatore Citraro [DM2 Meeting] at salvatore.citraro@phd.unipi.it

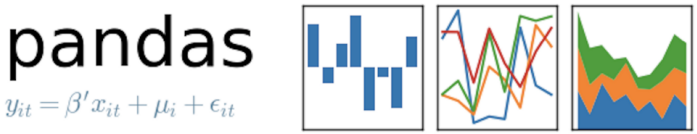
Topics

- **Module 1: Imbalanced Learning and Anomaly Detection**
  - CRISP
  - Evaluation
  - Imbalanced Learning
  - Anomaly Detection
- **Module2: Advanced Classification Methods**
  - Naive Bayes Classifier
  - Rule-based Classifiers
  - Logistic Regression
  - Support Vector Machines
  - Ensemble
  - Neural Networks
- **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

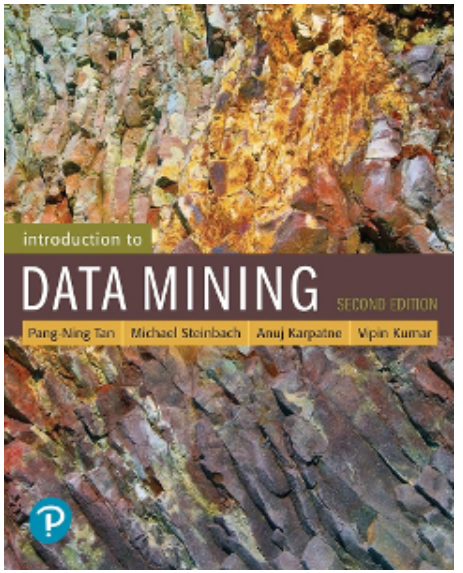
- Python
- Jupyter Notebook





Material

- Web Site: http://didawiki.cli.di.unipi.it/doku.php/dm/start
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. Addison Wesley, ISBN 0-321-32136-7, 2006, 2° Edition (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
- Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F. Guide to Intelligent Data Analysis. Springer Verlag, 1st Edition., 2010. ISBN 978-1-84882-259-7
- 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

DM2 Mark = 0.6*Oral + 0.4*Project

DM Mark = (DM1 + DM2) /2

Dataset

FMA: A Dataset For Music Analysis Data Set

Audio track (encoded as mp3) of each of the 106,574 tracks. It is on average 10 millions samples per track. Nine audio features (consisting of 518 attributes) for each of the 106,574 tracks. Given the metadata, multiple problems can be explored: recommendation, genre recognition, artist identification, year prediction, music annotation, unsupervized categorization. The dataset is split into four sizes: small, medium, large, full.

- The dataset for the project can be found at: https://github.com/mdeff/fma
- Detailed guidelines on the course webpage

Homework and Suggestions

Homework

- Declare Project Groups by next Monday 22° February adding your information at https://docs.google.com/spreadsheets/d/1RaAocJ2bCjCOYj4R068Rg6OLNNVmIG8YXu9puGIC1OU/edit?usp=sharing
- **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

salvatore.citraro@phd.unipi.it

Let’s start!