DATA MINING 2- Introduction

Riccardo Guidotti

a.a. 2021/2022

UNIVERSITÀ DI PISA

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

Topics

- **Module 1: Imbalanced Learning, Dimensionality Reduction and Anomaly Detection**
  - CRISP
  - Dimensionality Reduction
  - Imbalanced Learning
  - Anomaly Detection

- **Module2: Advanced Classification Methods**
  - Naive Bayes Classifier
  - Linear and Logistic Regression
  - Support Vector Machines
  - Neural Networks
  - Ensemble
  - Gradient Boosting
  - Rule-based Classifiers

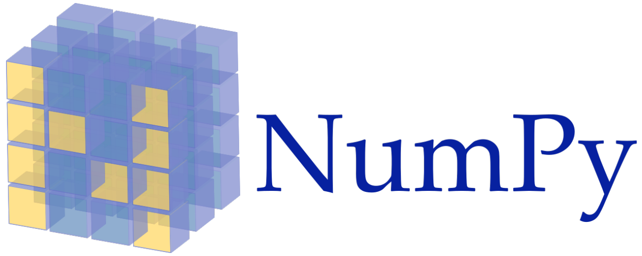
- **Module 3: Time Series**
  - Similarity
  - Approximation
  - Motif, Shapelets
  - Classification, Clustering

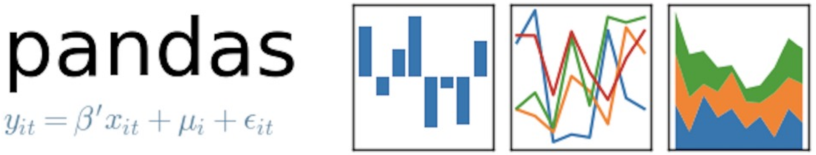
- **Module 4: Sequential Patterns and Advanced Clustering**
  - Sequential Pattern Mining
  - X-Means, OPTICS
  - Transactional Clustering

- **Module 5: Ethic Principles**
  - Explaianbility

Laboratory

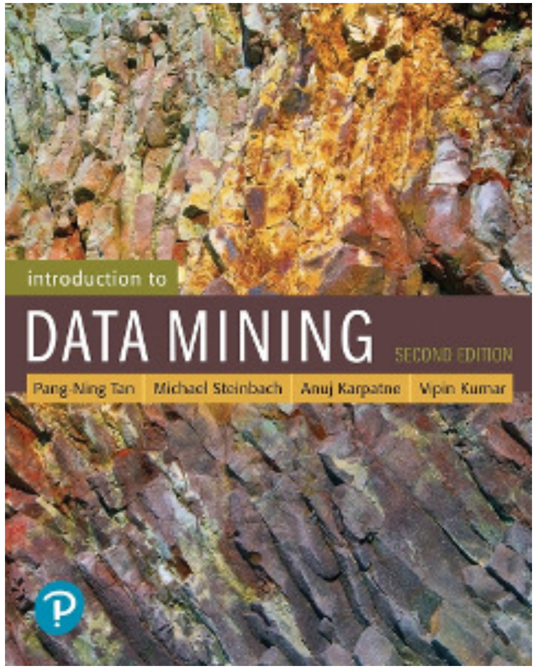
- Python
- Jupyter Notebook





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

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+smartphones#

- 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

https://www.wooclap.com/DMSURVEY

Let’s start!