DATA MINING 2

Course Overview

Riccardo Guidotti

UNIVERSITÀ DI PISA

Teachers

- **Riccardo Guidotti**
  - Computer Science Department
  - Email: riccardo.guidotti@unipi.it

- **Francesco Spinnato (Assistant)**
  - Scuola Normale Superiore
  - Email: francesco.spinnato@sns.it

Classes

- Classes
  - Monday, 09-11 (academic), Room Fib C1
  - Tuesday, 09-11 (academic), Room Fib C1

- Office Hours
  - TBD
  - Appointment [DM2 Meeting] at riccardo.guidotti@unipi.it

- Teaching Assistant
  - Francesco Spinnato [DM2 Meeting] at francesco.spinnato@sns.it

Topics

DM1

- Introduction to Data Mining
- Data Understanding
- Data Preparation
- Clustering
- Foundations of Classification
- Association Rule Mining
- Sequential Pattern Mining

DM2

- Imbalanced Learning
- Dimensionality Reduction
- Anomaly Detection
- Advanced Classification/Regression
- Time Series Analysis
- Transactional Clustering
- Explainability

Topics

- **Module 1: Advanced Data Pre-processing**
  - Imbalanced Learning
  - Dimensionality Reduction
  - Anomaly Detection
- **Module 2: Advanced Classification & Regression**
  - Logistic Regression
  - Support Vector Machines
  - Neural Networks
  - Ensemble Methods
  - Gradient Boosting
  - Rule-based Classifiers
- **Module 3: Time Series Analysis**
  - Time Series Similarity
  - Approximation
  - Motif, Shapelets
  - Classification, Clustering
- **Module 4: Transactional Data & XAI**
  - Sequential Pattern Mining
  - Transactional Clustering
  - Explainability

Laboratory

- Python
- Jupyter Notebook

python™

ANACONDA®

NumPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

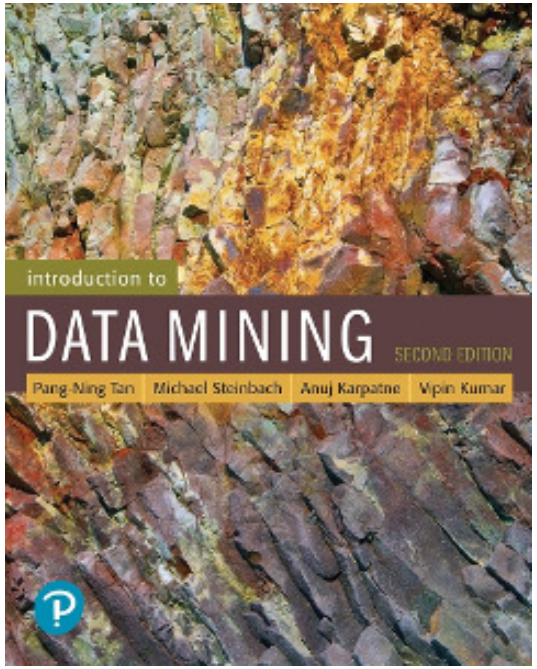
scikit
learn

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 presented during the classes
  - A single report to be sent periodically and one week before the oral exam
  - Groups composed of up to 3 people (DM1), people (DM2)
- Oral or Written Exam
  - Short discussion of the project (group presentation, where possible), plus
  - Questions on all topics presented during the classes
  - Exercises and questions about all topics

DM1 Mark = 0.6*Oral + 0.4*Project
DM2 Mark = 0.6*Oral + 0.4*Project
DM Mark = (DM1 + DM2) /2

Dataset - RAVDES

- **Ryerson Audio-Visual Database of Emotional Speech**
- Song (RAVDESS) contains audio of 24 professional actors (12 female, 12 male), vocalizing two statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.
- The dataset for the project can be found on the web page of the course.
- Detailed guidelines for the project will be presented next lecture and made on the web page of the course.

Dataset - RAVDES for DM2

1. RAVDESS FEATURES

- Features were extracted from the 2452 wav files. **Features are extracted by dividing each time series into 4 windows** instead of from the entire time series like for the DM1 dataset. There are now 434 features.

2. RAVDESS TIME SERIES

- Contains the original time series data (raw audio waveforms).

Homework and Suggestions

Homework

- Declare Project Groups by next Tuesday 28th 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!