DATA MINING 2
Introduction

Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

Classes

- Classes
  - Monday, 09-11 (academic?), Room C
  - Wednesday, 16-18 (sharp?), Room C1

- 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

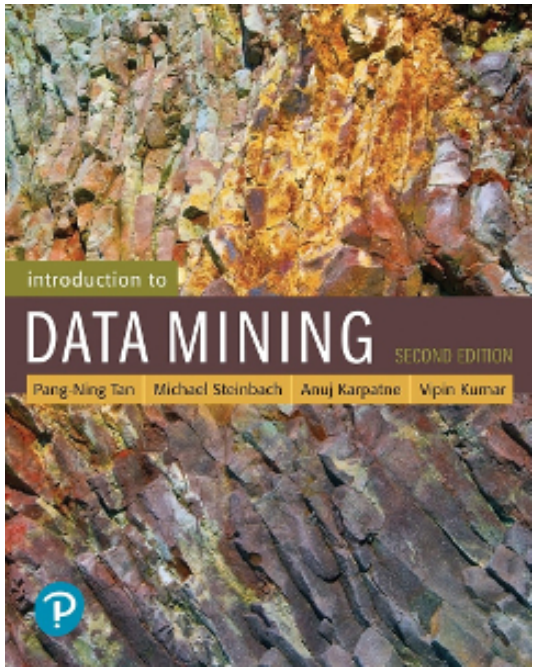
- Basic Classification Methods
  - Instance-based
  - Naive Bayes
  - Linear and Logistic Regression
  - Imbalanced Learning
  - Dimensionality Reduction
- Advanced Classification Methods
  - Support Vector Machines
  - (Deep) Neural Networks
  - Ensemble Classifiers
- Time Series
  - Distances and Clustering
  - Forecasting, Classification
- Sequential Patterns
  - Definitions
  - Mining
  - Constraints
- Outlier Analysis
- Advanced Clustering Methods
  - Expectation Maximization
  - Transactional Clustering
- Ethics Principles
  - Privacy
  - Explainablity

Laboratory

- Python
- Jupyter Notebook
- Ad-hoc Tools
  - SPFM (sequential patterns)
  - ELKI (outlier detection)

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

DM2 Mark = (Written + Project)/2 $\pm$ Oral

DM Mark = (DM1 + DM2) /2

- Written
  - Continuous assessment with 5 periodical tests during the course
  - Exercises and questions about all topics
- 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

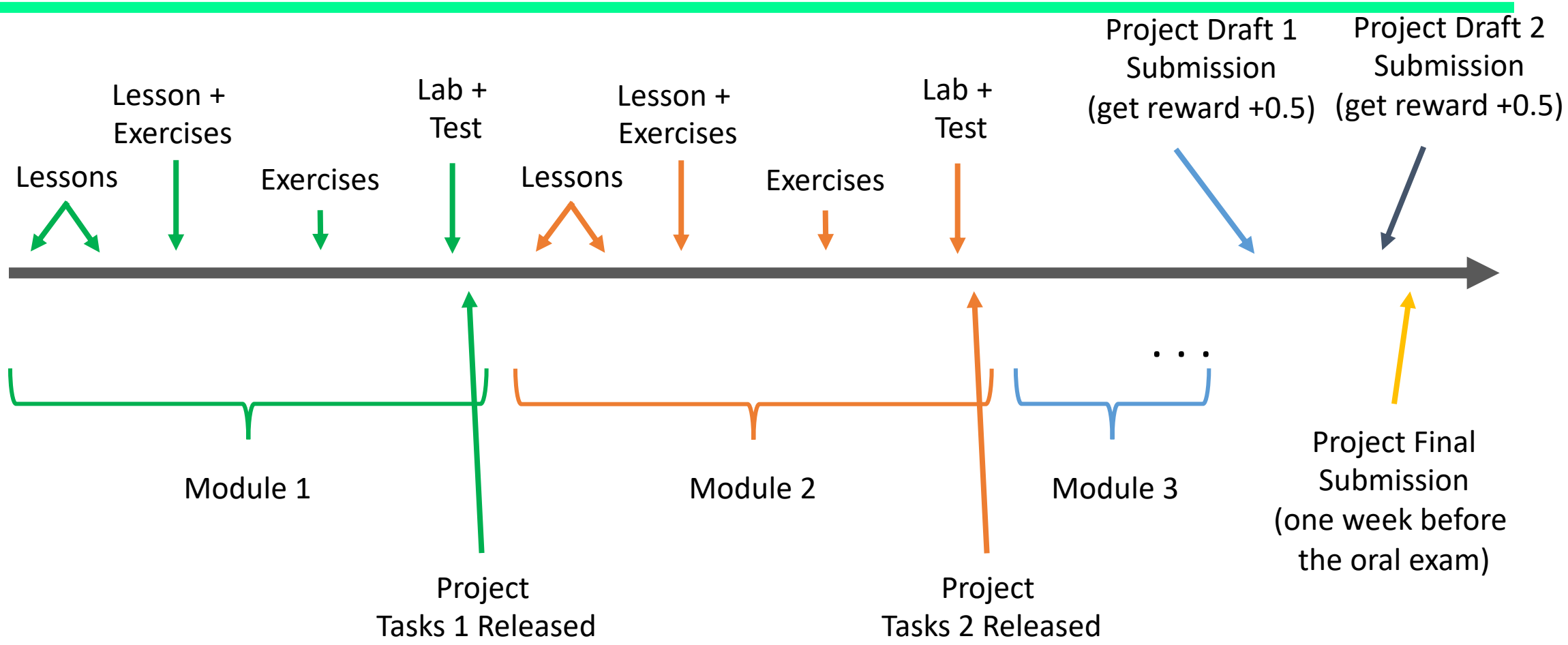
Exam Options

- Case 1 – Classic
  - Project submitted one week before the oral exam
  - Written exam at exam sessions
  - Oral exam for the theoretical aspects and for the project
- Case 2 – Recommended
  - Project submitted periodically and last version one week before the oral exam
  - Replace written exams and oral theoretical aspects with continuous assessment tests if sum of scores >= 18.
  - Oral exam *only* for the project
- Notes
  - If you fail periodical submission you can still follow case 2 but you lose the reward.
  - If you fail continuous assessment you have to move to case 1
  - If you go for case 2 to you *cannot* ask for oral theoretical aspects to improve the score because theoretical aspects are tested with the continuous assessment

Continuous Assessment Rules

- 5 Tests, one for each module of the course
- 6/7 questions for each test.
- Different tests generated randomly with shuffled questions and answers.
- 30 minutes per test at the end of regular lectures including exam preparation and deliver, i.e., 20/25 minutes for the exam
- Admitted usage of calculator:
  - no smartphone
  - if you do not bring the calculator you can do the math using your fingers
- The test is super easy thus the rules are very strict:
  - You take zero at the test if your are surprised to talk with others, look into others exams, copy from others, suggests to others, use your smarthphone
  - You cannot leave the room during the exam but you have to wait until the end for not creating noise in the class.

Course Timeline



Dataset

Occupancy Detection Data Set

- Experimental data used for binary classification (room occupancy) from Temperature,Humidity,Light and CO2. Ground-truth occupancy was obtained from time stamped pictures that were taken every minute.
- The dataset for the project can be found at: http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

Homework and Suggestions

Homework

- Declare Project Groups by next Monday 24° February adding your information at https://docs.google.com/spreadsheets/d/1_57y5ELInFsCFkaVrf0_rhMm3K1wvnIwKFgSZcXFukQ/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!