

Projects

Sandro Cumani

sandro.cumani@polito.it

Politecnico di Torino

The projects consist in binary classification tasks

There are 3 tasks (you have to choose one):

- Pulsar detection
- Wine quality detection
- Gender recognition from speech

Projects

For each dataset you have a `Train.txt` file that contains training data and a `Test.txt` file that contains test data

Each row of the data files corresponds to a sample

Features are separated by commas

The last column is the class label (0 or 1)

File `INFO.txt` contains the description of the features and additional information

Additional information may be present in the project folder

The protocol is the same for all datasets, and **MUST** be followed:

- Train data can be used to estimate model parameters and for cross-validation
- Validation sets can be extracted from the **training data only** (possibly using K-fold cross-validation if needed)
- The test set can be used **exclusively for evaluation**. All samples must be evaluated, and the score of a sample **must not depend** on the values of other test samples.

Pulsar detection

The first project requires identifying pulsars among pulsar candidates

The dataset is taken from the UCI repository¹

There are 8 features, that represent statistics extracted from radio signals collected by radio telescopes.

Classes are highly imbalanced.

Expect low error rates (in the order of few percent points or lower)

¹Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science

Wine quality detection

The second project requires discriminating between good and bad quality wines

The dataset is also taken from the UCI repository

The original dataset consists of 10 classes (quality 1 to 10)

For the project, the dataset has been binarized, collecting all wines with low quality (lower than 6) into class 0, and good quality (greater than 6) into class 1

Wines with quality 6 have been discarded to simplify the task

Wine quality detection

The dataset contains both red and white wines (originally separated, they have been merged)

There are 11 features, that represent physical properties of the wine

Classes are partially balanced

Classes are harder to separate than in the previous project, expect higher error rates (in the order of ten percent or more)

Gender identification

The third project consists in gender identification from high-level features

The dataset consists of male and female speaker embeddings

A speaker embedding is a small-dimensional, fixed sized representation of an utterance

Features are continuous values that represent a point in the m -dimensional embedding space

The embeddings have already been computed, you have to classify them

Gender identification

Each sample consists of 12 continuous features

Classes are balanced

To make the problem more tractable², and to avoid potential privacy issues, the dataset consists of synthetic samples that behave similarly to real speaker embeddings

The dataset is fairly easy to classify, expect error rates in the order of some percent points

²Real embeddings usually have from 200 to 500 dimensions

You should select a task

Analyze the problem, the kind of features, their ranges, their distributions, ...

Devise suited approaches for solving the classification task (refer to the models we discussed during lectures and laboratories)

Employ the training data as you deem appropriate (model training, hyper-parameter tuning, ...)

Evaluate different approaches on the test set following the proposed protocol

All the steps should be detailed in the report

Perform an analysis of the obtained result

- How good are the different models?
- Which model performed better?
- Was it expected? Why?
- Which choices made during training were good, which were sub-optimal? (e.g. did you select good values for hyper-parameters? Could different values had led to better performance?, ...)
- What are the trade-offs of different error types?
- ...

Conclude the report summarizing your findings

Summarizing, the report should contain

- A description of the problem and of the dataset, together with an analysis of the dataset features
- An analysis of different methods that can solve the problem
- A comparison of the effectiveness of the different methods
- A critical analysis of the results

Even if some techniques do not work well, you are encouraged to add them, with a justification for the poor performance.

You also have to **provide the code** that you used to implement the different algorithms (Python)

Since this course presents the basis of Machine Learning, **avoid** using ML libraries or ML toolboxes for the project (using toolboxes will result in lower marks — one of the goal of the course is that you learn how to implement the approaches)

The laboratories are already organized as to allow you to implement many of the techniques that we will discuss

You can, of course, re-use the code developed during the labs (including snippets provided by us)

If you are in doubt whether you can use some library or not, **ASK**

If you want to participate to an exam session, you have to submit the report by the official exam date

Oral examinations will start 7 – 10 days after, the timetable will be provided through the teaching portal

The report must be submitted through the teaching portal, section “Work Submission” (Elaborati)

The format should be a .zip file, containing

- The report in pdf format
- The source code (python source files, no jupyter notebook or similars)

The file name should be <student id>_<exam-date>.zip

Projects

Projects may be done in groups of up to 2 people

For group projects, both authors must upload their own .zip file

The report must contain the names of **both** authors

You can keep the same report for different oral sessions

The report mark will be disclosed during the oral exam

If you fail or withdraw from an oral session, you can also upload a new report if you wish to improve the report mark:

- If you withdraw before getting the report mark, you can re-submit the report on the same task
- After the report mark has been disclosed (to you or to your team mate), if you want to improve the mark you have to submit a report on a **different** task

The oral examination will include a discussion of the project, and questions that can cover the whole program