Projects

Sandro Cumani

sandro.cumani@polito.it

Politecnico di Torino

Projects

The projects consist in binary classification tasks

There are 4 tasks (you have to choose one):

- Gender identification
- Language detection
- Fingerprint spoofing detection
- Biometric identity verification

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)

Each project folder contains a pdf file with additional information

Projects

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.

Gender identification

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

The dataset consists of male and female speaker embeddings extracted from face images

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

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 not balanced

To make the problem more tractable[1], and to avoid potential privacy issues, the dataset consists of synthetic samples that behave similarly to real embeddings

The primary metric for the dataset is normalized actual DCF with $(\pi_T = 0.5, C_{fn} = 1, C_{fp} = 1)$

Expect costs in the order of 0.1 (accuracy in the order of 10%) or less

Refer to the description of the project for more information

NOTE: The task is different from the gender detection task of last year (the 2022 gender detection task cannot be submitted anymore)

[1] Real embeddings usually have few hundreds dimensions

Language detection

The second project consists in a language detection task

The goal is to detect whether an utterance is spoken in a target (fixed) language

The dataset consists of utterance embeddings extracted from audio sources

An utterance embedding is a small-dimensional, fixed sized representation of a speech segment

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

Language detection

Each sample consists of 6 continuous features

Classes are not balanced, the target language (class 1) has significantly less samples

As for the previous project, also in this case the dataset consists of synthetic samples that behave similarly to real embeddings

The primary metric for the dataset is the average of two normalized actual DCF, corresponding to $(\pi_T = 0.5, C_{fn} = 1, C_{fp} = 1)$ and $(\pi_T = 0.1, C_{fn} = 1, C_{fp} = 1)$

Expect costs in the order of 0.3 (accuracy in the order of 10%) or less

Refer to the description of the project for more information

Fingerprint spoofing detection

The third project consists in a fingerprint spoofing detection task

The goal is to detect whether a fingerprint image is authentic (class 1) or spoofed (artificially created replica of a fingerprint, class 0)

The dataset consists of embeddings extracted from fingerprint images

An embedding is a small-dimensional, fixed sized representation of an image

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

Fingerprint spoofing detection

Each sample consists of 10 continuous features

Classes are slightly imbalanced

As for the previous project, also in this case the dataset consists of synthetic samples that behave similarly to real embeddings

The primary metric for the dataset is actual DCF with $(\pi_T = 0.5, C_{fn} = 1, C_{fp} = 10)$

Expect costs in the order of 0.3 (accuracy in the order of 5%) or less

Refer to the description of the project for more information

Biometric identity verification

The fourth project consists in a biometric identity verification task

The goal is to detect whether a pair of embeddings belongs to the same person

The dataset consists of pairs (trials) of embeddings extracted from audio sources

An 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

Biometric identity verification

Each sample consists of 10 continuous features (5 for each embedding of the pair)

Classes are slightly imbalanced

As for the previous project, also in this case the dataset consists of synthetic samples that behave similarly to real embeddings

The primary metric for the dataset is actual DCF with $(\pi_T = 0.1, C_{fn} = 1, C_{fp} = 1)$

Expect costs in the order of 0.3 (accuracy in the order of 5%) or less

Refer to the description of the project for more information

Report

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, calibration, ...)

Report

Analyze the performance of the different models on a validation set:

- How good are the different models?
- Which model performs better?
- Was it expected? Why?
- What are the trade-offs of different error types?
- ...

Select a candidate model

Report

Evaluate the candidate model on the test set following the proposed protocol:

- How well does the model perform?
- Are the scores well-calibrated?
- How would the model perform for different applications?

Perform a (post-)analysis on the evaluation set, evaluating also alternative models that were analyzed in the previous stage:

- How do the results of the selected model compare with those on the validation set?
- How good were the choices made during train (hyperparameters, pre-processing, ...)?
- How good would alternative different models have been?

Report

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

The project pdfs provide additional information on the methodology

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)

Projects

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

Projects

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

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 similar)

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

Projects

You can keep the report mark for different exam sessions, up to the February 2024 session (included)

Insufficient or rejected project marks:

- If you submit a report but withdraw before getting the report mark (by sending an email after submitting but before the disclosure of the report marks), you can re-submit the report on the same task
- After the report mark has been disclosed (either to you or to your team mate), if you want to (or must, if the report mark was insufficient) improve the mark you have to submit a report on a different task

For more information, refer to the slides `0_Intro.pdf` available on the portal