Machine Learning Project
WS24/25

Machine Learning Group
Fakulty IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Milestone 2: Model Selection and Evaluation

Goal

The goal of this milestone is to evaluate the performance of algorithms of your project. As before, this sheet only serves as a coarse guideline and might not apply to your project. If you don't know how to proceed, please contact the course organizer.

How to submit

The report must be submitted via ISIS. Do **not** include code in your report. Reports that are longer than ten pages will not be accepted.

Put the code of your project in a folder `MS2` on the cluster (for detailed instructions see milestone 1).

Requirements

Implementation (1 point)

Write a function

```
def train_apply(method  = 'method_name',
                dataset = 'dataset_name',  ...)
```

that

- loads the data,
- trains the learning algorithm,
- returns the predicted labels for the test data.

Your function may have additional parameters, all of which must have default values. The function must be contained in a module `train_apply`. If the training phase includes model selection (e.g. using cross validation), then this must also be included in this function.

You do not need to implement the learning algorithms yourself. You may use public implementations and the scikit-learn toolbox.

Report (9 points)

Please include the following topics in your report. Each result should be *interpreted*.

Methodology (2 points)

Explain what you have done.

1. A short description of the methods (not more than two pages).
2. A description of the chosen pre-processing and feature selection (you can refer to your report from Milestone 1).
3. Explain the meaning of the hyperparameters and how you have chosen them (model selection). Do **not** simply copy values from literature but justify each choice.

Empirical estimate of the generalization error (4 points)

1. What is a good error measure for your particular problem?
2. Compute estimates of the generalization error using bar plots or violin plots, i.e. the expected error rates on new data (not used during training).
3. For a given error measure, showcase best, median and worst test cases for your method and baselines. Regarding the worst cases: what went wrong?
4. For most application, both the true positive rate and the false positive rate are important. Which trade-offs are achievable?

Discussion (3 points)

1. How expensive is it to train the classifiers? Is it possible to efficiently re-train the classifiers as new data becomes available?
2. Is it possible to obtain a confidence measure for the predictions of your algorithms? How reliable is it?
3. Draw a conclusion: are the learning algorithms suitable for this application? Which one is better?