Course name: Machine Learning, Blok4

Course code: 880083-M-6

Academic Year: 2022-2022

Çiçek Güven
Department of Cognitive Science and Artificial Intelligence
Tilburg University

Machine Learning Assignment

Introduction

For this assignment you will participate in groups of 4 students in a learning challenge. You will need to submit your **predictions**, as well as a report describing your **methods and results** and the **code** which you used to generate the predictions.

The assignment is worth 30% of your grade.

The assignment grade will be based on the quality of your work as judged by the instructor based on your report and code.

Alongside this assignment desciption a tabular data set will be provided where the target variable is named "transaction". You will using the other columns for learning the function that maps them to the transaction.

Report

Your report should be **3 page maximum** in **PDF format** and should include the following:

- Group number and name of all group members
- Description of your experiments and results including:
  - features used/not used
  - learning model and algorithm used
  - parameter tuning
  - description of method or system built to perform classification

- discussion of experiments run and performance of your solution (this may include analysis such as confusion matrix, accuracy per class, etc.)
- details of specification of the work done by group members
- Reference or appendices (if applicable)
- Flowchart, visualisations and the tables can go to the appendix, and you can refer to them from the report.

Code

Your code should be a plain python script (.py not a notebook) which can be run to generate your solutions. No data files should be included.

It is recommended the code is commented in such a way that alignment of the code and the report is easity visible.

Submission

You can submit your assignment via the canvas assignment submission. The submission must contain the following files:

- Report (.pdf)
- Code (.py)
- Predictions (.csv)

The grading detailed here is based on the report, but your code should be aligned with the report (e.g. when we review your code, we should be able to see the parts for HP tuning, training, etc.). Code without report or report without code will not be evaluated.
Make sure that you submit your .csv file for your predictions.
You will generate only one report (max 3 pages excluding references and appendix) You can include your figures and tables in the Appendix.

You can NOT compare a Neural Network with x hidden layer and a Neural Network with y hidden layer as two different models. In that case, your work will be evaluated for only using one model.

Group work

Your report needs to contain a detailed description of who did what, so make sure to keep track of this information.

Note: it is not acceptable to just say *All members worked together and contributed equally.*

If there are any problems with collaboration, such as serious disagreements, a group member not contributing, or a group dissolving, make sure to inform the course coordinator as soon as possible via email.

Code reuse rules

Remember this is a group assignment. You are **not allowed** to collaborate or share code with students outside your group. Submissions will be checked for plagiarism. If you are found breaking the above rules you will be reported to the Board of Examiners for fraud.

You are **allowed to use**:

- code examples provided by the instructor during the course
- open source libraries available for Python
- open source code found on Github, as long as it is credited in your script and report with a link to the source.

Tasks and Dataset

Binary Classification Task

In this assignment, you are asked to perform a binary classification task. You need to start with reserving 20% of the data for testing purposes.

You will be training 3 models of your choice and compare their performances. You can choose any two models (Neural Networks,logistic regression, Support Vector Machines (SVM), K-Nearest Neighbor, etc) for comparison. You will split your training data into training and validation sets to train your models and finding the optimal parameters, respectively. You will test the performance of the three models on the same test dataset that you reserved.

You will use:

- **training part** to train your model,
- **validation part** to find the optimal parameters of your model,
- **test part** to evaluate the model with the optimal parameters found on the validation set.

Method

Summarise the method(s) that will be applied to solve the problem on high level and motivate your choices (why these models, but not others for this problem).

Among the three models, which one is the baseline, do you have a preference among the models to compare to the baseline? Explain why, by comparing the models.

Illustrate the research methodology with a flowchart or similar visualization

There are four important restrictions on the method used:

- The method should be fully automatic, that is, re-running your code should re-create your prediction file.
- The method should not be hard-coded in any way. That is, if a separate set of data points are provided as input to the code, it should run and predict the labels without any changes in the code.
- Every software component used should be open-source and possible to install locally. This means that you cannot use proprietary closed-source software or access to a web service to carry out any data processing.
- The method should not use any external dataset which overlaps with the provided data. If you wish to make use of external data in your solution ask the instructor via the course forum to confirm that this data is allowed.

Grading

The models - 1.5 points:

Descriptions of the models, motivation, and comparison.

Briefly explain the model and why it is chosen for this task.

Data preperation- 1 point:

How did you split the data?
How do you represent your data? Have you applied any feature engineering (extraction/transformation)? If yes, explain what they are; if no, explain why not. How did you split your data into training/validation sets?

Model training - 3.5 points:

Which hyperparameters are tuned?
For each hyperparameter you tuned, what were the candidate values you tried?

Present the impact of using different HP (choose at least 2 (types of) hyper-parameters) values on the performance. Provide the specs of the final model chosen after HP tuning. Give any additional information about training (e.g. optimizers, loss functions, regularization, etc.) if there is any. Explain why you used/did not use these.

Results and Discussion of the results - 4 points:

Report performances and confusion matrix possibly with the aid of visualisations Comparison of the three models for their performances on the test set. Argue the transparency and reproducibility of your approach. Make a judgement on the performance ranking of competing models, based on robust evaluation metrics Do the models generalize well? Take a critical approach, review the strengths and weaknesses of the method implementation and current results.