Machine Learning
Programming Assignment II-b

Ujjwal Sharma, Athanasios Efthymiou and dr. Stevan Rudinac

The following assignments will test your understanding of topics covered in the first four weeks of the course. These assignments **will count towards your grade** and should be submitted through Canvas by **26.11.2020 at 08:59 (CET)**. You can choose to work individually or in pairs. You can get at most 5 points for these assignments, which is 5% of your final grade.

Submission

You can only submit a Jupyter Notebook (*.ipynb) for this homework. To test the code we will use Anaconda Python 3.7. Please state the names and student ID's of the authors (at most two) at the top of the submitted file. Please rename your notebook to add the name of the TA for the group you are assigned, e.g. assignment2b_{first_student_name}_{second_student_name}_{ujjwal_or_thanos}.ipynb. If you plan on working alone, please name the file as assignment2b_{student_name}_{ujjwal_or_thanos}.ipynb.

Instructions

- Alongside the code for your experiments, you are also required to present a report summarizing the observations and results of each of the experiments. You can use text and graphs/plots (`matplotlib`) for these reports. You should place these report blocks within the Jupyter Notebook in separate text cells. Plots can be appropriately placed near the text explanation. Your final submission should be a single Jupyter Notebook with code and report blocks.
- While it is perfectly acceptable to brainstorm and discuss solutions with other colleagues, please do not copy code. We will check all submissions for code similarity with each other and with openly-available solutions on the web.
- Please ensure that all code blocks are functional before you finalize your submission. Points will NOT be awarded for exercises where code blocks are non-functional.

1 Implementation Details

In this assignment, you will put into use two additional classifiers you have seen this week, the *Decision Tree Classifier* and the *Random Forest Classifier*. As before, you will be using the sklearn `Pipeline` functionality to encapsulate your preprocessing transformations as well as classification models into a single estimator. In the following assignments, you should perform preprocessing, model fitting and prediction operations only with a `Pipeline` estimator.

If you are asked to analyze an effect, we expect a written textual block summarizing your analysis. This can include plots if they strengthen your argument.

Any grid search should also be performed on the `Pipeline`, not on standalone estimators or transforms.

✉ u.sharma@uva.nl, a.efthymiou@uva.nl, s.rudinac@uva.nl

2 Data

For this assignment, you will continue to use the same dataset that was provided with Assignment 2A. If you are attempting this assignment without attempting part 2A, we would advise you to go back and look at the previous assignment for more information on the data.

3 Models

The model structure introduced in Assignment 2A uses a `Pipeline` to wrap a *scaler* and a *classifier*. For this assignment, you will use the same setup with two new tree-based classifiers. As such, you are only required to modify the classifier component of the pipeline.

In this assignment, you are asked to use the `DecisionTreeClassifier` and `RandomForestClassifier` as the primary classifiers. For these models, you must perform the following experiments:

1. Initialize a `DecisionTreeClassifier` model and perform the following tasks:
   (a) Fit your classifier on scaled as well as unscaled versions of the data and report the model scores.
   (b) Analyze the effects of data preprocessing operations like scaling on the classification performance of a decision tree classifier.
   (c) Are these models resilient to overfitting when model hyper-parameters have not been carefully selected? Supplement your explanation with suitable figures/tables if necessary.
   (d) Perform a grid search on the decision tree model parameters `['max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split']` to evaluate the optimal values for these parameters. Report model hyper-parameters for your best classifier model.
   (e) Report all evaluation metrics described in the Model Evaluation section below.

   In less than 50 words, explain your observations and your assessment of the classifier. You can use a text box (i.e. Markdown Cell) in Jupyter to write down your analysis.
2. Initialize a `RandomForestClassifier` model and perform the following tasks:
   (a) Fit your classifier on scaled as well as unscaled versions of the data and report the model scores.
   (b) Analyze the effects of data preprocessing operations like scaling on the classification performance of a random forest classifier and comment on its tendency to overfit to data.
   (c) Answer the following questions:
      i. How are decision tree classifiers different from random forests on a structural level? (max. 50 words)
      ii. Where would you choose decision trees over random forests and vice-versa? Demonstrate this using an appropriate example from your data. (max. 50 words)
   (d) Perform a grid search on the random forest model parameters `'max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split', 'n_estimators'` to evaluate the optimal values for these parameters. Report model parameters for your best classifier model.
   (e) Report all evaluation metrics described in the Model Evaluation section below.

   In no more than 50 words, explain your observations and your assessment of the classifier. You can use a text box (i.e. Markdown Cell) in Jupyter to write down your analysis.

4 Model Evaluation

For all models in 2A and 2B, you must report:

1. Classification metrics like accuracy, recall, F1 and micro/macro/weighted averaged precision/recall/F1 statistics.
2. ROC curves for each of the models.
3. Answer the following questions (with plots if needed):
   (a) Is accuracy an appropriate evaluation metric for this classification task? Justify in less than 20 words.
   (b) Can you select an optimal model for this task from the ROC curves? Justify in less than 20 words.

These metrics will be discussed at the beginning of Week 5.

5 Grading

Component	Points
Decision Trees + Analysis	2
Random Forests + Analysis	2
Evaluation Metrics	1