Sarah Schröder, Alexander Schulz Bielefeld University

Introduction to Machine Learning (WS 2023/24)
3rd Project

Released: Tuesday, 09.01.2024.
Due: Please solve the exercises in groups of three and submit your report and your code as an executable python script (`.py`) to Moodle by **Tuesday, 23.01.2024, 11:59pm**.

- Please note that the course language is **English**. However, you might hand-in your report in German as well as in English, but be consistent. We will not substract any points for errors regarding language as long as your report is understandable.
- We provide a LaTeX template `report_template.tex` which might help you.
- There is a Q&A tutorial for this project on Monday, 15.01.2024. You can use this session to work on this sheet and/or ask your tutor if you get into trouble.
- Submit your *pdf* and your code as *py*-file (you can export a py-file out of jupyter-notebook) to the LernraumPlus!
- If you use any code from the internet put a link to the source as an comment into your code for reference.
- The project will be discussed during the tutorials on Monday, 29.01.2024.
- If you have any questions, please ask your tutor or write an email to intromachlearn@techfak.uni-bielefeld.de.

In this project you will work on a real world data and two artificial datasets. There will be three parts and the report quality that will be graded.

Work on the tasks described below and write a **full-text** report on **max. seven** pages. Additionally, you can put all tables and figures, which do not fit into the page limit into the appendix. Please make sure to refer to them in your text. Your report should contain

- a description of the data sets used (now that you are partly working with real world data, your description might be a little more detailed here as you have information about what the features actually mean),
- descriptions of models and techniques used (unless they were explained in Project1 or 2),
- a documentation of your experimental set-up (What choices did you make? Why?),
- your results and an analysis (can you explain your results? what are the takeaways?),
- and, of course, a short introduction and conclusion.

When writing your report, please make sure that your descriptions are complete and precise. Based on your text we should be able to reproduce your pipeline and your results. Besides, you will have quite a lot of results. Consider one part of the task it to present them in a consist way. For example, you can consider visualizing your results in plots or creating tables.

When putting together your report pay attention to the structure. This is very important, as it is hard grasp work described in a badly structured report.

Hint: When creating plots using matplotlib, you can save the current figure by `plt.savefig('path/to/file.eps', format='eps')`. If you are using LaTeX for the first time `www.tablesgenerator.com` might be a helpful resource for easily creating clean looking tables.

1 Feature Selection

(*14 Points*)

In the lecture you learned about three types of feature selection. In the following, consider the F measure[1] for a filter, Sequential Feature Selector[2] as wrapper and Lasso as an embedded method.

Apply each of them to the real world data set from Project2 to select the two and the six most important features. Use one regressor of your choice as a baseline and the evaluation methods from the last project (including learning curves). Utilize the R2 score when reporting and analyzing your results. Also take a look at which features are selected and which are not by the different techniques.

2 Random Forest and Feature Importances

(*17 Points*)

(a) Train and evaluate with cross-validation a random forest classifier, and the other classifiers you know from the lecture on `dataset1.npz` (Naive Bayes, Logistic Regression, kNN).

(b) Visualize the data by plotting each combination of two features. Analyze the feature importances of the random forest with respect to the data. Rerun your experiments on a suitable subset of the features.

3 Challenge

(*14 Points*)

This exercise constitutes a challenge: Apply the learned concepts of the lecture and potentially think of new solutions in order to achieve an F1 score (use class 1 as 'positive' class) of at least 0.95 for the test set of `dataset2.npz`, without using it for training.

4 Report Quality

(*35 Points*)

In addition to the implementation, presentation and analysis of your results, we will score the completeness and overall quality of the report. Consider the bullet points from the first page.

5 Bonus Task - Clustering

(*10 Points*)

With this task you can gather bonus points (they are not included in calculating the threshold for passing).
Load the three given pictures and compress them by clustering the pixels of each picture according to their color values. Then, represent each pixel of a cluster in the mean color of the cluster. Visualize your results for different reasonable numbers of clusters using at least one evaluation score.
Hints:

- You do not need to actually compress the pictures so they use less disk space. Here, we only go the first step of that and assign new color values to each pixel.
- You can use the library `skimage (you might need to install it via pip install scikit-image)`
- Have a look into `loadAndDisplayImagesExampe.py`
- Calculation may need some time (up to a minute per picture)
- For the bonus task, you can add a separate chapter to your report. Briefly explain what you did and which methods you used, before presenting your results. You do not need to explain the methods.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html`

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html`