Johannes Brinkrolf, Valerie Vaquet Bielefeld University

Introduction to Machine Learning (WS 2020/21)
3. Assignment

Released: Friday, 08.01.2021.
Due: Please solve the exercises in groups of three and submit your report and your code as an executable python script to LernraumPlus by **Sunday, 31.01.2021, 11:59pm**.

- Please note that the course language is **English**. German hand-ins are not graded. However, don't worry: 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.
- Write all names of your group members and the tutor's name on the first page (have a look at given template).
- Mark each paragraph with the initials of the responsible group member. You find a new command `\initials{}`, therefore, in the template.
- There is a Q&A tutorial for this project on Monday, 25.01.2021. 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, 08.02.2021.
- If you have any questions, please ask your tutor or write an email to intromachlearn@techfak.uni-bielefeld.de.

1 Real word data

In this exercise you work with real word data. More precisely, it is a data set about house prices. Load the data with the field names `X`, `y` and `features`. `X` gives you the data matrix, `y` the corresponding outputs. You can find the feature names in `features`.

Your task is to apply and compare linear regression models, to investigate whether a feature selection is reasonable, and to write a short report. More precisely, work on the following tasks:

(a) Train at least three different regression models[1] and evaluate them with MSE and R2-score. As in the last past project, use cross-validation for evaluating your models.

(b) Try out pre-processing techniques.

(c) In the lecture you learned three types of feature selection (see slide 34 in slides3.pdf). Choose two of them and implement at least one method for each. Experimentally evaluate your feature selections.

(d) For the best performing combination of model and pre-processing steps, implement and plot the so-called learning curve on your own (see further explanation below).

Prepare your results in a **full-text** report on **max. three to five** pages. Your report should contain

- an introduction and a description of the given data,
- descriptions of the classifiers and the evaluation method you used,
- an overview of the three principles for feature selection, and a description of those methods you applied
- a documentation of your experimental set-up (what choices regarding models, pre-processing and feature selection did you make? Why?),
- a description, a comparison, and an analysis of your results,
- a discussion and a 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 result. 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.

[1] e. g. linear regression, ridge regression, Huber etc. Note, that the Laplacian regressor is not implemented in sklearn.

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.

Learning Curve: Learning curves show how well the model performs when trained with a growing amount of data samples. First of all, it is necessary to split the data set into a train and a test set. In order to compute the learning curve, we repeatedly train a model with a growing number of training samples (from the train set) starting by one and stopping with the number of training samples. We evaluate each of these models on the part of the training set which was already used for training, and on the entire test set. Finally, we plot the results over the number of samples used for training.