Assignment 10

Machine Learning, Summer term 2018
Norman Hendrich, Marc Bestmann, Philipp Ruppel
June 18, 2018

Solutions due by June 30 and July-08

Machine Learning Project, 10+20 points

Assignments 10 are 11 are the final assignment for this lecture. They are different from the previous assignments; you will not need to present any results in class. Instead, your task is to **investigate a data-set by yourself**, and to write two reports about your studies.

Note that you can still come to the exercises on 25/27 June and 02/04 July to discuss your project with the teaching assistants and other students. This is also a chance to present in class for those students who have not yet presented.

On 09/11 July the exercises will take place normally, so that we can discuss (some of) your ML project results.

Data-sets

We use data-sets from **Kaggle** (`www.kaggle.com`). "Kaggle is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models." (Wikipedia).

Please choose *one* of the data-sets available at Kaggle for your study. Note that some data-sets may need pre-processing (e.g. for textual features), while others can be used directly. Take your time for your decision: the data-sets involve different sub-problems (natural language processing, time-series, etc.) and the respective prediction problems vary in difficulty. Which data-set is a good choice depends on your personal interests and also on which software and toolbox you need to use. There is plenty of information on the data-sets together with helpful hints, scripts, and tutorials on the Kaggle forum and scripts pages.

To access the data-sets, one student per group needs to register at Kaggle. Some of the challenges are not active anymore but can be accessed in the section for completed challenges.

You can use whichever toolboxes, software libraries and programming or scripting languages you like. However, you must fully document all software that you used.

Assignment 10: Data Analysis, (10 points), due June/30

In the first part of your study, you have to analyze the basic characteristics of your chosen data-set. The goal is to sharpen your general understanding of the data. This is an **unsupervised** task: do not take into account too much the predictive modeling task of Assignment 11.

You have to write a report of up to **2000 words** about your study. Please structure your report as follows:

a. Short Introduction. Give a *short* overview over the data-set. Explain the features and their types.

b. Data Analysis. Describe your findings of the data analysis along with the *basic statistics* of the data-set: means, medians, variances, empirical probabilities etc. Also identify class imbalances, non-normalized features, and include statistics for appropriate subgroups.

c. Identify potentially *interesting features* and subgroups of features. Are the features which are correlated with each other? Can you identify groups of uncorrelated features, for example by clustering features with respect to the correlation coefficient as a similarity function? Which features seem to have a lot of explanatory or predictive potential?

d. Run at leat one *clustering* method on the data; or explain why clustering won't help on your data-set. If necessary, preprocess the data. You may focus on specific feature subgroups for clustering (this means that you only use some of the features to cluster your data). Interpet the found clusters and identify meaningful clusters.

e. Run at least one *dimensionality reduction* method; or explain why this won't work on your data-set. Can you simplify the problem with dimesnionality reduction methods? What might be a good choice for the diemnsion parameter?

f. Are there any other aspects of the data that you find interesting?

g. Conclusion. Provide concluding remarks about your findings. Describe what your findings suggest for the predictive modeling task in Assignment 11.

h. References. When you use special algorithms, techniques, or software libraries, provide references to the papers, yournals, and websites.

What to hand in: you need to hand in your report as a PDF file, together with your complete code for data analysis as a ZIP file to your teaching assistant by email. The file needs to make clear that you have done the data analysis yourself.

Assignment 11, Predictive Modeling and Report (20 points, due July/08)

Each Kaggle data-set comes with a well-defined predictive modeling tasks. In this assignment, you are asked to perform a study on the predictive modeling taks of your chosen data-set.

Please write a short report (again, up to 2000 words) about your study:

a. Short Introduction. Describe the prediction tasks and the evaluation metric used in the Kaggle competition. Use this evaluation metric in your study (this might be an unusual metric which you might need to implement yourself).

b. Predictive Modeling Steps. Describe your data preparation steps including normalization.

c. Describe clearly which data you use for training, validation, and testing, and how you optimize parameters.

d. Investigate at least three different *prediction methods*. For instance, for classification you might use SVMs, logistic regression, kNN, decision trees, etc. Among them must be at least one method which was presented in class. Also investigate and document different parametrizations (kernel type, distance measure, etc.)

e. Illustrate your results with adequate plots.

f. Discuss your findings: which model would you recommend to use? Are the overall results satisfactory? Could your predictive model be used in practive? Use a lot of skepticism and point to limitations, potential problems, and shortcomings in your approach. Can you hypthesize any causal effects? Or do important explanatory variables seem to be missing?

g. If possible, compare your restults to the results of other participants in the Kaggle competition.

h. Conclustions. Provide short concluding remarks about your findings. Which types of methods achieved the best results? Where would you continue future work?

What to hand in: your full report as a PDF file, plus the complete code and figures as a ZIP file. Again, this file needs to make clear that you have done thze predictive modeling yourself.

General Remarks

The report should be written in English (German only in exceptional cases). If possible, use LaTeX to write your report.

The goal is NOT to present the most competitive results (if you can do so, this is great, but we don't expect this). Instead, someone who reads your report should be able to understand what you did, judge whether your approach makes sense, whether your results seem correct, and evaluate your findings.

More importantly, the reader needs to be able to **reproduce** the results from reading only your report (without looking at your code). Therefore, you need to provide accurate descriptions of what you did, including details on the chosen methods, parameters, train and test data, etc.

Provide *meaningful* plots together with your data. Think well which are plots that are helpful to convey your story. There is no point to include a collection of 50 uncommented plots. Avoid tables, they are hard to digest.

Test data: Each Kagge data-set contains official test data. The correct test labels are not disclosed. You can evaluate your models on this data by making submissions to Kaggle. In your report, you have to provide results on the Kaggle test data. However, you can usually make only a limited number of submissions to Kaggle per day. Thus in case you require more test runs, it might make sense to split the training data into your personal training and test data, and use this personal test partition for further evaluation of your models. This allows you to perform several train/test splits and report results over multiple runs.

Use the same *evaluation metric* as used in the Kaggle competition. This might involve implementing it yourself.