ET003A/ET006A – Assignment: Non-NN Machine Learning

Sebastian Bader

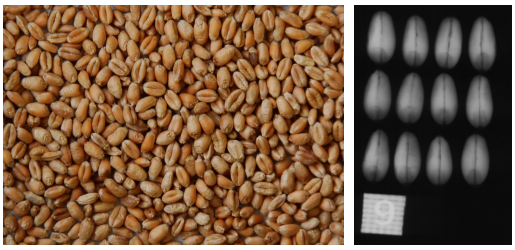
1 Background and aim

Machine learning is a great tool to build models based on data, rather than pre-defined rules and relations. However, it is normally not clear in advance what machine learning approaches will result in sufficiently good (or at the extreme, the best) result. Therefore, the field of machine learning builds often on trial and error, where multiple approaches need to be tested and evaluated.

The goal of this assignment is to train and evaluate a number of (non-neural network) classification models on a given data set. Through completion of this assignment, understanding and skills related to the following learning objectives are demonstrated:

- apply methods to import, combine, annotate and convert data to appropriate formats for data analysis.
- select, motivate and apply common methods and algorithms for machine learning for typical use cases and present the results in an appropriate way.
- design and perform performance validation for machine learning systems.

2 Task description



2.1 The data set

The data set used in this assignment contains a number of samples regarding three families of wheat (named Kama, Rosa and Canadian). Each sample has a label of which

family of wheat it relates to, as well as 7 features that are obtained from an x-ray analysis of the grain. The seven numeric (real-numbered) features are:

1. area (A)
2. perimeter (P)
3. compactness ($C = 4 * \pi * A / P^2$)
4. length of the kernel
5. width of the kernel
6. asymmetry coefficient
7. length of the kernel groove

In total the data set contains 210 samples, with 70 samples per class. The data set is provided as a csv file.

2.2 Matlab Classification Learner App

To start with, you can explore the provided data set with the Matlab Classification Learner App. The Matlab Classification Learner App is a tool with graphical user interface that quickly allows to test a number of classification models on a data set. Once you have imported the data set, you can try out a number of standard models in the App to identify suitable candidates for the next step. You can find some information on how to get started with the App here: https://se.mathworks.com/help/stats/classificationlearner-app.html

2.3 Implementation and evaluation

Based on your investigation in the Matlab Classification Learner App, select two promising models for further analysis in Matlab. Create a Matlab Live Script (.mlx file) that

- Imports and shuffles the data set
- Reserves a part of the data for final testing
- Trains the selected models using cross-validation
- Calculates the models' accuracy
- Optimizes hyperparameters of the models
- Evaluates the performance of the optimized model using the test set
- Plots confusion matrices

You are of course welcome to add other functionality to your script, which could include plotting of a group scatter matrix in order to investigate and visualize clusters, or an analysis on the influence of feature selection on the model performance.

Matlab Live Scripts can, in addition to code, also contain formatted text data. This allows us to use the script at the same time as a documentation of your work. Add reasonable descriptions to your live script to make it easy for an external reader to follow your work. In addition to documenting what you are doing, include explanations on

1. Why it is necessary to hold out on a separate test set that is not used until the very end.
2. What the differences is between cross-validation and a hold-out method of a validation set, as well as what advantages cross-validation brings.

You can find information about Matlab Live Scripts and their formatting alternatives here: https://se.mathworks.com/help/matlab/live-scripts-and-functions.html.

3 Examination and feedback

For examination of this assignment, the documented .mlx file needs to be handed in (there is an inbox on the moodle page). The assignment is graded on an A-F scale, where the following aspects will be taken into account:

- Correctness and completeness
- Logical structuring
- Clarity and format of documentation
- Demonstrated understanding of key concepts

Written feedback will be provided to each student in conjunction with the grading with respect to these aspects.