Machine Learning 2
Data Science

Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Version November 26, 2019

Project

The subject of the project is to compare the results of two different machine learning methods on a data set of your choice.

The project is not compulsory but it is worth 30% of the marks towards the course.

You will work in groups of 2 to 4 Students, analysing a medium sized data set using supervised learning. The deadline for handing in your work via Moodle is **Sunday 12th January 2020**.

Part of the project is to find a suitable data set. This should be in the standard matrix form (`data.frame`), have at least 500 observations and preferably between 1000 and 2000 observations. The data should be appropriate for a supervised learning problem, which can be either a regression or classification problem. The data should contain at least 5 predictor variables and an outcome variable. A Website with many suitable data sets is the *Machine Learning Repository* at the Center for Machine Learning and Intelligent Systems: `http://archive.ics.uci.edu/ml/datasets.php`

You will compare the results of two machine learning methods. At least one of the methods should be from Machine Learning 2, a list of acceptable methods is given below.

R will be your primary data analysis environment. If you choose to train a neural network on your data, then you may use alternative software to fit the data. The model evaluation and comparison should be done using *R*.

Your project must be your own work. Copying of external sources or other ML projects, including from previous semesters, will be treated as plagiarism, and could result in zero marks being awarded.

Split your data into three parts: 60% training data, 20% validation data and 20% test data. Your validation data can be used for finding the best hyperparameters and model choice etc. If you use cross validation to find the best hyperparameters you may combine the training and validation data.
Only use the test data to compare the two machine learning methods.

Your project report will include:

- a short description of your data,
- a mathematical overview of the two ML methods used,
- a description of your fitting process including, a summary of how you arrived at your final model, the choice of hyperparameters and how made this choice,
- an appropriate assessment of the predicted values and a fair comparison of the two methods.
- appropriate graphical presentation.

To submit your work, you should upload in Moodle:

- Your report in a standard format (such as PDF format),
- Your data set and
- Your *R* code in a script file. The *R* script should include code to read in the data, indicate the main parts using comments and run without errors.

Your two ML methods should be from the following list, one must be marked with ML2.

- Logistic regression (ML1)
- Linear (and/or quadratic) discriminant analysis (ML1)
- Ridge regression and/or the lasso (ML1)
- Tree models (ML1)
- One or more ensemble methods (ML1)
- Naive Bayes classification (ML2)
- Non-linear models (e.g spline smoothing) (ML2)
- Generalised additive models (ML2)
- Support vector machines (ML2)
- Projection pursuit regression (ML2)
- Neural networks (ML2)

Note that two bullet points should be chosen, so for example ridge regression and the lasso counts as one method.

It is a good idea to check with the lecturer that the data are appropriate before analysing the data. Each group should send an email to the lecturer on or before 11th December. State which students are in your group and where the data originates. A simple synopsis of the dataset should be included such as the number of observations, the variables with their data type (nominal, ordinal, discrete or continuous). Please indicate your chosen ML methods. When you send the email, have the data available to send if requested, in a form that can be easily read into *R*.

Good luck!