Machine Learning 2
Data Science

Winter Semester 2023/24 **Prof. Tim Downie**

Version November 21, 2023

Assessed Project

You will analyse a data set using two machine learning methods. You will compare the performance of these two methods on your data, and write up the results in a report.

The project is not compulsory but it is worth 35% of the marks towards the course. Students who have submitted a project in previous years do not have to submit another project the marks carried over to this semester. If you choose to work on a second project, it will be the marks from this year that will count.

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 **Tuesday 9th January 2024**.

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 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, that have been covered in ML 1 or ML 2. At least one of the methods should be from ML 2, a list of acceptable methods is given below.

R will be the data analysis environment including the model evaluation and comparison.

Your project must be your own work. You may use websites, books etc. for information about the theory and implementation of the methods. These sources should be referenced. Copying of external sources with respect to analysing your data set or copying of other ML2 projects, including from previous semesters, will be treated as plagiarism, and could result in zero marks being awarded.

Artificial intelligence software: you may use AI as a tool if you want to. Be aware that AI answers are quite often wrong or have inaccuracies. You are responsible for the content in your project, so unedited copy and pasting from such software is not recommended. Any section where you have used AI software should be referenced.

Regression or classification? You should choose whether to use regression methods or classification methods according to your outcome variable. If for example your outcome variable is income in US dollars, use regression methods rather than convert to the binary outcomes $\leqslant 50K\$$ and $> 50K\$$.

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 short 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 you made this choice,
- an appropriate assessment of the predicted values and a fair comparison of the two methods using the test data.
- appropriate graphical presentation.
- a bibliography of sources used, referenced in the main text.

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 from ML2.

- Logistic regression (ML1)
- Ridge regression and/or lasso regression (ML1)
- Tree models (ML1)
- One or more ensemble method (ML1)
- Non-linear models (e.g spline smoothing) (ML2)
- Generalised additive models (ML2)
- Linear (and/or quadratic) discriminant analysis (ML2)
- Naive Bayes classification (ML2)
- Support vector machine (ML2)
- Support vector regression (An extension SVM applied to regression data) (ML2)
- Projection pursuit regression (ML2)

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

You should check with the lecturer that your data are appropriate before starting the analysis. **Each group should send an email to the lecturer with the following information on or before Friday 8th December.** Include inn the email:

- which students are in your group,
- where you have found the data,
- 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).
- your chosen ML methods.
- Do not attach the data, but do have the file available to send, if requested, in a form that can be easily read into *R* as a *data frame*.

Good luck with you project!