IN-STK5000 Project 1 - Preliminary Report

Tobias Opsahl, Alva Hørlyk, Ece Cetinoglu, (Project 10)

September 2021

Firstly, we are going to do data exploring. We will read the files, name the columns, do some simple plots and look for correlations. After this initial phase, we will begin task 1. We will do this by hypothesis testing and exploring correlation between variables, interpreting the results and conclude. For the genes, we need to do variable selection. We will test multiple methods, and verify the method on synthetic data. This data will be produced in a way that it mimics the original data, but where we know the true effect of the genes on the response. We will make it so many of the genes are not generated to have an effect on the response, but add noise so the effect is not too obvious. We will try methods as forward and backward features selection and principal component analysis. We might also try and test shrinkage methods.

For the second task, we will try many models with different responses. The most obvious response is 'death', but we will probably try to look at other symptoms as well. We will look at regression, KNN and shrinkage methods. We will do crossvalidation on all of the models, and see if any of them are good at predicting. We will also try bootstrapping, and different form of bootstrapping evaluations. If the regression models end up being useful, we can try to explain the results.

For the third task, we will do a bit of research to try to find important fairness and privacy. We will look at similar studies and compare the precautions they have done.

In the end, we will write a detailed report. Here we will precisely inform what we ended up doing, why we did it, what the results were, how the methods works and how we verified them.

It is also worth to mention that we will probably get new ideas for models and strategies while working, so some parts of this outline might change.