Danmarks Tekniske Universitet

Report 2

Course:
02450: Machine Learning

Author(s):

Anders H. Opstrup (s160148)
Gu Jinshan (s161944)
Huayu Zheng(s162077)

DTU Compute
Institut for Matematik og Computer Science

02450: Machine Learning
Tue Herlau

8 November 2016

Contents

1 Introduction

As described in the conclusion of the first report the group have had some concerns about the dataset (forest fires). Therefor the group decided to change the dataset. The dataset was too small to make any interesting analysis. The group have chosen the SPAM dataset from http://statweb.stanford.edu/ tibs/ ElemStatLearn/. This dataset has around 50 attributes, compared to 12 the forest fires dataset has and the SPAM dataset has around 5000 records where the forest fires dataset only has around 500.

This report covers the second assignment for course *02450: Introduction to Machine Learning and Data Mining*. The assignments purpose is to solve a relevant regression problem and classification problem for our dataset. The group decided to predict the variable $word_freq_credit$ which is the frequency of the word credit used in an email, which could be an indication of SPAM mail. For the classification problem the group decided to identify SPAM from not-SPAM.

2 Regression

Frequent use of the word credit could been an indication of spam, the goal for the regression problem is to predict the variable word-freq-credit.

2.1 Applying linear regression to the data set

All the parameters have been normalized, before doing the sequential feature selection and linear regression. A 10-fold cross validation with a 10-fold internal cross validation method have been used for selecting the features for the model. The figure below (figure 1.) shows the selected features at each fold.

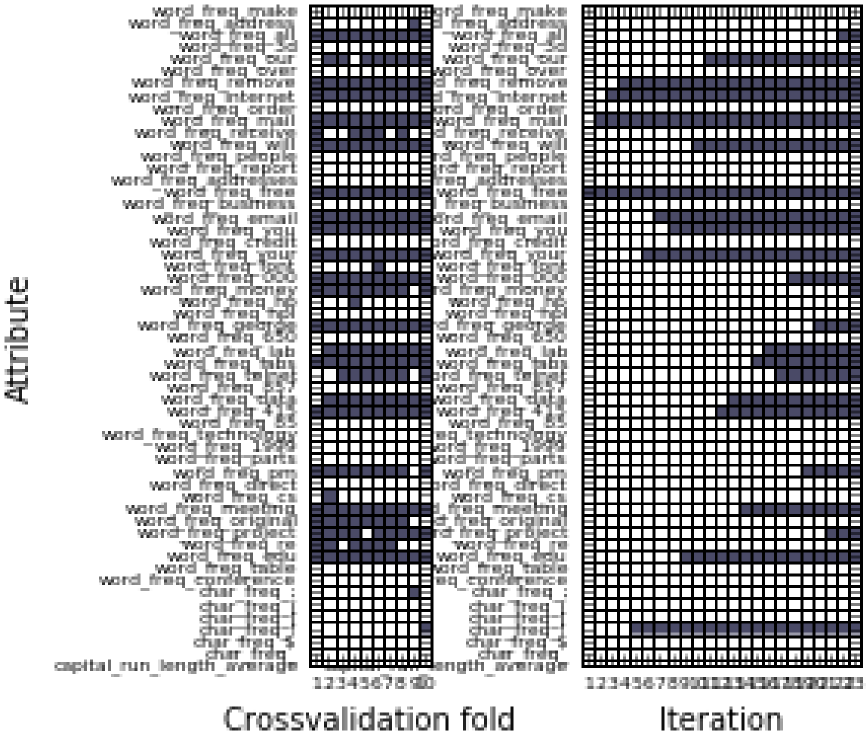


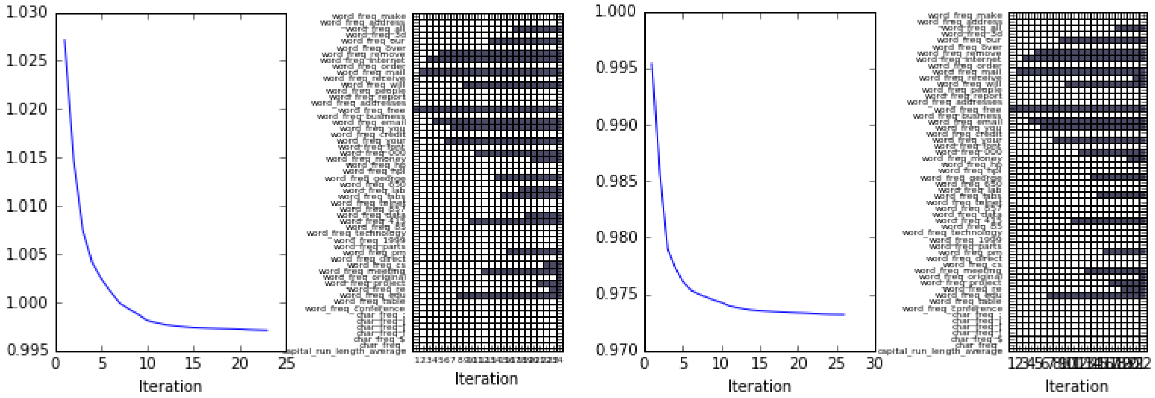
Figure 1: Selected features

By doing feature selection we reduce the amount of data, by only selecting the attributes which correlates to out regression problem. This makes modelling the data less prone to over fitting and reduces the overall complexity of the model. However when comparing the regression errors with feature selection and with out feature selection, the difference seems almost indistinguishable see table (figure 1.), thus the model with feature selection is far less complex as described.

Measure	Without feature selection	With feature selection
Test error	0.938505700248	0.940197192727
Training error	0.923006656194	0.928276304747

Table 1: Error difference between with and without feature selection

The figure below (figure 2) shows the regression for the first fold and for the 10th fold. Here it illustrates how the squared error gets lower and lower for each training session we do with the model.



(a) First fold

(b) 10th fold

Figure 2: First fold compared with 10th fold of the linear regression

2.2 Artificial Neural Network

An Artificial Neural Network (ANN) have been trained to solve the regression problem. For ease of comparison with the linear regression a 10-fold cross validation has been used as well. The network was trained with two hidden units and five networks trained for each fold. With a mean-square error of 0.913393150004, the network does preform a bit better than the linear regression but not significantly better. As seen below (figure 3.) the network does preform quite well for some folds but after some folds it makes some errors again, this could be a sign of possibly gain in efficiency if the network gets optimised a bit more, but here we need to keep in mind the danger for over fitting the model.

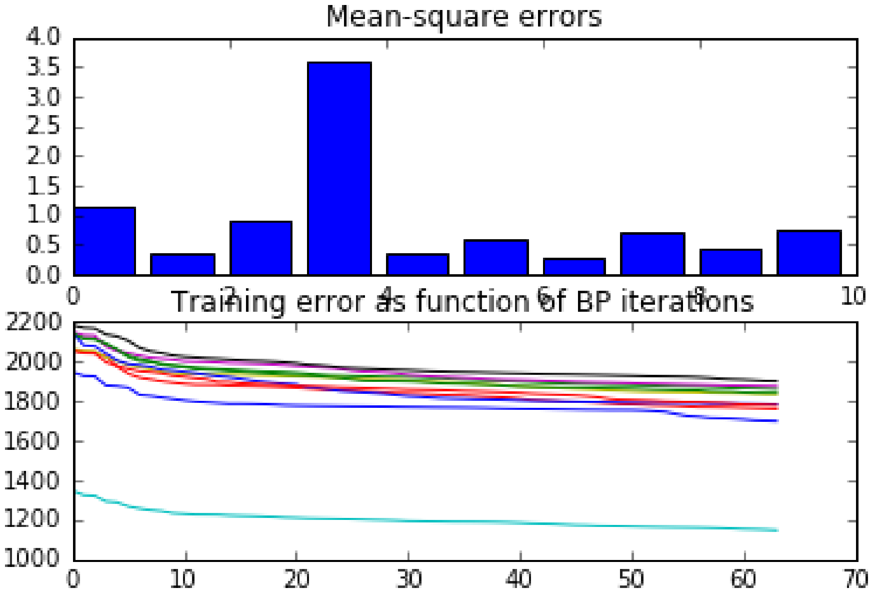


Figure 3: Errors for each fold of the ANN

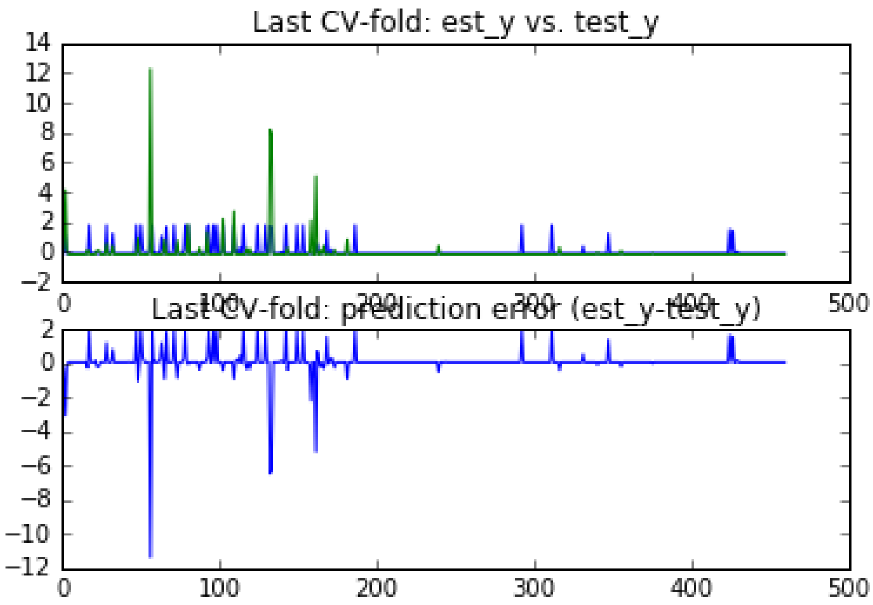


Figure 4: Comparing estimated values and test values, from the ANN

2.3 Comparing Artificial Neural Network with Linear regression

For comparing the two methods of regression a series of t-tests have been run on the results from each method. The mean squared errors of the ANN and the mean squared test errors from the Linear Regression was used for the t-tests. With a t-value of 2.89173 and corresponding p-value of 0.01783, there is a strong indication that the Artificial Neural Network is a better fit for regression on the dataset.

3 Classification

The 58th attribute was used to solve the classification problem for the dataset, which was detecting spam from non-spam. The 58th attribute is binary and indicates spam or non-spam, spam = 1 and non-spam = 0.

Some earlier classification analysis have been done to the dataset before with a classification rate 7%. The methods for obtaining this rate is unfornatly unknown.

3.1 Classification on the dataset

The group decided to analyse the data with a decision tree, K nearest neighbour and an artificial neural network. Furthermore a two level cross validation was used for selecting parameters, a 10-fold was used for both outer and inner fold.

3.1.1 Decision Tree

The parameters for the decision tree was based off a tree depth ranging from 1-20. Below (figure 5.) shows a boxplot for the optimal levels for each fold.

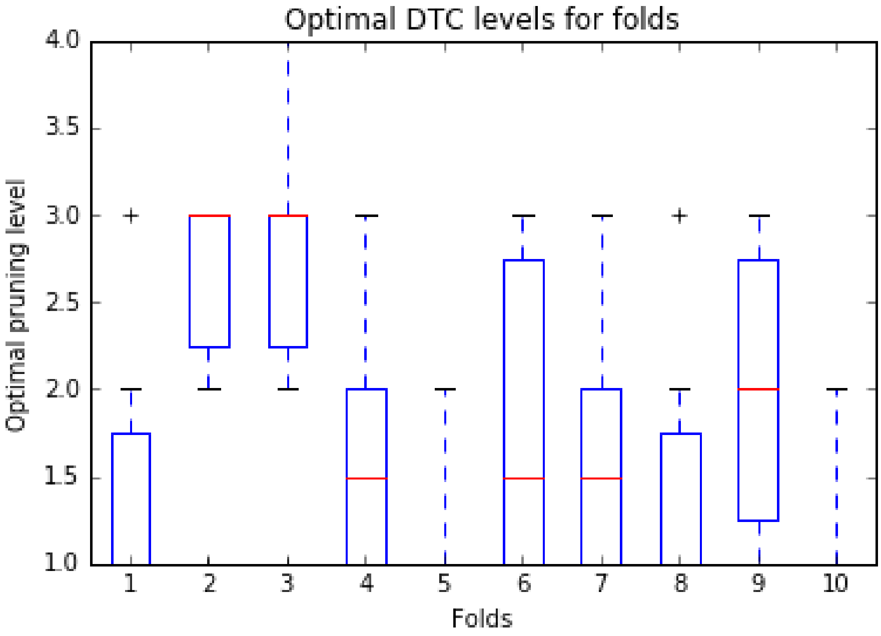


Figure 5: Boxplot for the decision tree

3.1.2 Artificial Neural Network

The artificial neural network was trained with 5 networks per fold, with at 10 fold cross validation. The network gave a good low test error rate at 6.88955955862%. Below is the result of the ANN illustrated.

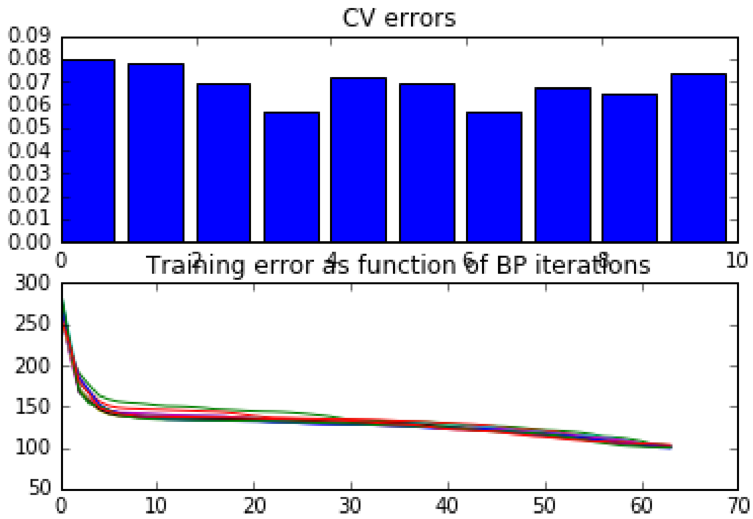


Figure 6: Errors for each fold of the ANN

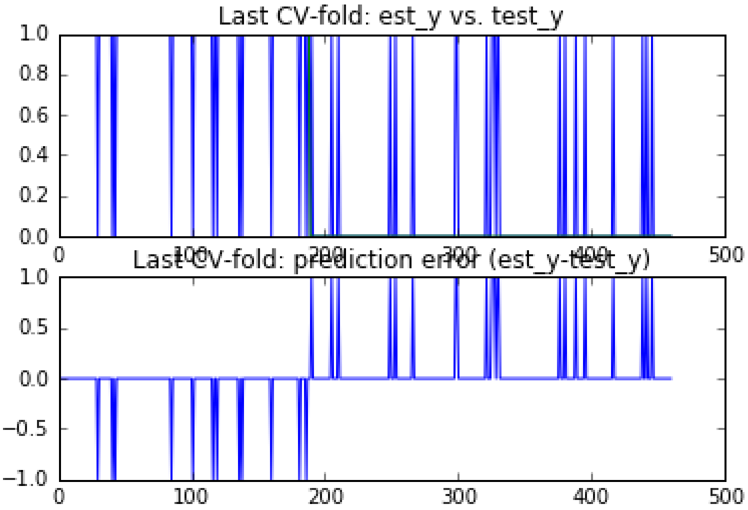


Figure 7: Comparing estimated values and test values, from the ANN

3.1.3 K-Nearest Neighbour

The K-Nearest Neighbour analysis was carried out in the same way as the decision tree, below (figure 8.) is the result of the analysis with KNN. The best error rate of the KNN analysis was 8.83459555812%.

3.1.4 Comparing the K-Nearest Neighbour with the Artificial Neural Network

For comparing the two best preforming methods a series of t-test have been carried out on the result for each method. The t-value of the test was 5.4900656372629992 and the p-value was 0.00038510731768533901. With the result of the t-test, the Artificial Neural Network is the preferred method for the classification problem on the dataset.

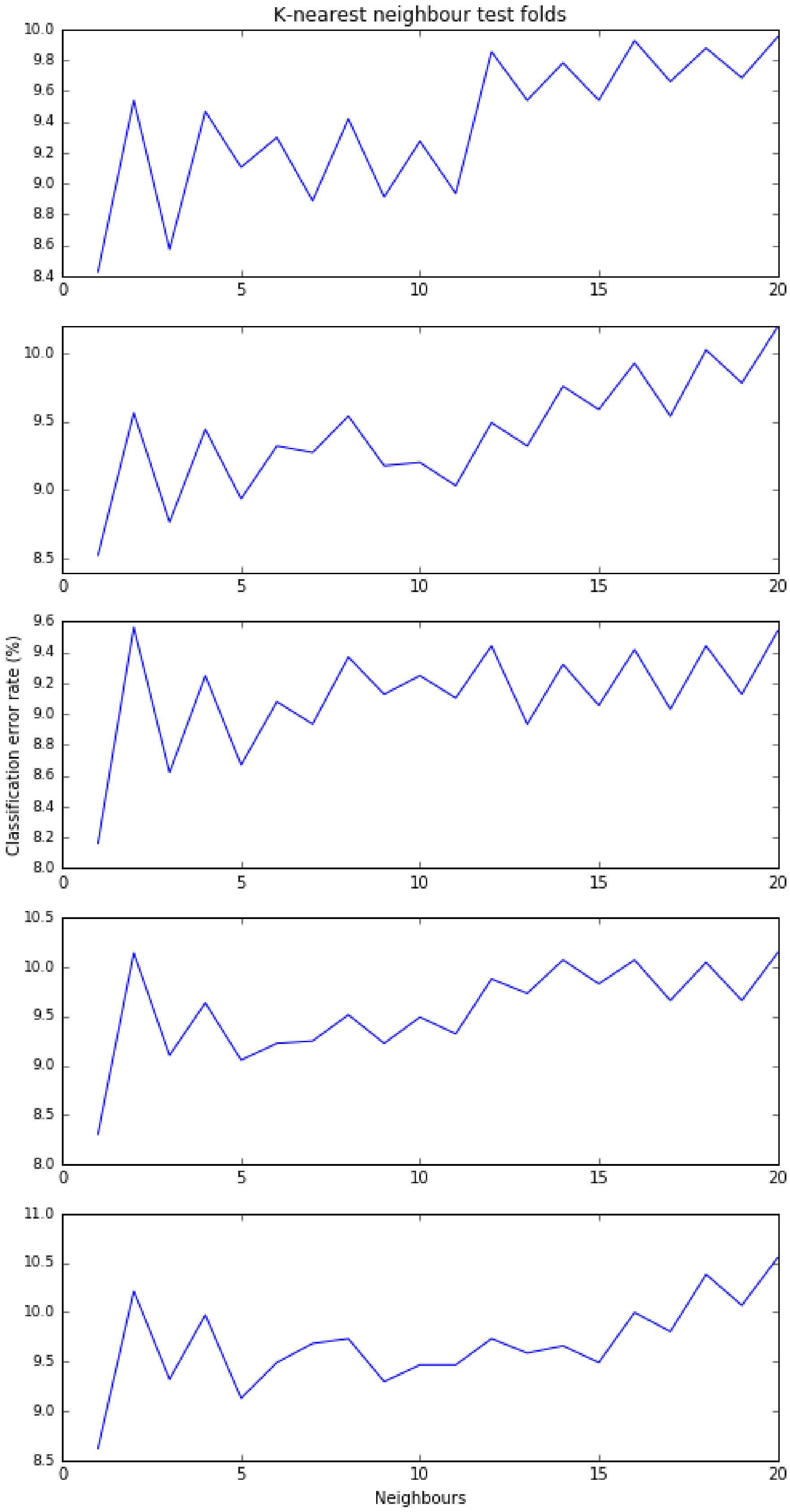


Figure 8: K-Nearest Neighbour error for each fold

4 Conclusion

In the regression part we want to predict the variable word-freq-credit. But error rate is quite high. For the linear regression part, the error rate if above 0.9 even when we model it with feature selection. The modeling result using Artificial Neural Network is better than linear regression, but still has a high mean-square error of 0.9134.

However, when we do the decision problem of checking whether an email is SPAM or not. The prediction result is much better. For both ANN and KNN method we have a test error rate below 10%. And ANN perform better with 6.89% error rate. This shows that predicting spam email with these attributes is somehow effective.