

# Machine Learning Assignment 1

*Agustín Valencia - aguva779*

*11/19/2019*

## Assignment 1. Spam classification with nearest neighbors

**2. Use logistic regression to classify the training and test data by the classification principle  $\hat{Y} = 1$  if  $p(Y = 1|X) > 0.5$ , otherwise  $\hat{Y} = 0$  and report the confusion matrices and the misclassification rates for train and test data. Analyze the obtained results.**

Evaluating the model with training data :

```
## Classification Performance : train set - trigger = 0.5
## TPR = 83.48083 % - TNR = 84.86906 % - FPR = 16.51917 % - FNR = 15.13094 %
## Misclassification Rate = 15.47445 %
```

Now, with unseen data it can be observed that the misclassification rate increased, though numbers still consistent.

```
## Classification Performance : test set - trigger = 0.5
## TPR = 74.92711 % - TNR = 84.2259 % - FPR = 25.07289 % - FNR = 15.7741 %
## Misclassification Rate = 18.10219 %
```

**3. Use logistic regression to classify the test data by the classification principle  $\hat{Y} = 1$  if  $p(Y = 1|X) > 0.8$ , otherwise  $\hat{Y} = 0$**

Setting a higher trigger implies that the classifier will be more selective, then it is expected to decrease the amount of mails being labeled as spam.

The training stats:

```
## Classification Performance : train set - trigger = 0.8
## TPR = 87.10938 % - TNR = 80.61041 % - FPR = 12.89062 % - FNR = 19.38959 %
## Misclassification Rate = 18.17518 %
```

Testing stats:

```
## Classification Performance : test set - trigger = 0.8
## TPR = 76.30522 % - TNR = 79.57181 % - FPR = 23.69478 % - FNR = 20.42819 %
## Misclassification Rate = 21.0219 %
```

Although the misclassification rate has increased, the false positive rate, i.e., the amount of valid email being sent to the spambox, has decreased, which from a user perspective could be more valuable than a higher accuracy on true positives.

**4. Use standard `knn()` with  $K = 30$  from package *knn*, report the misclassification rates for the training and test data and compare the results with step 2.**

```
## Classification Performance : train knn - k = 30
## TPR = 70.42802 % - TNR = 91.00467 % - FPR = 29.57198 % - FNR = 8.995327 %
## Misclassification Rate = 16.71533 %

## Classification Performance : test knn - k = 30
## TPR = 48.97541 % - TNR = 79.59184 % - FPR = 51.02459 % - FNR = 20.40816 %
## Misclassification Rate = 31.31387 %
```

5. Repeat step 4 for  $K=1$  and compare results with step 4. What effects does the decrease of  $K$  lead to and why?

```
## Classification Performance : train knn - k = 1
## TPR = 100 % - TNR = 100 % - FPR = 0 % - FNR = 0 %
## Misclassification Rate = 0 %

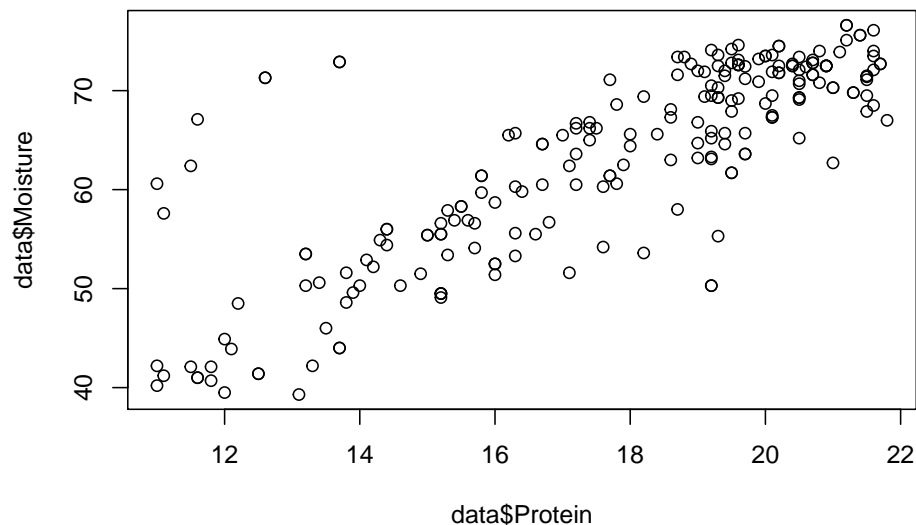
## Classification Performance : test knn - k = 1
## TPR = 43.25323 % - TNR = 77.68396 % - FPR = 56.74677 % - FNR = 22.31604 %
## Misclassification Rate = 35.91241 %
```

If we assign  $k=1$  training misclassification is 0%, this means we are overfitting our model, thus the misclassification for the testing set may be bigger than other scenarios.

### Assignment 3. Feature selection by cross-validation in a linear model.

### Assignment 4. Linear regression and regularization

1. Import data and create a plot of Moisture versus Protein. Do you think these data are described well by a linear model?

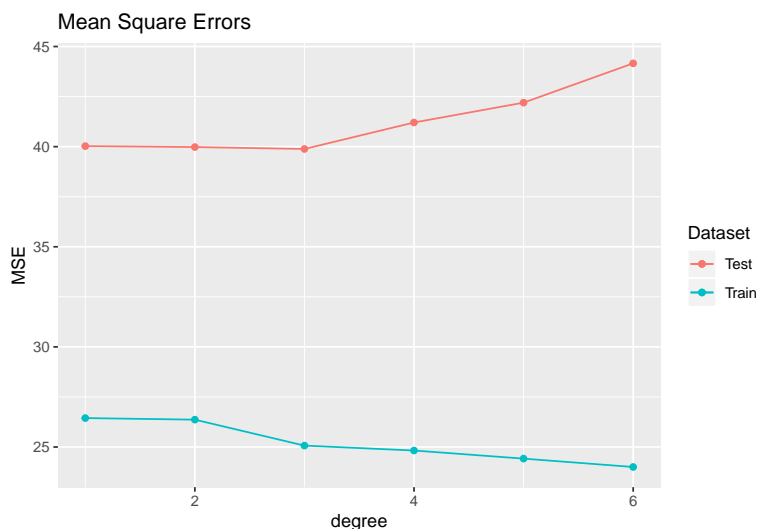


By the plot, although there are some outliers, it seems that the data could be approximated by a linear model.

2. Consider model  $M_i$  in which Moisture is normally distributed and the expected Moisture is polynomial

*PUT DOWN SOME MATH BAAAAAM !!*

3. Divide the data (50/50) and fit models  $M_i, i = 1, \dots, 6$ . For each model, record the training and validation MSE and present a plot showing how training and validation MSE depend on  $i$ . Which model is best according to this plot? How do MSE values change and why? Interpret this picture in bias-variance tradeoff.



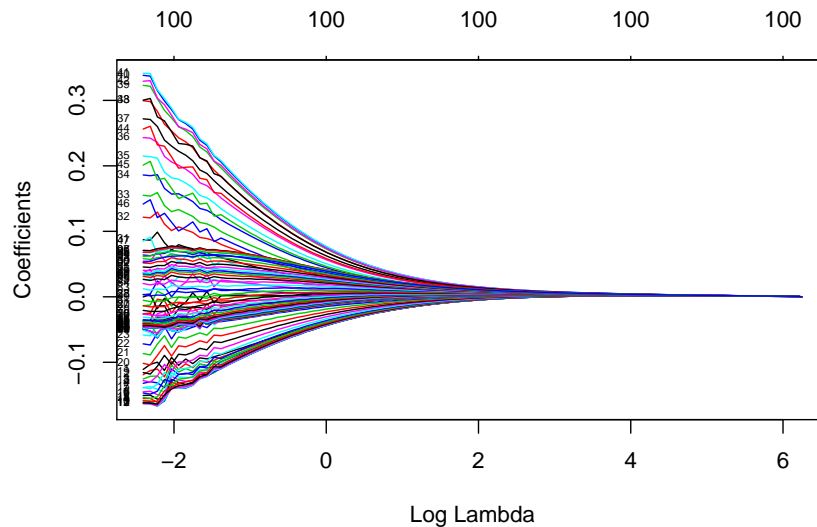
4. Perform variable selection of a linear model in which Fat is response and Channel1-Channel100 are predictors by using stepAIC. Comment on how many variables were selected.

After running stepAIC we get that the amount of selected variables is :

```
## There were selected 64 variables
```

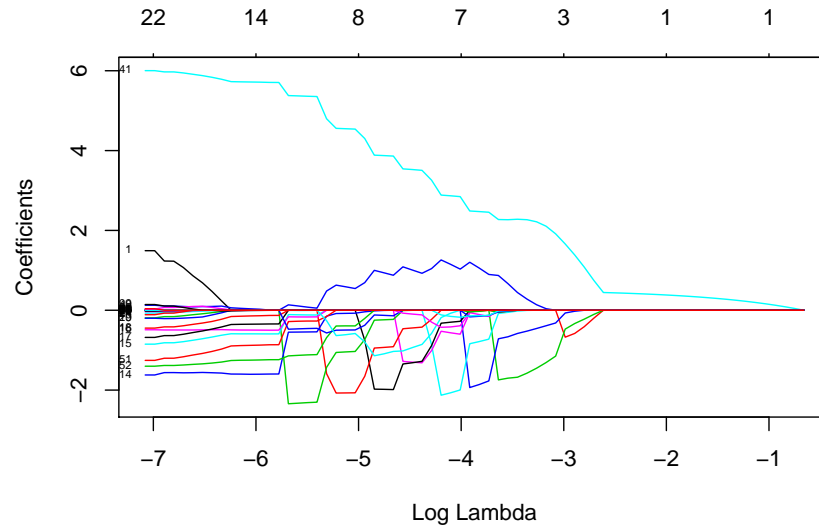
Thus, taking into account that one of them is the intercept, we have 63 selected variables out of 100.

5. Fit a Ridge regression model with the same predictor and response variables. Present a plot showing how model coefficients depend on the log of the penalty factor  $\lambda$  and report how the coefficients change with  $\lambda$



It can be seen that when using Ridge regression among the  $\lambda$  increasing, coefficients converge to zero, though the amount of parameters still 100 since Ridge do not drop them.

6. Repeat step 5 but fit with LASSO instead of the Ridge regression and compare the plots from steps 5 and 6. Conclusions?



LASSO converges to zero much faster than Ridge regression and it also drops variables.

7. Use cross-validation to find optimal LASSO model (make sure that case  $\lambda = 0$  is also considered by the procedure), report the optimal  $\lambda$  and how many variables were chosen by the model and make conclusions. Present also a plot showing the dependence of the CV score and comment how the CV score changes  $\lambda$

8. Compare the results from steps 4 and 7.