group1\_block1\_lab1\_report\_732A99

23 November 2019

# Assignment 1 - Spam classification with nearest neighbors

# 1. setting working directory and importing data

install.packages(“readxl”) library(readxl)

setwd(“/Users/joris/Documents/Joris/Link?ping University/Semester 1/Period 2/732A99 - Machine Learning/Lab1”)

#spam\_data <- read\_excel(“spambase.xlsx”) spam\_data <- read\_xlsx(“spambase.xlsx”)

# dividing the data into training and test sets (50/50)

n=dim(spam\_data)[1] set.seed(12345) id=sample(1:n, floor(n\*0.5)) train=spam\_data[id,] test=spam\_data[-id,]

# Run the model

spam\_logistic <- glm(train$Spam~., family = binomial, data = train) predicted <- predict(spam\_logistic, test, type = “response”)

# 2. Classify on a 0.5 split

predicted\_class <- as.numeric(predicted > 0.5) # confusion matrix table(predicted\_class, test$Spam)

# 3. Classify on a 0.8 split

predicted\_class <- as.numeric(predicted > 0.8) # confusion matrix table(predicted\_class, test$Spam)

# 4. K-nearest neighbor with K = 30

install.packages(“kknn”) library(kknn)

predicted <- kknn(trainfitted.values > 0.5) # confusion matrix table(predicted\_class, test$Spam)

# 5. K-nearest neighbor with K = 1

predicted <- kknn(trainfitted.values > 0.5) # confusion matrix table(predicted\_class, test$Spam)

# ——————————————————————————————

# Assignment 03

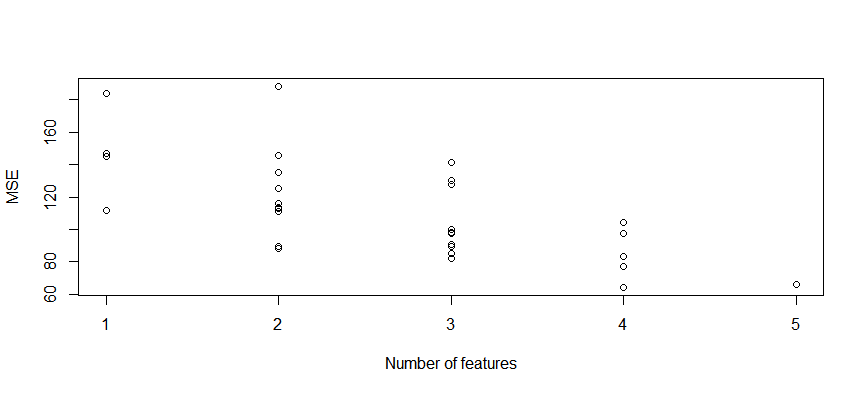
### Test your function on data set swiss available in the standard R repository. Fertility should be Y and all other variables should be X. Nfolds should be X.

**Report the resulting plot and interpret it**

* On the X axis of the plot are shown the number of features used at a time. The Y axis shows the minimal Mean Squared Error (MSE) obtained in the different folds.
* As we increase the number of features used in our model, the MSE reduces.
* We can also see in the 1 variable case, that there should be specifically one feature that reduces drastically the MSE.
* Another interesting characteristic of this plot, is that the min MSE was obtained with 4 features, instead of 5.

**Report the optimal subset of features and comment whether it is reasonable that these specific features have largest impact on the target**

The best result was obtained using features 1 (Agriculture), 3 (Education), 4 (Catholic) and 5 (Infant Mortality). This result, makes sense for variables 1, 3 and 4 as these variables are related with education and lifestyle. But variable 5 does not make sense, because infant mortality should not have any effect on fertility.



## $CV  
## [1] 64.48673  
##   
## $Features  
## [1] 1 0 1 1 1

# ———————————————————————

## Assignment 4. Linear regression and regularization

Plot of Protein vs Moisture:

There is a clear linear relationship between Protein and Moisture and it could be described well by a linear model.

**Consider model in which Moisture is normally distributed, and the expected Moisture is a polynomial function of Protein including the polynomial terms up to power .**  
Probabilistic model for :

$$M\_i \sim N(\sum\_{j=0}^{i} w\_j x^j, \; \sigma^2),\; i \in \{1, 2, 3, 4, 5, 6\} \\$$

We use mean of squared error (MSE) because we want to make our model to be as close as possible to the training data . So the lower the value of MSE, the better the model fits the training data. That way, we can select the best model by looking at the lowest MSE value.  
MSE for :

According to the plot, if we look at MSE values on training set, is the best model. However, if we look at the validation MSE, has the smallest errors. This can be explained by bias-variance trade-off. Models with higher degree of polynomial have lower bias but higher variance, so the model would overfit the training data and resulting in higher prediction errror when presented with validation data. If the models have lower degree of polynomial, it would underfit the data and have higher training MSE. However, due to the lower variance, they generalize better than more complex models hence the lower validation MSE.

**Variable selection of a linear model in which Fat is response and Channel1-Channel100 are predictors by using stepAIC:**

There are 63 variables selected out of 100 variables.

**Ridge regression:**

**LASSO:**

Based on two plots above, we can see that the ridge regression still use all of 63 variables as the value of increased, although the value of the coefficients got smaller but not reaching 0. In the LASSO model, some coefficients reached to 0 even when is still not relatively big. When or , there were 10 variables left that have nonzero coefficient. In conclusion, LASSO models use only several variables as the value of grows while ridge regression models use all variables.

**Cross-validation to find the optimal LASSO model:**

The value of chosen is which is the same as using all 63 variables. In conclusion, the best model for predicting Fat is by using 63 variables that were selected by stepAIC. As seen on the plot, the value of MSE increased as the value of increased too.

# Appendix