# Machine Learning Assignment 1

*Agustín Valencia*

*11/19/2019*

## Assignment 1. Spam classification with nearest neighbors

1. Importing the data:

```
data <- read.xlsx("data/spambase.xlsx")
n = dim(data)[1]
set.seed(12345)
id = sample(1:n, floor(n*0.5))
train = data[id,]
test = data[-id,]
```

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.

```
# util function
get_performance <- function(targets, predictions) {
    t <- table(targets, predictions)
    print("Confusion Matrix")
    print(t)
    tn <- t[1,1]
    tp <- t[2,2]
    fp <- t[1,2]
    fn <- t[2,1]
    total <- dim(test)[1]
    tpr <- tp/total * 100
    tnr <- tn/total * 100
    fpr <- fp/total * 100
    fnr <- fn/total * 100

    cat("Classification performance:\n")
    cat("TPR = ", tpr, "%\n")
    cat("TNR = ", tnr, "%\n")
    cat("FPR = ", fpr, "%\n")
    cat("FNR = ", fnr, "%\n")
}

# fit the model
fit <- glm(Spam ~ . , data = train, family = "binomial")

# performance on training data
pred_train <- predict(fit, newdata = train)
pred_train_at_05 <- as.integer(pred_train > 0.5)
targets <- train$Spam
get_performance(targets, pred_train_at_05)
```

```
## [1] "Confusion Matrix"
##        predictions
## targets   0   1
##       0 875  56
##       1 156 283
## Classification performance:
## TPR =  20.65693 %
## TNR =  63.86861 %
## FPR =  4.087591 %
## FNR =  11.38686 %
```

```
# performance on test data
pred_test <- predict(fit, newdata = test)
pred_test_at_05  <- as.integer(pred_test > 0.5)
targets <- test$Spam
get_performance(targets, pred_test_at_05)
```

```
## [1] "Confusion Matrix"
##        predictions
## targets   0   1
##       0 865  86
##       1 162 257
## Classification performance:
## TPR =  18.75912 %
## TNR =  63.13869 %
## FPR =  6.277372 %
## FNR =  11.82482 %
```