

# Computer Lab X Block X

732A99 Machine Learning

Sofie Jörgensen (sofjo281), Yunan Dong (yundo963) and Shahrukh Iqbal (shaiq681)

08 November 2020

## Statement Of Contribution

We collectively acknowledge that the current lab report is a collaborative work of Sofie Jörgensen(sofjo281), Yunan Dong(yundo963) and Shahrukh Buland Iqbal(shaiq681). In this report Sofie was responsible for Assignment 1, Yunan for Assignment 2 and Shahrukh for Assignment 3.

## Assignment 1. Handwritten digit recognition with K-Nearest Neighbor

In this assignment, we will explore the non-parametric model k-Nearest Neighbor by classifying hand-written digits and evaluate the prediction quality. For this purpose, we will be using the data set **optdigits.csv** where 43 people have contributed to 3822 handwritten digits, where one digit is stored in a matrix  $M$  of size  $8 \times 8$ , where each element in  $M$  is a positive integer in the range 0 to 16. In more detail, there are a total of 65 columns, where each of the 3822 rows represent the elements in  $M$ , distributed on 64 columns. The last column indicates the actual digit, and the elements are integers from 0 to 9.

Comment references.

### 1.

The first step is to load the data set **optdigits.csv** in R and then split the data into 50% training, 25% validation and 25% test set according to the instructions given in lecture 1c, slide 31.

### 2.

Now that the data is divided into training, validation and test set respectively, we are now ready to fit a k-Nearest Neighbor with  $k = 30$ . For this purpose, we are going to build a standard unweighted 30-Nearest Neighbor classifier, by using the **kknn()** function with setting **kernel = "rectangular"**, provided in the **kknn** package. This procedure is done in order to estimate confusion matrices and misclassification errors for the training and the test data. We are especially interested in measuring the prediction quality, both for different digits and overall performance.

## Assignment 2. Title

## Assignment 3. Title

## Appendix

```

# 1.1

# Import data
optdigits <- read.csv(file = "optdigits.csv")
optdigits <- cbind(y=as.factor(optdigits[,65]), optdigits[,1:64])

# Number of rows
n <- dim(optdigits)[1]

set.seed(12345)

# Sample training ids
id_train <- sample(1:n, floor(n*0.5))

# Divide data into training set
train <- optdigits[id_train,]

# Remaining data is validation and test set
temp <- optdigits[-id_train,]
n_temp <- dim(temp)[1]

# Sample validation ids
id_valid <- sample(1:n_temp, floor(n_temp*0.5))

# Divide data into validation set
validation <- temp[id_valid,]

# Divide data into test set
test <- temp[-id_valid,]

# 1.2
library(kknn)

# Fit 30-Nearest Neighbor
model_train <- kknn(formula = formula(train),
                    train = train,
                    test = train,
                    k = 30,
                    kernel = "rectangular")

model_test <- kknn(formula = formula(train),
                  train = train,
                  test = test, k = 30,
                  kernel = "rectangular")

# 1.2 and so on
# Insert code
# 2.1
# Insert code
# 3.1
# Insert code

```