Week 7 - AYUPod - K-Nearest Neighbors

Contents

(Source: seattletimes.com)

KNn for Classification

```
# dummy all categorical variables
# normalize all continuous variables to range 0 and 1
knn_prepared = function(d)
{
  library(tidyverse)
  library(fastDummies)
  d_numeric = d %>% summarise_if(is.numeric, function(x){(x-min(x))/(max(x)-min(x))})

  d_category = d %>% select_if(~!is.numeric(.))
  d_category_dummy = dummy_cols(d_category, remove_first_dummy = TRUE, remove_selected_columns=TRUE)

  return(as_tibble(cbind(d_numeric, d_category_dummy)))

}
```

```
library(tidyverse)
library(caret)
library(class)
```

```
d <- read_csv('german_credit.csv')

d <- rename(d, target=class) # rename the target variable as target
d$target = as.factor(d$target)

df = select(d, -target)

df = knn_prepared(df)
df$target = d$target



library(caret)
set.seed(2020)
splitIndex <- createDataPartition(df$target, p = .70,
                                  list = FALSE)
df_train <- df[ splitIndex,]
df_test <- df[-splitIndex,]

pred = knn(select(df_train, -target), test = select(df_test, -target), cl = df_train$target)
cm <- confusionMatrix(data = pred, reference = df_test$target, positive = "1")
cm$overall[1]
```

```
##  Accuracy
## 0.7133333
```

Question:

- Train a KNN with `k = 3` on the training data to predict the claim cost category (i.e., `claim_cost_category` is your target variable).
- Calculate the accuracy of the decision tree on the test data.

KNN for Regression

```
library(FNN)
d <- read_csv('german_credit.csv')
d <- rename(d, target=credit_amount) # Set credit_amout as the target variable
df = select(d, -target) # select the set of predictors for pre-processing

df = knn_prepared(df)
df$target = d$target



library(caret)
set.seed(2020)
splitIndex <- createDataPartition(df$target, p = .70,
                                  list = FALSE)
df_train <- df[ splitIndex,]
df_test <- df[-splitIndex,]
```

```
pred = knn.reg(train = select(df_train, -target), test = select(df_test, -target), y = df_train$target,
postResample(pred = pred$pred, obs = df_test$target)
```

```
##         RMSE     Rsquared          MAE
## 2412.4396412    0.2091544 1597.6510000
```

Question

- Train a KNN with `k = 5` on the training data to predict the ultimate claim cost (i.e., `UltimateIncurredClaimCost` is your target variable).
- Calculate the RMSE, Rsquared and MAE of the model on the test data.