Week 7 - AYUPod - K-Nearest Neighbors

Table of 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$target, k = 10)  
postResample(pred = pred$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.