Project 1 - k-NN Implementation

X167441 - Ola Tranum Arnegaard

10/03/2019

Objective

-Use k-NN to predict PaymentMethods for a Telecompany's customers given other bill features.

Data

The data consists of the Customer ID + 20 features for each customer. The dataset contains info about a Telecompany's customers. My goal was to predict a categorical variable from qualitative (and one (boolean) categorical) variables, based on a given Knn algorithm.

Analysis

Functions

Because the k-NN algorithm use the distance between the data points as scores, it is important that all distances are normalized. Normalization:

```r
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
```

Code

Read data:

```r
data <- read.csv("/Users/ola/Desktop/Fall2019/UC-Berkeley/Practical ML (R)/TelecompanyChurn/churn.csv",
```

Choose the desired columns from the dataset. Initially, I used the *MonthlyCharges* and *TotalCharges*; however, later, I added more variables. If desired, more features may also be added.

```r
data_chosen_col <- data.frame(data[c(18,3,6)], data[19:20])
```

Remove N/A cells as well as normalizing:

```r
data_not_na <- na.omit(data_chosen_col)
data_normalized <- as.data.frame(lapply(data_not_na[2:5], normalize))
```

Separate into train and test sets ($_{75\%}$/25%):

```r
data_train <- data_normalized[1:5500,] #around 75%
data_test <- data_normalized[5501:7032,]
```

Separate the labels (the predictive unknown)

```r
data_train_labels <- data_not_na[1:5500, 1] #returns a vector: data_train_labels[1:10]
data_test_labels <- data_not_na[5501:7032, 1]
```

Run k-NN algorithm with k=75 (~sqrt(lenght(train_data))):

```r
library(class)
prediction <- knn(train = data_train, test = data_test, cl = data_train_labels, k = 75)
```

Create crosstable summary:

```r
library(gmodels)
CrossTable(x = data_test_labels, y = prediction, prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1532 
## 
## 
##                           | prediction 
##          data_test_labels | Bank transfer (automatic) |   Credit card (automatic) |          Electro
## --------------------------|---------------------------|---------------------------|-----------------
## Bank transfer (automatic) |                        54 |                        85 | 
##                           |                     0.161 |                     0.254 | 
##                           |                     0.295 |                     0.370 | 
##                           |                     0.035 |                     0.055 | 
## --------------------------|---------------------------|---------------------------|-----------------
##   Credit card (automatic) |                        67 |                        66 | 
##                           |                     0.208 |                     0.205 | 
##                           |                     0.366 |                     0.287 | 
##                           |                     0.044 |                     0.043 | 
## --------------------------|---------------------------|---------------------------|-----------------
##          Electronic check |                        42 |                        48 | 
##                           |                     0.081 |                     0.093 | 
##                           |                     0.230 |                     0.209 | 
##                           |                     0.027 |                     0.031 | 
## --------------------------|---------------------------|---------------------------|-----------------
##              Mailed check |                        20 |                        31 | 
##                           |                     0.056 |                     0.087 | 
##                          |                     0.109 |                     0.135 |
##                          |                     0.013 |                     0.020 |
## -------------------------|---------------------------|---------------------------|-----------------
##             Column Total |                       183 |                       230 |
##                          |                     0.119 |                     0.150 |
## -------------------------|---------------------------|---------------------------|-----------------
##
##
```

Result

As seen in the crosstable, the predictions does not ork, i.e. by running through all the training data for each of our test data, the distances to the (max_of)K nearest neighbours on our chosen set of features, does not predict the unknown variable,*PaymentMethod*, very well. One example is the predicted *BankTransfer (automatic)* category, where we correctly predicted 55 of a total of 180, but were supposed to predict a total of 335. Because this is a Knn algorithm, it is important to have features that group very well with the category we want to predict (high covariance/low normalized distance). Therefore, one possibility would be to add and remove different features from the full dataset to the ones we use in this script, and then look for improvements.