Knn - Ola Tranum Arnegaard

Knn

-Predict PaymentMethod for induviduals by Telecompany 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.

Functions

Normalization:

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

Code

Read data:

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

Choose the columns from the dataset that I wish to use. Initially, I only used the *MonthlyCharges* and *TotalCharges*, but the predictions were not good. Therefore, I added more variables. If desired, it is possible to choose additional features by adding more columns in the line below.

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

Remove N/A cells as well as normalizing:

```
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 ($_{75\%}/25\%$):

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

Separate put the labels (the predictive unknown)

```
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 Knn algorithm with k=75 (~sqrt(lenght(train_data))):

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

See crosstable summary:

```
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 |
##                            |                     0.297 |                     0.368 |
##                            |                     0.035 |                     0.055 |
## -------------------------|---------------------------|---------------------------|-----------------
##   Credit card (automatic) |                        65 |                        67 |
##                            |                     0.202 |                     0.208 |
##                            |                     0.357 |                     0.290 |
##                            |                     0.042 |                     0.044 |
## -------------------------|---------------------------|---------------------------|-----------------
##          Electronic check |                        44 |                        48 |
##                            |                     0.085 |                     0.093 |
##                            |                     0.242 |                     0.208 |
##                            |                     0.029 |                     0.031 |
## -------------------------|---------------------------|---------------------------|-----------------
##              Mailed check |                        19 |                        31 |
##                            |                     0.053 |                     0.087 |
##                            |                     0.104 |                     0.134 |
##                            |                     0.012 |                     0.020 |
## -------------------------|---------------------------|---------------------------|-----------------
##              Column Total |                       182 |                       231 |
##                            |                     0.119 |                     0.151 |
## -------------------------|---------------------------|---------------------------|-----------------
##
##
```

As seen in the table, the prediction is not very good, 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.