Assignment 1

Submitted by Maryam Heidari

ALY 6020- Predictive Analytics- CRN: 20770

Northeastern University

Date of Submission: 24 February 2020

Introduction:

In this assignment, we are going to get familiar with KNN method which is one of the simplest methods of classification.

Part A:

In the part one, I am going to use the example in the chapter 3 of Machine Learning with R (Lantz) and follow the five steps. In this example I used the Wisconsin Breast Cancer Diagnostic dataset from the UCI.

The result of each patient is in the M column and you can see that in the below table:

```
  B   M
357 211
```

To avoid any misunderstanding, I renamed the B and M with "Benign", "Malignant". Now, the summary table of this column is:

```
   Benign Malignant
     62.9      37.1
```

Since I have different parameters with different scales, I need to normalize the data set.

```r
# normalize these features
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
wbcd_n <- as.data.frame(lapply(wbcd[3:32], normalize))
```

After normalizing the data set, I divided it to two data sets which are train and test, so I can use the train one to make a KNN model and use the test one to evaluate the accuracy of the model. And also, I need to store these class labels in factor vectors, to generate the KNN model.

```
wbcd_train <- wbcd_n[1:469, ]
wbcd_test <- wbcd_n[470:568, ]
wbcd_train_labels <- wbcd[1:469, 2]
wbcd_test_labels <- wbcd[470:568, 2]
```

Now, I can make a model. As my train data includes 469 instances, I used k = 21, an odd number roughly equal to the square root of 469. With a two-category outcome, using an odd number eliminates the chance of ending with a tie vote.

```
install.packages("class")
library(class)

wbcd_test_pred <- knn(train = wbcd_train, test = wbcd_test,
                      cl = wbcd_train_labels, k = 21)
```

Then I used the CrossTable() function to evaluate how good is my model.

```
library(gmodels)
CrossTable(x = wbcd_test_labels, y = wbcd_test_pred, prop.chisq=FALSE)
```

And the result is:

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  99


                 | wbcd_test_pred
wbcd_test_labels |    Benign | Malignant | Row Total |
-----------------|-----------|-----------|-----------|
          Benign |        76 |         0 |        76 |
                 |     1.000 |     0.000 |     0.768 |
                 |     0.974 |     0.000 |           |
                 |     0.768 |     0.000 |           |
-----------------|-----------|-----------|-----------|
       Malignant |         2 |        21 |        23 |
                 |     0.087 |     0.913 |     0.232 |
                 |     0.026 |     1.000 |           |
                 |     0.020 |     0.212 |           |
-----------------|-----------|-----------|-----------|
    Column Total |        78 |        21 |        99 |
                 |     0.788 |     0.212 |           |
-----------------|-----------|-----------|-----------|
```

As you can see, the percentages of true negative results is 76%, the percentages of true positive results is 23%, the percentages of the false positive is 21%, and the percentages of false negative results is 2%.

The desire is to make false negative close to zero, so to improve the model, I standardize the data set, and repeat the steps and the result is:

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  99


                 | wbcd_test_pred
wbcd_test_labels |    Benign | Malignant | Row Total |
-----------------|-----------|-----------|-----------|
         Benign |        76 |         0 |        76 |
                |     1.000 |     0.000 |     0.768 |
                |     0.974 |     0.000 |           |
                |     0.768 |     0.000 |           |
-----------------|-----------|-----------|-----------|
       Malignant |         2 |        21 |        23 |
                |     0.087 |     0.913 |     0.232 |
                |     0.026 |     1.000 |           |
                |     0.020 |     0.212 |           |
-----------------|-----------|-----------|-----------|
    Column Total |        78 |        21 |        99 |
                |     0.788 |     0.212 |           |
-----------------|-----------|-----------|-----------|
```

The result is as same as the last table, and unfortunately it does not improve.

Part B:

In this part, we need to repeat all of the steps for new data set. I find a data set about the heart disease.

The result is in the target column, the table of this column is:

```
  0   1
138 165
```

I renamed them by "no heart disease" and "heart disease" to avoid misunderstanding, and the table of parentages is:

```
no heart disease    heart disease
            45.5             54.5
```

I normalized, divided the data set to the train and test and follow the steps.

As my train data includes 303 instances, I used k = 17, an odd number roughly equal to the square root of 303. With a two-category outcome, using an odd number eliminates the chance of ending with a tie vote.

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  61


                | dt_test_pred
  dt_test_labels | no heart disease |    heart disease |        Row Total |
-----------------|------------------|------------------|------------------|
no heart disease |               23 |                8 |               31 |
                 |            0.742 |            0.258 |            0.508 |
                 |            0.793 |            0.250 |                  |
                 |            0.377 |            0.131 |                  |
-----------------|------------------|------------------|------------------|
   heart disease |                6 |               24 |               30 |
                 |            0.200 |            0.800 |            0.492 |
                 |            0.207 |            0.750 |                  |
                 |            0.098 |            0.393 |                  |
-----------------|------------------|------------------|------------------|
    Column Total |               29 |               32 |               61 |
                 |            0.475 |            0.525 |                  |
-----------------|------------------|------------------|------------------|
```

As you can see, the percentages of true negative results is 23%, the percentages of true positive results is 8%, the percentages of the false positive is 24%, and the percentages of false negative results is 6%.

The desire is to make false negative close to zero, so to improve the model, I standardize the data set, and repeat the steps and the result is:

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  61


                | dt_test_pred
  dt_test_labels | no heart disease |    heart disease |        Row Total |
-----------------|------------------|------------------|------------------|
no heart disease |               19 |                6 |               25 |
                 |            0.760 |            0.240 |            0.410 |
                 |            0.864 |            0.154 |                  |
                 |            0.311 |            0.098 |                  |
-----------------|------------------|------------------|------------------|
   heart disease |                3 |               33 |               36 |
                 |            0.083 |            0.917 |            0.590 |
                 |            0.136 |            0.846 |                  |
                 |            0.049 |            0.541 |                  |
-----------------|------------------|------------------|------------------|
    Column Total |               22 |               39 |               61 |
                 |            0.361 |            0.639 |                  |
-----------------|------------------|------------------|------------------|
```

After standardizing, the percentage of false negative changed from 6% to 3% which is so much better.

Then I changed the K from 17 to 21, And the result is:

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  61


                | dt_test_pred
  dt_test_labels | no heart disease |    heart disease |        Row Total |
-----------------|------------------|------------------|------------------|
no heart disease |               19 |                6 |               25 |
                 |            0.760 |            0.240 |            0.410 |
                 |            0.905 |            0.150 |                  |
                 |            0.311 |            0.098 |                  |
-----------------|------------------|------------------|------------------|
   heart disease |                2 |               34 |               36 |
                 |            0.056 |            0.944 |            0.590 |
                 |            0.095 |            0.850 |                  |
                 |            0.033 |            0.557 |                  |
-----------------|------------------|------------------|------------------|
    Column Total |               21 |               40 |               61 |
                 |            0.344 |            0.656 |                  |
-----------------|------------------|------------------|------------------|
```

the percentage of false negative changed from 3% to 2% which is so much better. But after that even by increasing the K, I do not see any improvement.

Conclusion:

I learned about the KNN method, how it , how to choose your K number, and how I can improve it.

References:

Ch. 3 of Machine Learning with R (Lantz), in pp. 75-87

http://archive.ics.uci.edu/ml

Kaggle.com