Regression_VasanthakumarKalaikkovan

Vasanthakumar Kalaikkovan

29/05/2021

Regression Algorithm

Regression algorithms are used to predict numeric quantity while classification algorithms predict categorical outcomes. A spam filter is an example use case for a classification algorithm. The input dataset is emails labeled as either spam (i.e. junk emails) or ham (i.e. good emails). The classification algorithm uses features extracted from the emails to learn which emails fall into which category.In this problem, you will use the nearest neighbors algorithm to fit a model on two simplified datasets. The first dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables (You worked with this dataset last week!). The second dataset (found in trinary-classifier-data.csv) is similar to the first dataset except that the label variable can be 0, 1, or 2.Note that in real-world datasets, your labels are usually not numbers, but text-based descriptions of the categories (e.g. spam or ham). In practice, you will encode categorical variables into numeric values.

```
library(class)
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 4.0.5
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

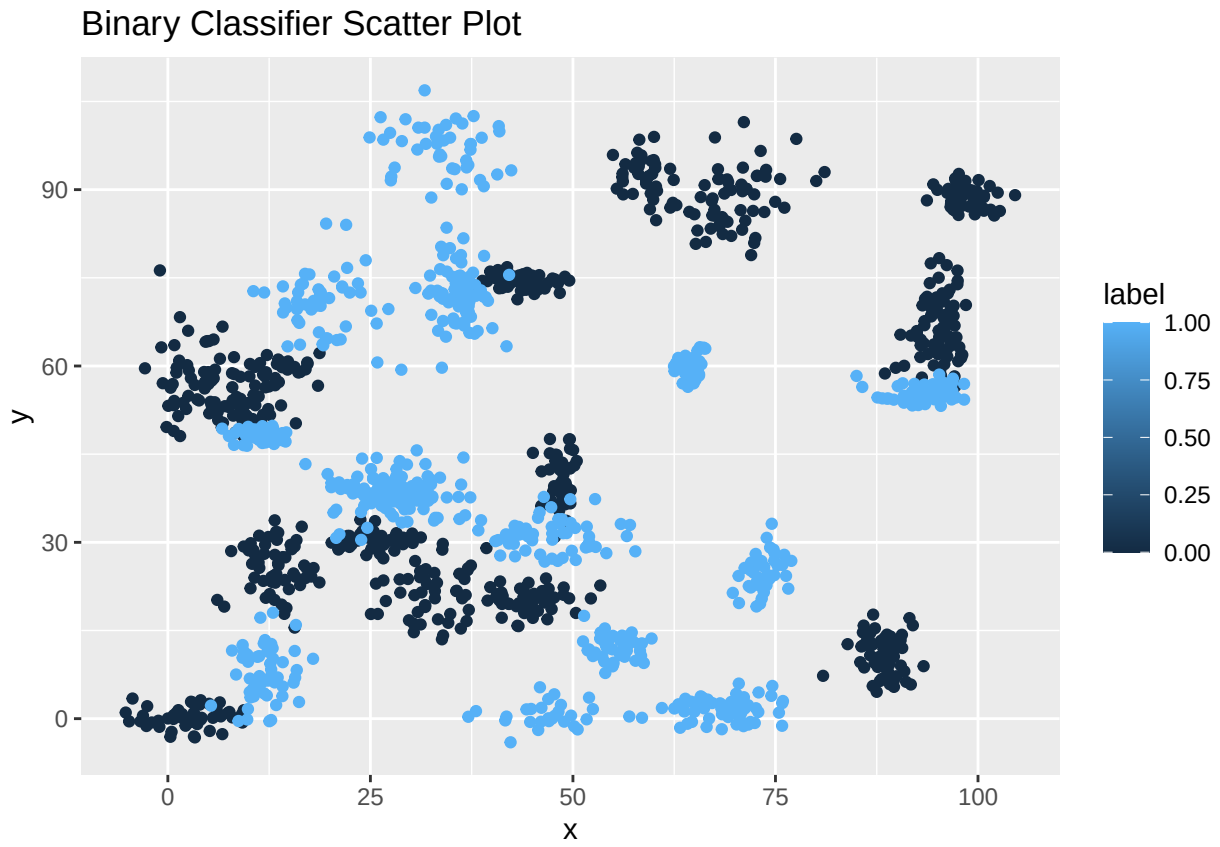
```
getwd()
```

```
## [1] "E:/Repos/StatisticsR/DSC520-Statistics/week11"
```

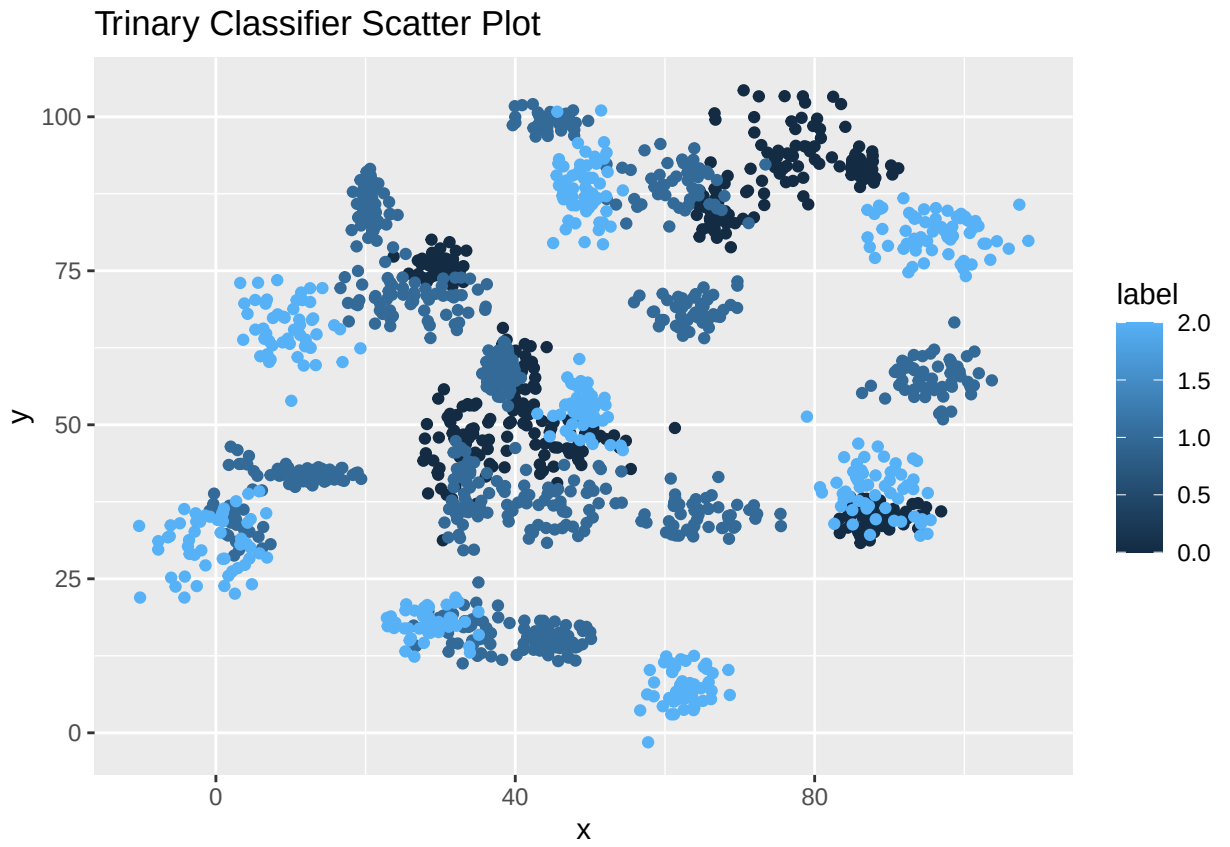
```
setwd("E://Repos/StatisticsR/DSC520-Statistics/week11")

binary_df <- read.csv("binary-classifier-data.csv")
trinary_df <- read.csv("trinary-classifier-data.csv")
```

Plot the data from each dataset using a scatter plot.

```
ggplot(binary_df,aes(x=x,y=y,color=label))+geom_point()+labs(title="Binary Classifier Scatter Plot")
```




```
ggplot(trinary_df,aes(x=x,y=y,color=label))+geom_point()+labs(title="Trinary Classifier Scatter Plot")
```



The k nearest neighbors algorithm categorizes an input value by looking at the labels for the k nearest points and assigning a category based on the most common label. In this problem, you will determine which points are nearest by calculating the Euclidean distance between two points. As a refresher, the Euclidean distance between two points: i.Fitting a model is when you use the input data to create a predictive model. There are various metrics you can use to determine how well your model fits the data. For this problem, you will focus on a single metric, accuracy. Accuracy is simply the percentage of how often the model predicts the correct result. If the model always predicts the correct result, it is 100% accurate. If the model always predicts the incorrect result, it is 0% accurate. ii.Fit a k nearest neighbors' model for each dataset for k=3, k=5, k=10, k=15, k=20, and k=25. Compute the accuracy of the resulting models for each value of k. Plot the results in a graph where the x-axis is the different values of k and the y-axis is the accuracy of the model.

```
k <- c(3,5,10,15,20,25)
binary_accuracy <- NULL

for(i in 1:6)
{
  cat("k=NN Binary Classisfier:",k[i])
  binary_knn<-knn(train=binary_df,test=binary_df,cl=as.factor(binary_df$label),k=k[i])
```

```
  binary_table <- CrossTable(x=binary_df$label,y=binary_knn,prop.chisq = FALSE)
  binary_accuracy[i] <-binary_table$prop.tbl[1,1]+binary_table$prop.tbl[2,2]
}
```

```
## k=NN Binary Classisfier: 3
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       757 |        10 |       767 |
##                 |     0.987 |     0.013 |     0.512 |
##                 |     0.986 |     0.014 |           |
##                 |     0.505 |     0.007 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        11 |       720 |       731 |
##                 |     0.015 |     0.985 |     0.488 |
##                 |     0.014 |     0.986 |           |
##                 |     0.007 |     0.481 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       768 |       730 |      1498 |
##                 |     0.513 |     0.487 |           |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 5
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
```

```
##               0 |       756 |        11 |       767 |
##                 |     0.986 |     0.014 |     0.512 |
##                 |     0.981 |     0.015 |           |
##                 |     0.505 |     0.007 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        15 |       716 |       731 |
##                 |     0.021 |     0.979 |     0.488 |
##                 |     0.019 |     0.985 |           |
##                 |     0.010 |     0.478 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       771 |       727 |      1498 |
##                 |     0.515 |     0.485 |           |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 10
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       752 |        15 |       767 |
##                 |     0.980 |     0.020 |     0.512 |
##                 |     0.979 |     0.021 |           |
##                 |     0.502 |     0.010 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        16 |       715 |       731 |
##                 |     0.022 |     0.978 |     0.488 |
##                 |     0.021 |     0.979 |           |
##                 |     0.011 |     0.477 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       768 |       730 |      1498 |
##                 |     0.513 |     0.487 |           |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 15
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
```

```
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       752 |        15 |       767 |
##                 |     0.980 |     0.020 |     0.512 |
##                 |     0.977 |     0.021 |           |
##                 |     0.502 |     0.010 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        18 |       713 |       731 |
##                 |     0.025 |     0.975 |     0.488 |
##                 |     0.023 |     0.979 |           |
##                 |     0.012 |     0.476 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       770 |       728 |      1498 |
##                 |     0.514 |     0.486 |           |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 20
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       751 |        16 |       767 |
##                 |     0.979 |     0.021 |     0.512 |
##                 |     0.975 |     0.022 |           |
##                 |     0.501 |     0.011 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        19 |       712 |       731 |
##                 |     0.026 |     0.974 |     0.488 |
##                 |     0.025 |     0.978 |           |
##                 |     0.013 |     0.475 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       770 |       728 |      1498 |
##                 |     0.514 |     0.486 |           |
```

```
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 25
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       749 |        18 |       767 |
##                 |     0.977 |     0.023 |     0.512 |
##                 |     0.979 |     0.025 |           |
##                 |     0.500 |     0.012 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        16 |       715 |       731 |
##                 |     0.022 |     0.978 |     0.488 |
##                 |     0.021 |     0.975 |           |
##                 |     0.011 |     0.477 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       765 |       733 |      1498 |
##                 |     0.511 |     0.489 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```
trinary_accuracy <- NULL

for(i in 1:6)
{
  cat("k=NN Trinary Classisfier:",k[i])
  trinary_knn<-knn(train=trinary_df,test=trinary_df,cl=as.factor(trinary_df$label),k=k[i])
  trinary_table <- CrossTable(x=trinary_df$label,y=trinary_knn,prop.chisq = FALSE)
  trinary_accuracy[i] <-trinary_table$prop.tbl[1,1]+trinary_table$prop.tbl[2,2]
}
```

```
## k=NN Trinary Classisfier: 3
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
```

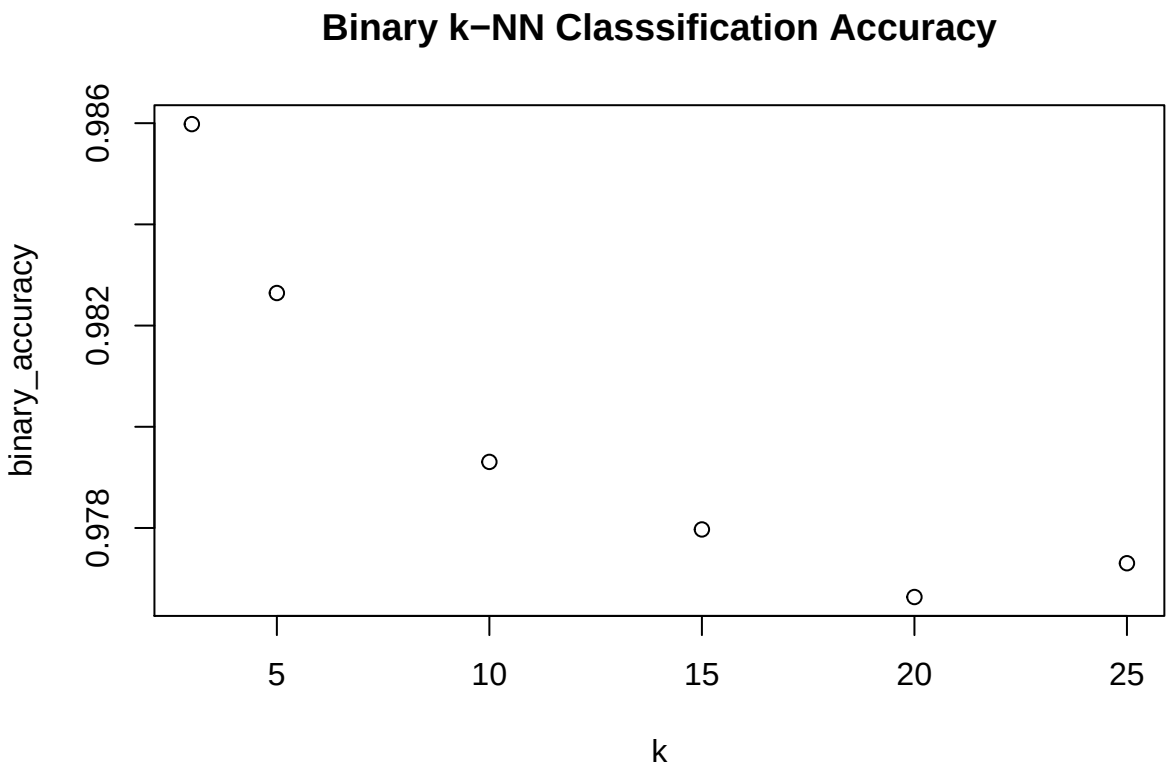
```
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                   | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       386 |         8 |         0 |       394 |
##                  |     0.980 |     0.020 |     0.000 |     0.251 |
##                  |     0.965 |     0.011 |     0.000 |           |
##                  |     0.246 |     0.005 |     0.000 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |         9 |       707 |         6 |       722 |
##                  |     0.012 |     0.979 |     0.008 |     0.460 |
##                  |     0.022 |     0.975 |     0.014 |           |
##                  |     0.006 |     0.451 |     0.004 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |         5 |        10 |       437 |       452 |
##                  |     0.011 |     0.022 |     0.967 |     0.288 |
##                  |     0.012 |     0.014 |     0.986 |           |
##                  |     0.003 |     0.006 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       400 |       725 |       443 |      1568 |
##                  |     0.255 |     0.462 |     0.283 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classifier: 5
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                   | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       376 |        18 |         0 |       394 |
##                  |     0.954 |     0.046 |     0.000 |     0.251 |
##                  |     0.952 |     0.025 |     0.000 |           |
##                  |     0.240 |     0.011 |     0.000 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        13 |       695 |        14 |       722 |
##                  |     0.018 |     0.963 |     0.019 |     0.460 |
##                  |     0.033 |     0.957 |     0.031 |           |
```

```
##                  |     0.008 |     0.443 |     0.009 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |         6 |        13 |       433 |       452 |
##                  |     0.013 |     0.029 |     0.958 |     0.288 |
##                  |     0.015 |     0.018 |     0.969 |           |
##                  |     0.004 |     0.008 |     0.276 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       395 |       726 |       447 |      1568 |
##                  |     0.252 |     0.463 |     0.285 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classisfier: 10
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                  | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       366 |        25 |         3 |       394 |
##                  |     0.929 |     0.063 |     0.008 |     0.251 |
##                  |     0.920 |     0.034 |     0.007 |           |
##                  |     0.233 |     0.016 |     0.002 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        18 |       684 |        20 |       722 |
##                  |     0.025 |     0.947 |     0.028 |     0.460 |
##                  |     0.045 |     0.934 |     0.046 |           |
##                  |     0.011 |     0.436 |     0.013 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        14 |        23 |       415 |       452 |
##                  |     0.031 |     0.051 |     0.918 |     0.288 |
##                  |     0.035 |     0.031 |     0.947 |           |
##                  |     0.009 |     0.015 |     0.265 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       398 |       732 |       438 |      1568 |
##                  |     0.254 |     0.467 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classisfier: 15
##
##    Cell Contents
## |-------------------------|
## |                       N |
```

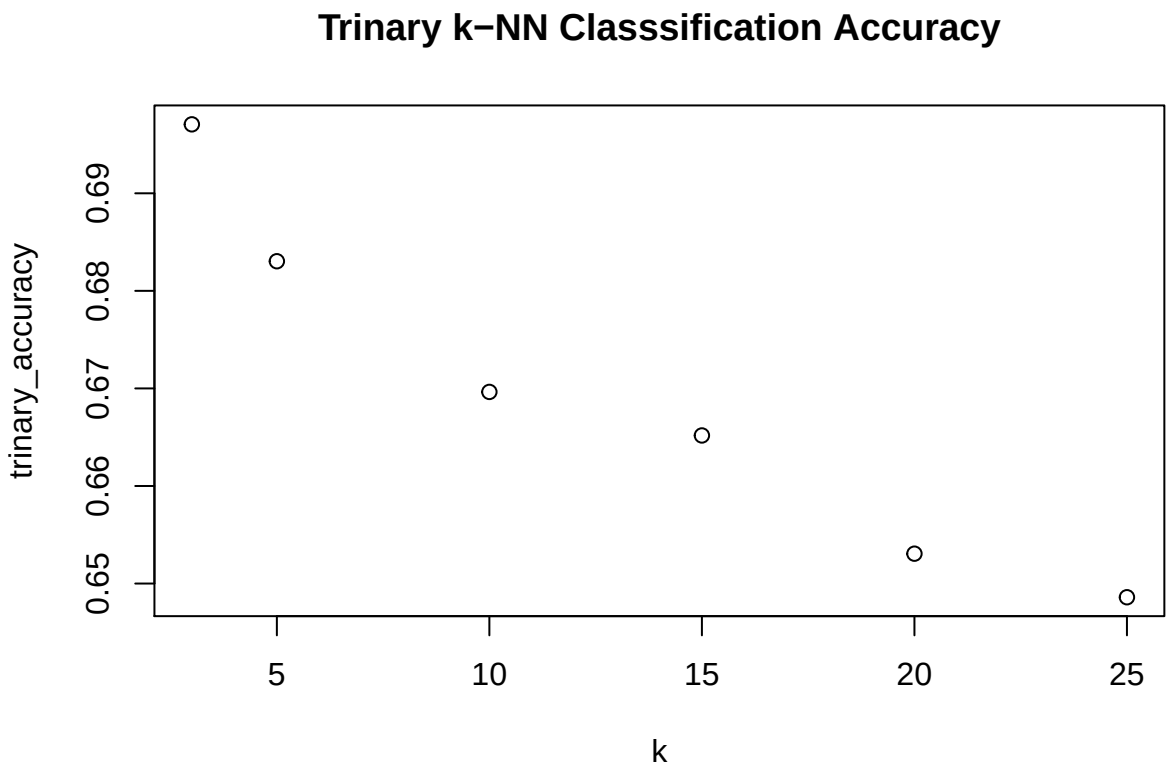
```
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                  | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       364 |        26 |         4 |       394 |
##                  |     0.924 |     0.066 |     0.010 |     0.251 |
##                  |     0.899 |     0.036 |     0.009 |           |
##                  |     0.232 |     0.017 |     0.003 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        21 |       679 |        22 |       722 |
##                  |     0.029 |     0.940 |     0.030 |     0.460 |
##                  |     0.052 |     0.937 |     0.050 |           |
##                  |     0.013 |     0.433 |     0.014 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        20 |        20 |       412 |       452 |
##                  |     0.044 |     0.044 |     0.912 |     0.288 |
##                  |     0.049 |     0.028 |     0.941 |           |
##                  |     0.013 |     0.013 |     0.263 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       405 |       725 |       438 |      1568 |
##                  |     0.258 |     0.462 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classisfier: 20
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                  | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       350 |        39 |         5 |       394 |
##                  |     0.888 |     0.099 |     0.013 |     0.251 |
##                  |     0.879 |     0.053 |     0.011 |           |
##                  |     0.223 |     0.025 |     0.003 |           |
## -----------------|-----------|-----------|-----------|-----------|
```

```
##                1 |        24 |       674 |        24 |       722 |
##                  |     0.033 |     0.934 |     0.033 |     0.460 |
##                  |     0.060 |     0.918 |     0.055 |           |
##                  |     0.015 |     0.430 |     0.015 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        24 |        21 |       407 |       452 |
##                  |     0.053 |     0.046 |     0.900 |     0.288 |
##                  |     0.060 |     0.029 |     0.933 |           |
##                  |     0.015 |     0.013 |     0.260 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       398 |       734 |       436 |      1568 |
##                  |     0.254 |     0.468 |     0.278 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classisfier: 25
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                  | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       343 |        43 |         8 |       394 |
##                  |     0.871 |     0.109 |     0.020 |     0.251 |
##                  |     0.873 |     0.058 |     0.018 |           |
##                  |     0.219 |     0.027 |     0.005 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        24 |       674 |        24 |       722 |
##                  |     0.033 |     0.934 |     0.033 |     0.460 |
##                  |     0.061 |     0.913 |     0.055 |           |
##                  |     0.015 |     0.430 |     0.015 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        26 |        21 |       405 |       452 |
##                  |     0.058 |     0.046 |     0.896 |     0.288 |
##                  |     0.066 |     0.028 |     0.927 |           |
##                  |     0.017 |     0.013 |     0.258 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       393 |       738 |       437 |      1568 |
##                  |     0.251 |     0.471 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
```

```
plot(k,binary_accuracy,main="Binary k-NN Classsification Accuracy")
```




```
plot(k,trinary_accuracy,main="Trinary k-NN Classsification Accuracy")
```


Trinary k−NN Classsification Accuracy
trinary_accuracy
0.65
0.66
0.67
0.68
0.69
5
10
15
20
25
k