KNN on Oncogenic Cells

Adwesh Behera - 20025006

11 September 2018

Description

This is solution to “R - Programming: K-Nearest Neighbour Assignment” on Topcoder.

Problem Statement

Routine breast cancer screening allows the disease to be diagnosed and treated prior to it causing noticeable symptoms. The process of early detection involves examining the breast tissue for abnormal lumps or masses. If a lump is found, a fine-needle aspiration biopsy is performed, which uses a hollow needle to extract a small sample of cells from the mass. A clinician then examines the cells under a microscope to determine whether the mass is likely to be malignant or benign.

If machine learning could automate the identification of cancerous cells, it would provide considerable benefit to the health system. Automated processes are likely to improve the efficiency of the detection process, allowing physicians to spend less time diagnosing and more time treating the disease. An automated screening system might also provide greater detection accuracy by removing the inherently subjective human component from the process.

Apply the k-NN algorithm to perform diagnosis Benign or Malignant

The Dataset

The dataset for the above problem was imported from https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/wisc_bc_data.csv

s

```r
data = read.csv("wisc_bc_data.csv")
str(data)
```

```
## 'data.frame':    569 obs. of  32 variables:
##  $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 8449
##  $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

```r
data = data[-1]    #Removal of patient indices
```

Including Libraries

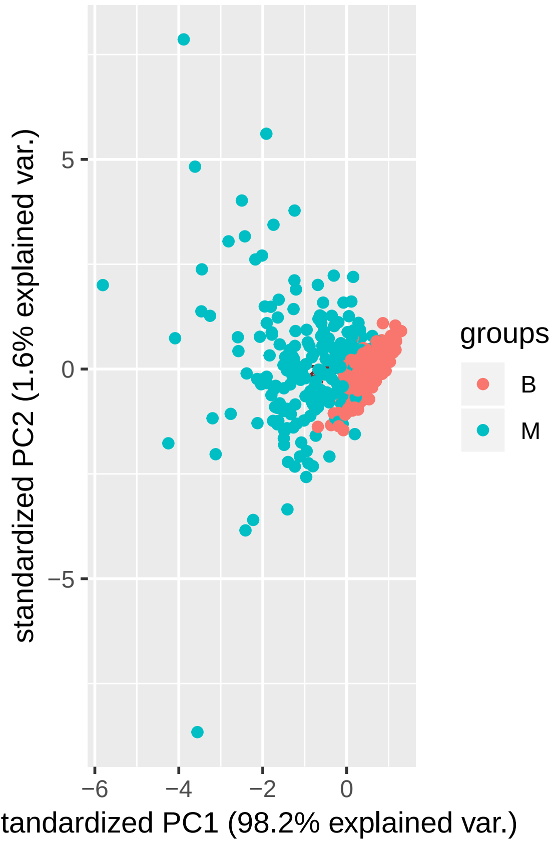
The following libraries were needed to be installed and imported for the assignment.

```r
library(pca3d)
library(class)
library(gmodels)
library(ggbiplot)
```

```
## Loading required package: ggplot2

## Loading required package: plyr

## Loading required package: scales

## Loading required package: grid
```

Principal Component Analysis

```r
#pca3d(princomp(data[,c(2:31)]), group = data$diagnosis)
ggbiplot(princomp(data[,c(2:31)]), groups = data$diagnosis, varname.size = 0)
```



The graph above shows the existance of spacial separation in the euclidian space between the two tumour catrgories.

Pre-Processing : Normalization

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

data_n = as.data.frame(lapply(data[2:31], normalize))
summary(data_n)
```

```
##   radius_mean      texture_mean    perimeter_mean     area_mean     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.2233   1st Qu.:0.2185   1st Qu.:0.2168   1st Qu.:0.1174  
##  Median :0.3024   Median :0.3088   Median :0.2933   Median :0.1729  
##  Mean   :0.3382   Mean   :0.3240   Mean   :0.3329   Mean   :0.2169  
##  3rd Qu.:0.4164   3rd Qu.:0.4089   3rd Qu.:0.4168   3rd Qu.:0.2711  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  smoothness_mean  compactness_mean concavity_mean    concave.points_mean
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000     
##  1st Qu.:0.3046   1st Qu.:0.1397   1st Qu.:0.06926   1st Qu.:0.1009     
##  Median :0.3904   Median :0.2247   Median :0.14419   Median :0.1665     
##  Mean   :0.3948   Mean   :0.2606   Mean   :0.20806   Mean   :0.2431     
##  3rd Qu.:0.4755   3rd Qu.:0.3405   3rd Qu.:0.30623   3rd Qu.:0.3678     
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000     
##  symmetry_mean    fractal_dimension_mean   radius_se      
##  Min.   :0.0000   Min.   :0.0000         Min.   :0.00000  
##  1st Qu.:0.2823   1st Qu.:0.1630         1st Qu.:0.04378  
##  Median :0.3697   Median :0.2439         Median :0.07702  
##  Mean   :0.3796   Mean   :0.2704         Mean   :0.10635  
##  3rd Qu.:0.4530   3rd Qu.:0.3404         3rd Qu.:0.13304  
##  Max.   :1.0000   Max.   :1.0000         Max.   :1.00000  
##    texture_se      perimeter_se        area_se        smoothness_se   
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.1047   1st Qu.:0.04000   1st Qu.:0.02064   1st Qu.:0.1175  
##  Median :0.1653   Median :0.07209   Median :0.03311   Median :0.1586  
##  Mean   :0.1893   Mean   :0.09938   Mean   :0.06264   Mean   :0.1811  
##  3rd Qu.:0.2462   3rd Qu.:0.12251   3rd Qu.:0.07170   3rd Qu.:0.2187  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000  
##  compactness_se     concavity_se     concave.points_se  symmetry_se    
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000    Min.   :0.0000  
##  1st Qu.:0.08132   1st Qu.:0.03811   1st Qu.:0.1447    1st Qu.:0.1024  
##  Median :0.13667   Median :0.06538   Median :0.2070    Median :0.1526  
##  Mean   :0.17444   Mean   :0.08054   Mean   :0.2235    Mean   :0.1781  
##  3rd Qu.:0.22680   3rd Qu.:0.10619   3rd Qu.:0.2787    3rd Qu.:0.2195  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000    Max.   :1.0000  
##  fractal_dimension_se  radius_worst    texture_worst    perimeter_worst 
##  Min.   :0.00000      Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.04675      1st Qu.:0.1807   1st Qu.:0.2415   1st Qu.:0.1678  
##  Median :0.07919      Median :0.2504   Median :0.3569   Median :0.2353  
##  Mean   :0.10019      Mean   :0.2967   Mean   :0.3640   Mean   :0.2831  
##  3rd Qu.:0.12656      3rd Qu.:0.3863   3rd Qu.:0.4717   3rd Qu.:0.3735  
##  Max.   :1.00000      Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##    area_worst      smoothness_worst compactness_worst concavity_worst  
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000    Min.   :0.00000  
##  1st Qu.:0.08113   1st Qu.:0.3000   1st Qu.:0.1163    1st Qu.:0.09145  
##  Median :0.12321   Median :0.3971   Median :0.1791    Median :0.18107  
##  Mean   :0.17091   Mean   :0.4041   Mean   :0.2202    Mean   :0.21740  
##  3rd Qu.:0.22090   3rd Qu.:0.4942   3rd Qu.:0.3025    3rd Qu.:0.30583  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000    Max.   :1.00000  
##  concave.points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000       Min.   :0.0000   Min.   :0.0000         
##  1st Qu.:0.2231       1st Qu.:0.1851   1st Qu.:0.1077         
##  Median :0.3434       Median :0.2478   Median :0.1640         
##  Mean   :0.3938       Mean   :0.2633   Mean   :0.1896         
##  3rd Qu.:0.5546       3rd Qu.:0.3182   3rd Qu.:0.2429         
##  Max.   :1.0000       Max.   :1.0000   Max.   :1.0000         
```

Dividing Training and Test Data

We take N_sample rows as test cases and use the rest to train the classifier.

```r
N_sample = 100
s_index = sample(1:569, N_sample)
data_Train = data_n[-s_index,]
data_Test = data_n[s_index,]
```

KNN

```
N_K = 25
data_knn = knn(train = data_Train, test = data_Test, cl = data[-s_index, 1], k = N_K)
```

Accuracy Analysis

The confusion matrix is shown below.

```
CrossTable(x = data[s_index,1], data_knn)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  100 
## 
## 
##                  | data_knn 
## data[s_index, 1] |         B |         M | Row Total | 
## -----------------|-----------|-----------|-----------|
##                B |        62 |         0 |        62 | 
##                  |    12.555 |    22.320 |           | 
##                  |     1.000 |     0.000 |     0.620 | 
##                  |     0.969 |     0.000 |           | 
##                  |     0.620 |     0.000 |           | 
## -----------------|-----------|-----------|-----------|
##                M |         2 |        36 |        38 | 
##                  |    20.484 |    36.417 |           | 
##                  |     0.053 |     0.947 |     0.380 | 
##                  |     0.031 |     1.000 |           | 
##                  |     0.020 |     0.360 |           | 
## -----------------|-----------|-----------|-----------|
##     Column Total |        64 |        36 |       100 | 
##                  |     0.640 |     0.360 |           | 
## -----------------|-----------|-----------|-----------|
## 
## 
```