# kNN in R Implementation

Bhagirath Kumar Lader

23/04/2020

```
#k-NN classification in R
wbcd <- read.table("https://raw.githubusercontent.com/bhagi8289/datasets/master/wisc_bc_data.csv",sep =
#wbcd[1:5]

#str(wbcd)
wbcd <- wbcd[-1]
#str(wbcd)
table(wbcd$diagnosis)
```

```
##
##   B   M
## 357 212
```

```
wbcd$diagnosis <- factor(wbcd$diagnosis,levels=c("B","M"),labels = c("Benign","Malignant"))
table(wbcd$diagnosis)
```

```
##
##    Benign Malignant
##       357       212
```

```
round(prop.table(table(wbcd$diagnosis))*100,digits = 1)
```

```
##
##    Benign Malignant
##      62.7      37.3
```

```
summary(wbcd[c("radius_mean","area_mean","smoothness_mean")])
```

```
##   radius_mean       area_mean      smoothness_mean
##  Min.   : 6.981   Min.   : 143.5   Min.   :0.05263
##  1st Qu.:11.700   1st Qu.: 420.3   1st Qu.:0.08637
##  Median :13.370   Median : 551.1   Median :0.09587
##  Mean   :14.127   Mean   : 654.9   Mean   :0.09636
##  3rd Qu.:15.780   3rd Qu.: 782.7   3rd Qu.:0.10530
##  Max.   :28.110   Max.   :2501.0   Max.   :0.16340
```

```
normalize <- function(x) {
  return((x-min(x))/(max(x)-min(x)))
}
normalize(c(1,2,3,4,5))
```

```
## [1] 0.00 0.25 0.50 0.75 1.00
```

```
normalize(c(10,20,30,40,50))
```

```
## [1] 0.00 0.25 0.50 0.75 1.00
```

```
wbcd_n <- as.data.frame(lapply(wbcd[2:31],normalize))
str(wbcd_n)
```

```
## 'data.frame':    569 obs. of  30 variables:
##  $ radius_mean      : num  0.253 0.171 0.192 0.203 0.389 ...
##  $ texture_mean     : num  0.0906 0.3125 0.2408 0.1245 0.1184 ...
##  $ perimeter_mean   : num  0.242 0.176 0.187 0.202 0.372 ...
##  $ area_mean        : num  0.136 0.0861 0.0974 0.1024 0.2411 ...
##  $ smoothness_mean  : num  0.453 0.399 0.497 0.576 0.244 ...
##  $ compactness_mean : num  0.155 0.292 0.18 0.289 0.153 ...
##  $ concavity_mean   : num  0.0934 0.1496 0.0714 0.1086 0.0795 ...
##  $ points_mean      : num  0.184 0.131 0.123 0.238 0.132 ...
##  $ symmetry_mean    : num  0.454 0.435 0.33 0.359 0.334 ...
##  $ dimension_mean   : num  0.202 0.315 0.283 0.227 0.115 ...
##  $ radius_se        : num  0.0451 0.1228 0.0309 0.0822 0.0242 ...
##  $ texture_se       : num  0.0675 0.1849 0.2269 0.2172 0.0116 ...
##  $ perimeter_se     : num  0.043 0.1259 0.0276 0.0515 0.0274 ...
##  $ area_se          : num  0.0199 0.0379 0.0126 0.0365 0.0204 ...
##  $ smoothness_se    : num  0.215 0.196 0.117 0.325 0.112 ...
##  $ compactness_se   : num  0.0717 0.252 0.0533 0.2458 0.0946 ...
##  $ concavity_se     : num  0.0425 0.0847 0.0267 0.0552 0.0392 ...
##  $ points_se        : num  0.235 0.259 0.142 0.372 0.173 ...
##  $ symmetry_se      : num  0.16 0.382 0.131 0.111 0.121 ...
##  $ dimension_se     : num  0.0468 0.0837 0.045 0.088 0.0301 ...
##  $ radius_worst     : num  0.198 0.141 0.159 0.142 0.294 ...
##  $ texture_worst    : num  0.0965 0.291 0.3843 0.0999 0.0989 ...
##  $ perimeter_worst  : num  0.182 0.139 0.147 0.13 0.269 ...
##  $ area_worst       : num  0.0894 0.0589 0.0703 0.0611 0.1558 ...
##  $ smoothness_worst : num  0.445 0.331 0.434 0.433 0.274 ...
##  $ compactness_worst: num  0.0964 0.2175 0.1173 0.1503 0.142 ...
##  $ concavity_worst  : num  0.0992 0.153 0.0852 0.0692 0.1088 ...
##  $ points_worst     : num  0.323 0.272 0.255 0.296 0.281 ...
##  $ symmetry_worst   : num  0.249 0.271 0.282 0.106 0.182 ...
##  $ dimension_worst  : num  0.0831 0.1366 0.1559 0.084 0.0828 ...
```

```
summary(wbcd_n[c("radius_mean","area_mean","smoothness_mean")])
```

```
##   radius_mean       area_mean      smoothness_mean
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2233   1st Qu.:0.1174   1st Qu.:0.3046
##  Median :0.3024   Median :0.1729   Median :0.3904
##  Mean   :0.3382   Mean   :0.2169   Mean   :0.3948
##  3rd Qu.:0.4164   3rd Qu.:0.2711   3rd Qu.:0.4755
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
wbcd_train <- wbcd_n[1:469,]
wbcd_test <- wbcd_n[470:569,]
wbcd_train_labels <- wbcd[1:469,1]
wbcd_test_labels <- wbcd[470:569,1]
library(class)
wbcd_pred <- knn(train = wbcd_train,test = wbcd_test, cl = wbcd_train_labels, k=21)
wbcd_pred[1:5]
```

```
## [1] Benign    Benign    Benign    Benign    Malignant
## Levels: Benign Malignant
```

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

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_pred
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        61 |         0 |        61 |
##                  |     1.000 |     0.000 |     0.610 |
##                  |     0.968 |     0.000 |           |
##                  |     0.610 |     0.000 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         2 |        37 |        39 |
##                  |     0.051 |     0.949 |     0.390 |
##                  |     0.032 |     1.000 |           |
##                  |     0.020 |     0.370 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        63 |        37 |       100 |
##                  |     0.630 |     0.370 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
table(wbcd_test_labels)
```

```
## wbcd_test_labels
##    Benign Malignant
##        61        39
```

```r
table(wbcd_pred)
```

```
## wbcd_pred
##    Benign Malignant
##        63        37
```

```r
# To improve model performance, instead of min-max scaling, we use z-score standardization
# we use built-in scale() function
wbcd_z <- as.data.frame(scale(wbcd[-1]))
summary(wbcd_z[c("radius_mean","area_mean","smoothness_mean")])
```

```
##   radius_mean        area_mean       smoothness_mean
##  Min.   :-2.0279   Min.   :-1.4532   Min.   :-3.10935
##  1st Qu.:-0.6888   1st Qu.:-0.6666   1st Qu.:-0.71034
##  Median :-0.2149   Median :-0.2949   Median :-0.03486
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.4690   3rd Qu.: 0.3632   3rd Qu.: 0.63564
##  Max.   : 3.9678   Max.   : 5.2459   Max.   : 4.76672
```

```r
wbcd_train <- wbcd_z[1:469,]
wbcd_test <- wbcd_z[470:569,]
wbcd_train_labels <- wbcd[1:469,1]
wbcd_test_labels <- wbcd[470:569,1]
library(class)
wbcd_pred <- knn(train = wbcd_train,test = wbcd_test, cl = wbcd_train_labels, k=21)
wbcd_pred[1:5]
```

```
## [1] Benign    Benign    Benign    Benign    Malignant
## Levels: Benign Malignant
```

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

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                  | wbcd_pred
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        61 |         0 |        61 |
```

```
##                  |     1.000 |     0.000 |     0.610 |
##                  |     0.924 |     0.000 |           |
##                  |     0.610 |     0.000 |           |
## -----------------|-----------|-----------|-----------|
##        Malignant |         5 |        34 |        39 |
##                  |     0.128 |     0.872 |     0.390 |
##                  |     0.076 |     1.000 |           |
##                  |     0.050 |     0.340 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |        66 |        34 |       100 |
##                  |     0.660 |     0.340 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
table(wbcd_test_labels)
```

```
## wbcd_test_labels
##    Benign Malignant
##        61        39
```

```
table(wbcd_pred)
```

```
## wbcd_pred
##    Benign Malignant
##        66        34
```

```
# The result has a worse performance
```