Classification using knn algorithm in R (Predicting cancer results)

> getwd()

[1] "/home/brajesh"

> setwd("/home/brajesh/machine-learning/class-knn-r")

> getwd()

[1] "/home/brajesh/machine-learning/class-knn-r"

> wbcd<-read.csv("wisc\_bc\_data.csv",stringsAsFactors = FALSE)

> str(wbcd)

'data.frame': 569 obs. of 32 variables:

$ id : int 87139402 8910251 905520 868871 9012568 906539 925291 87880 862989 89827 ...

$ diagnosis : chr "B" "B" "B" "B" ...

$ radius\_mean : num 12.3 10.6 11 11.3 15.2 ...

$ texture\_mean : num 12.4 18.9 16.8 13.4 13.2 ...

$ perimeter\_mean : num 78.8 69.3 70.9 73 97.7 ...

$ area\_mean : num 464 346 373 385 712 ...

$ smoothness\_mean : num 0.1028 0.0969 0.1077 0.1164 0.0796 ...

> wbcd

id diagnosis radius\_mean texture\_mean

1 87139402 B 12.320 12.39

2 8910251 B 10.600 18.95

3 905520 B 11.040 16.83

4 868871 B 11.280 13.39

5 9012568 B 15.190 13.21

6 906539 B 11.570 19.04

7 925291 B 11.510 23.93

8 87880 M 13.810 23.75

9 862989 B 10.490 19.29

10 89827 B 11.060 14.96

11 91485 M 20.590 21.24

> wbcd<-wbcd[,-1]

> wbcd

diagnosis radius\_mean texture\_mean perimeter\_mean

1 B 12.320 12.39 78.85

2 B 10.600 18.95 69.28

3 B 11.040 16.83 70.92

4 B 11.280 13.39 73.00

5 B 15.190 13.21 97.65

6 B 11.570 19.04 74.20

7 B 11.510 23.93 74.52

8 M 13.810 23.75 91.56

9 B 10.490 19.29 67.41

10 B 11.060 14.96 71.49

11 M 20.590 21.24 137.80

> str(wbcd)

'data.frame': 569 obs. of 31 variables:

$ diagnosis : chr "B" "B" "B" "B" ...

$ radius\_mean : num 12.3 10.6 11 11.3 15.2 ...

$ texture\_mean : num 12.4 18.9 16.8 13.4 13.2 ...

$ perimeter\_mean : num 78.8 69.3 70.9 73 97.7 ...

$ area\_mean : num 464 346 373 385 712 ...

$ smoothness\_mean : num 0.1028 0.0969 0.1077 0.1164 0.0796 ...

> table(wbcd$diagnosis)

B M

357 212

> bm<-table(wbcd$diagnosis)

> prop.table(bm)

B M

0.6274165 0.3725835

> prop.table(bm)\*100

B M

62.74165 37.25835

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

> 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)))}

> wbcd\_n<-as.data.frame(lapply(wbcd[2:31], normalize))

> 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]

> install.packages("class")

Installing package into ‘/home/brajesh/R/x86\_64-pc-linux-gnu-library/3.4’

> library(class)

> wbcd\_test\_pred<-knn(train = wbcd\_train, test = wbcd\_test, cl=wbcd\_train\_labels, k=21)

> install.packages("gmodels")

Installing package into ‘/home/brajesh/R/x86\_64-pc-linux-gnu-library/3.4’

> library(gmodels)

> CrossTable(x=wbcd\_test\_labels, y=wbcd\_test\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_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 | |

-----------------|-----------|-----------|-----------|

> wbcd\_z<-as.data.frame(scale(wbcd[-1]))

> wbcd\_train<-wbcd\_z[1:469,]

> wbcd\_test<-wbcd\_z[470:569,]

> wbcd\_test\_pred<-knn(train = wbcd\_train, test = wbcd\_test, cl=wbcd\_train\_labels, k=21)

> CrossTable(x=wbcd\_test\_labels, y=wbcd\_test\_pred, prop.chisq = FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| wbcd\_test\_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 | |

-----------------|-----------|-----------|-----------|