Breast Cancer

Shashank Semwal

25 June 2018

INSERT LIBRARIES

```
library(class)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 3.4.4
```

Import DATA

```
wdbc<-read.table("C:/Users/Shashank/Documents/R/dataset/breast cancer/wdbc.data",sep=',')
dim(wdbc)
```

```
## [1] 569  32
```

Removing the lables of the data & creating samples

```
wdbc_sample=sample(nrow(wdbc),size = nrow(wdbc)*.7)
wdbc_train=wdbc[wdbc_sample,-c(1,2)]
wdbc_test=wdbc[-wdbc_sample,-c(1,2)]
```

Standarizing the data

```
wdbc_std_train<-as.data.frame(lapply(wdbc_train,function(x) (x-min(x))/(max(x)-min(x))))
wdbc_std_test<-as.data.frame(lapply(wdbc_test,function(x) (x-min(x))/(max(x)-min(x))))
dim(wdbc_std_test)
```

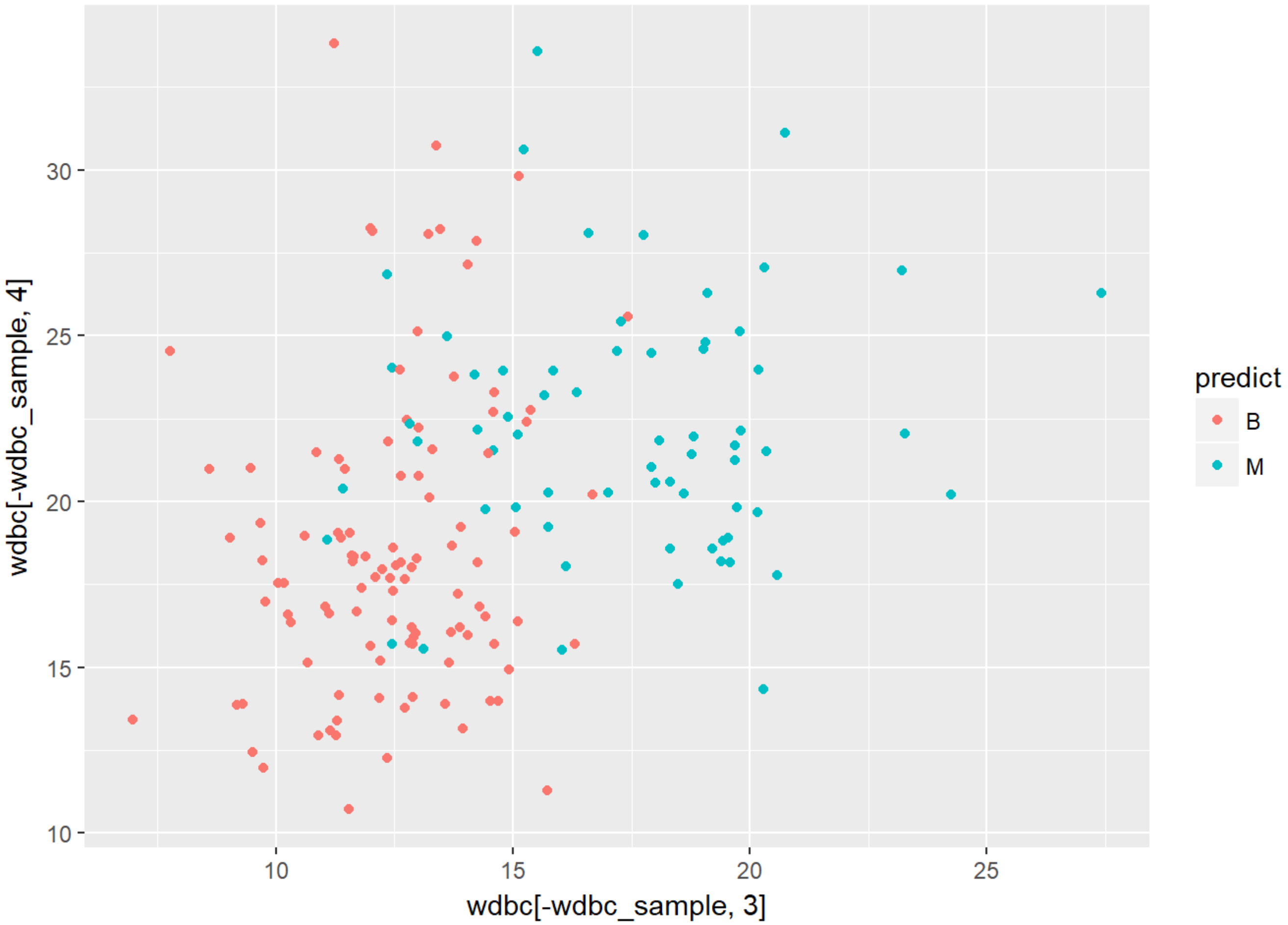
```
## [1] 171  30
```

```
predict<-knn(train =wdbc_std_train,test = wdbc_std_test,cl = wdbc[wdbc_sample,2],k = 3 )
```

```
confusionMatrix(predict,wdbc[-wdbc_sample,2])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 97  9
##          M  0 65
##
##                Accuracy : 0.9474
##                  95% CI : (0.9024, 0.9757)
##     No Information Rate : 0.5673
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8912
##  Mcnemar's Test P-Value : 0.007661
##
##             Sensitivity : 1.0000
##             Specificity : 0.8784
##          Pos Pred Value : 0.9151
##          Neg Pred Value : 1.0000
##              Prevalence : 0.5673
##          Detection Rate : 0.5673
##    Detection Prevalence : 0.6199
##       Balanced Accuracy : 0.9392
##
##        'Positive' Class : B
##
```

```
ggplot(,aes(wdbc[-wdbc_sample,3],wdbc[-wdbc_sample,4],col=predict))+
         geom_jitter(stat = 'identity')
```



CROSSTABLE

```
cross=CrossTable(wdbc[-wdbc_sample,2],predict)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  171 
## 
## 
##                       | predict 
## wdbc[-wdbc_sample, 2] |         B |         M | Row Total | 
## ----------------------|-----------|-----------|-----------|
##                     B |        97 |         0 |        97 | 
##                       |    22.610 |    36.871 |           | 
##                       |     1.000 |     0.000 |     0.567 | 
##                       |     0.915 |     0.000 |           | 
##                       |     0.567 |     0.000 |           | 
## ----------------------|-----------|-----------|-----------|
##                     M |         9 |        65 |        74 | 
##                       |    29.637 |    48.331 |           | 
##                       |     0.122 |     0.878 |     0.433 | 
##                       |     0.085 |     1.000 |           | 
##                       |     0.053 |     0.380 |           | 
## ----------------------|-----------|-----------|-----------|
##          Column Total |       106 |        65 |       171 | 
##                       |     0.620 |     0.380 |           | 
## ----------------------|-----------|-----------|-----------|
## 
## 
```

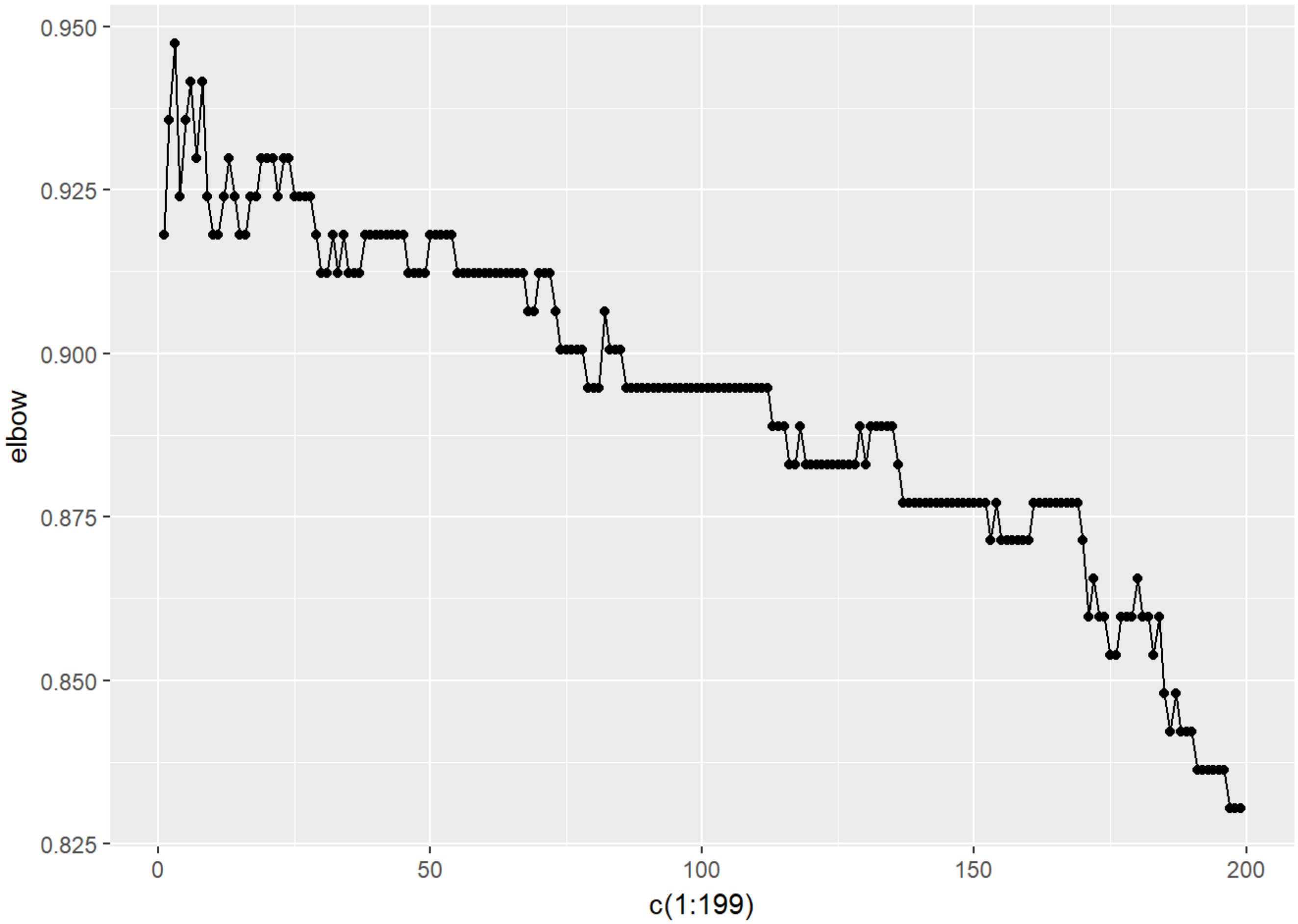
```
cross$t
```

```
##     y
## x    B  M
##   B 97  0
##   M  9 65
```

Elbow Chart

```
elbow<-c()
n=1
while (T) {
  predict<-knn(train =wdbc_std_train,test = wdbc_std_test,cl = wdbc[wdbc_sample,2],k = n )
  n=n+1
  cm=confusionMatrix(predict,wdbc[-wdbc_sample,2])
  elbow<-c(elbow,cm$overall[1])
  if(n==200){
    break
  }
}
```

```
ggplot(,aes(c(1:199),elbow))+geom_point(stat = 'identity')+
  geom_line()
```



```
# SO THE WE CAN CONCLUDE THAT WE CAN TAKE THE K VALUE FROM 25-50 IN ORDER TO GET BEST ACCURACY
```