Report on Breast Cancer Wisconsin (Prognostic) Data Set

Vijay Krishnavanshi

1 August 2016

DataSet

Dataset is from UCI Machine Learning Repository - Breast Cancer Wisconsin (Prognostic) Data Set. This dataset contains three files :

1) Wisconsin Breast Cancer Database
2) Wisconsin Diagnostic Breast Cancer (WDBC)
3) Wisconsin Prognostic Breast Cancer (WPBC)

This dataset is available online at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29.

Reading the dataset can be done similar to that of the CSV as the fields are "," separated and newline contains new records. Data is clean so not much of the preprocessing is required. File contain no headers.

```
WDBC_data=read.csv("wdbc.data",head=FALSE,sep = ",")
WPBC_data=read.csv("wpbc.data",head=FALSE,sep = ",")
bcw_data=read.csv("breast-cancer-wisconsin.data",head=FALSE,sep = ",")
```

Loading required libraries :

```
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
library(class)
```

Preprocessing the data

Using just the first dataset

```
bcw_data<-bcw_data[-(which(bcw_data$V7=="?")),]
```

Classification on Wisconsin Breast Cancer Database

```
Target Vaiable - Class ( 2 for Benign Tumor and 4 for Malignant Tumor )
```

```r
summary(bcw_data)
```

```
##        V1                  V2               V3               V4        
##  Min.   :   63375   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000  
##  1st Qu.:  877617   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000  
##  Median : 1171795   Median : 4.000   Median : 1.000   Median : 1.000  
##  Mean   : 1076720   Mean   : 4.442   Mean   : 3.151   Mean   : 3.215  
##  3rd Qu.: 1238705   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000  
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000  
##                                                                       
##        V5              V6               V7            V8        
##  Min.   : 1.00   Min.   : 1.000   1      :402   Min.   : 1.000  
##  1st Qu.: 1.00   1st Qu.: 2.000   10     :132   1st Qu.: 2.000  
##  Median : 1.00   Median : 2.000   2      : 30   Median : 3.000  
##  Mean   : 2.83   Mean   : 3.234   5      : 30   Mean   : 3.445  
##  3rd Qu.: 4.00   3rd Qu.: 4.000   3      : 28   3rd Qu.: 5.000  
##  Max.   :10.00   Max.   :10.000   8      : 21   Max.   :10.000  
##                                   (Other): 40                   
##        V9             V10              V11     
##  Min.   : 1.00   Min.   : 1.000   Min.   :2.0  
##  1st Qu.: 1.00   1st Qu.: 1.000   1st Qu.:2.0  
##  Median : 1.00   Median : 1.000   Median :2.0  
##  Mean   : 2.87   Mean   : 1.603   Mean   :2.7  
##  3rd Qu.: 4.00   3rd Qu.: 1.000   3rd Qu.:4.0  
##  Max.   :10.00   Max.   :10.000   Max.   :4.0  
## 
```

Data is skewed.

Preparing data for training a classifier

```r
### dimension of dataset
dim(bcw_data)
```

```
## [1] 683  11
```

```r
inTrain<-createDataPartition(y=bcw_data$V11,p=0.75,list=FALSE)
training<-bcw_data[inTrain,]
testing<-bcw_data[-inTrain,]
### dimension of the training set
dim(training)
```

```
## [1] 513  11
```

```r
### dimension of testing set
dim(testing)
```

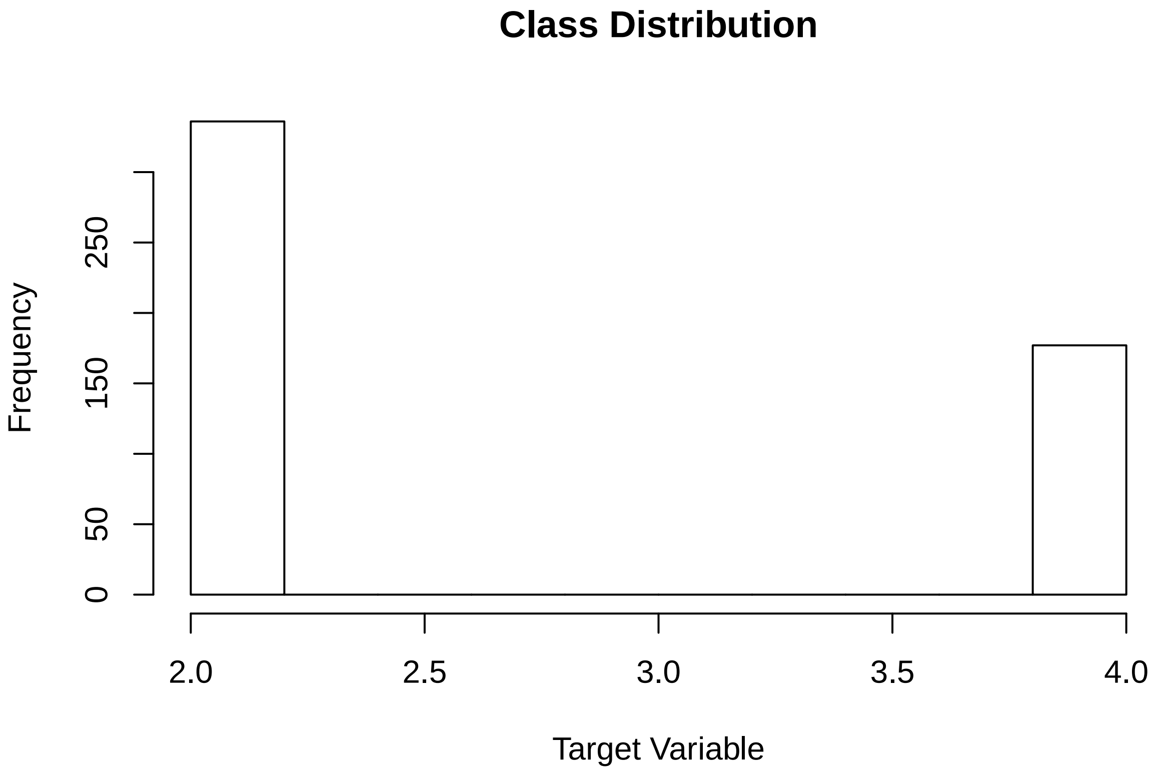
```
## [1] 170  11
```

Summary shows that data is skewedbut when we explore the target variable we see that the class is well balanced.

```r
summary(training)
```

```
##        V1                 V2               V3               V4
##  Min.   :   76389   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:  857774   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1165926   Median : 4.000   Median : 1.000   Median : 1.000
##  Mean   : 1063260   Mean   : 4.431   Mean   : 3.099   Mean   : 3.136
##  3rd Qu.: 1231387   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##        V5               V6               V7            V8
##  Min.   : 1.000   Min.   : 1.000   1      :296   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 2.000   10     :103   1st Qu.: 2.000
##  Median : 1.000   Median : 2.000   2      : 22   Median : 3.000
##  Mean   : 2.877   Mean   : 3.193   3      : 22   Mean   : 3.454
##  3rd Qu.: 4.000   3rd Qu.: 4.000   5      : 22   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :10.000   4      : 17   Max.   :10.000
##                                    (Other): 31
##        V9              V10              V11
##  Min.   : 1.000   Min.   : 1.000   Min.   :2.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
##  Median : 1.000   Median : 1.000   Median :2.00
##  Mean   : 2.865   Mean   : 1.561   Mean   :2.69
##  3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
##  Max.   :10.000   Max.   :10.000   Max.   :4.00
##
```

```r
hist(training$V11,main = "Class Distribution" , xlab = "Target Variable" )
```



##

Techniques that are available for Classification

1. Quadratic Discrimnant Analysis
2. Linear Discriminant Analysis
3. K - Nearest Neighbour - 1
4. K - Nearest Neighbour - 3
5. Logistic Regression

Logistic Regression and KNN was tried as they do not assume the gaussian distribution of the data. Variables being skewed this condition was not staisfied

```
set.seed(1)
knn.pred1=knn(training[,2:10],testing[,2:10],training[,11],k=1)
p=table(knn.pred1,testing[,11])
print(p)
```

```
##
## knn.pred1   2   4
##         2 105   6
##         4   3  56
```

```
accuracy=(p[1,1]+p[2,2])/length(testing$V11)
accuracy
```

```
## [1] 0.9470588
```

```
knn.pred2=knn(training[,2:10],testing[,2:10],training[,11],k=3)
p=table(knn.pred2,testing[,11])
print(p)
```

```
##
## knn.pred2   2   4
##         2 105   4
##         4   3  58
```

```
accuracy=(p[1,1]+p[2,2])/length(testing$V11)
accuracy
```

```
## [1] 0.9588235
```

K - Nearest Neighbour was selected as the classifier was erring more on the False Positive side on the False Negative Side.

Clustering

Techniques available:

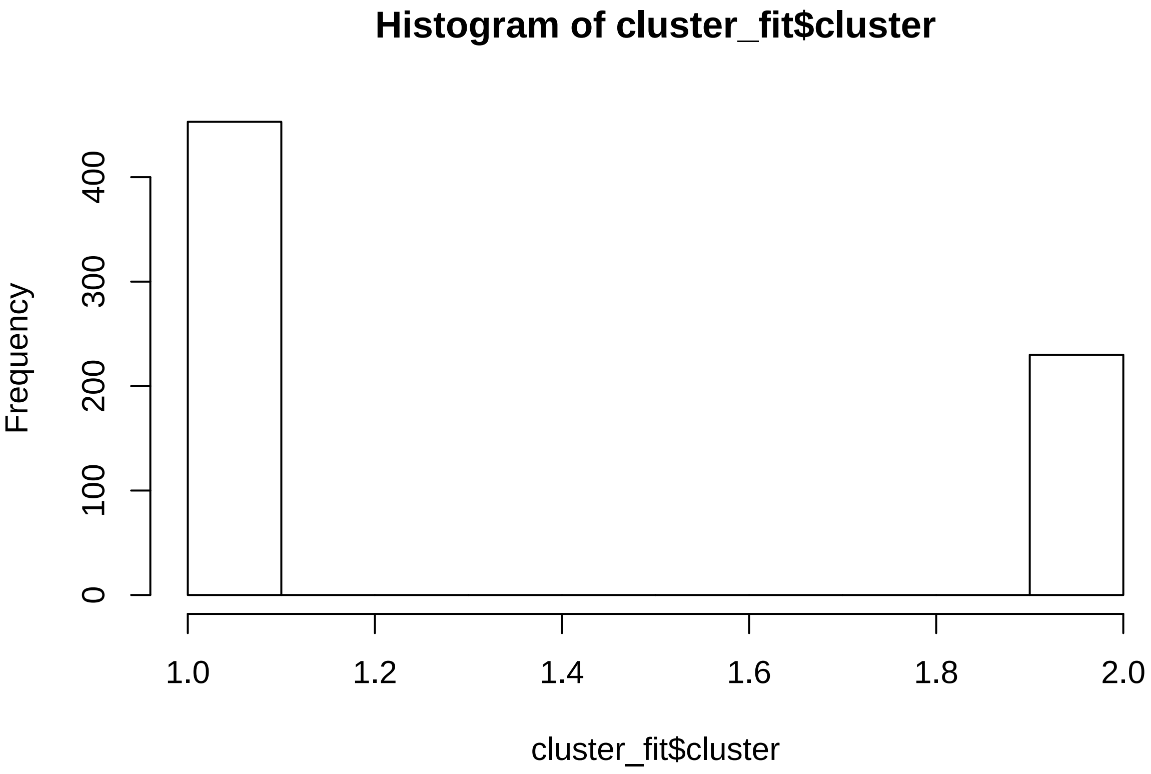
1. Principle Component Analysis
2. K - Means Algorithm

- Hartigan-Wong
- Lloyd
- Forgy
- MacQueen

Not used PCA as the number of feature was much less than the number of data points and k - means can better handle skewed data.

Data points were spanning over small range of values so Variability was also not an issue

```
cluster_fit=kmeans(bcw_data[,2:10],2,iter.max = 100)
hist(cluster_fit$cluster)
```



```
res=cluster_fit$cluster
p=table(ifelse(res==1,2,4),bcw_data[,11])
print(p)
```

```
##
##       2   4
##   2 435  18
##   4   9 221
```

```
accuracy=(p[1,1]+p[2,2])/length(cluster_fit$cluster)
accuracy
```

```
## [1] 0.9604685
```

Regression

```
summary(WPBC_data)
```

```
##        V1          V2            V3                V4
##  Min.   :   8423   N:151   Min.   :  1.00   Min.   :10.95
##  1st Qu.: 855745   R: 47   1st Qu.: 14.00   1st Qu.:15.05
##  Median : 886339           Median : 39.50   Median :17.29
##  Mean   :1990469           Mean   : 46.73   Mean   :17.41
##  3rd Qu.: 927996           3rd Qu.: 72.75   3rd Qu.:19.58
##  Max.   :9411300           Max.   :125.00   Max.   :27.22
##
##        V5              V6                V7                V8
##  Min.   :10.38   Min.   : 71.90   Min.   : 361.6   Min.   :0.07497
##  1st Qu.:19.41   1st Qu.: 98.16   1st Qu.: 702.5   1st Qu.:0.09390
##  Median :21.75   Median :113.70   Median : 929.1   Median :0.10190
##  Mean   :22.28   Mean   :114.86   Mean   : 970.0   Mean   :0.10268
##  3rd Qu.:24.66   3rd Qu.:129.65   3rd Qu.:1193.5   3rd Qu.:0.11098
##  Max.   :39.28   Max.   :182.10   Max.   :2250.0   Max.   :0.14470
##
##        V9               V10                V11                V12
##  Min.   :0.04605   Min.   :0.02398   Min.   :0.02031   Min.   :0.1308
##  1st Qu.:0.11020   1st Qu.:0.10685   1st Qu.:0.06367   1st Qu.:0.1741
##  Median :0.13175   Median :0.15135   Median :0.08607   Median :0.1893
##  Mean   :0.14265   Mean   :0.15624   Mean   :0.08678   Mean   :0.1928
##  3rd Qu.:0.17220   3rd Qu.:0.20050   3rd Qu.:0.10393   3rd Qu.:0.2093
##  Max.   :0.31140   Max.   :0.42680   Max.   :0.20120   Max.   :0.3040
##
##       V13               V14                V15               V16
##  Min.   :0.05025   Min.   :0.1938   Min.   :0.3621   Min.   : 1.153
##  1st Qu.:0.05672   1st Qu.:0.3882   1st Qu.:0.9213   1st Qu.: 2.743
##  Median :0.06171   Median :0.5333   Median :1.1685   Median : 3.767
##  Mean   :0.06271   Mean   :0.6033   Mean   :1.2645   Mean   : 4.255
##  3rd Qu.:0.06671   3rd Qu.:0.7509   3rd Qu.:1.4632   3rd Qu.: 5.213
##  Max.   :0.09744   Max.   :1.8190   Max.   :3.5030   Max.   :13.280
##
##       V17              V18                V19                 V20
##  Min.   : 13.99   Min.   :0.002667   Min.   :0.007347   Min.   :0.01094
##  1st Qu.: 35.37   1st Qu.:0.005001   1st Qu.:0.019803   1st Qu.:0.02681
##  Median : 58.45   Median :0.006193   Median :0.027880   Median :0.03691
##  Mean   : 70.23   Mean   :0.006762   Mean   :0.031199   Mean   :0.04075
##  3rd Qu.: 92.48   3rd Qu.:0.007973   3rd Qu.:0.038335   3rd Qu.:0.04897
##  Max.   :316.00   Max.   :0.031130   Max.   :0.135400   Max.   :0.14380
##
##       V21                V22                V23                V24
##  Min.   :0.005174   Min.   :0.007882   Min.   :0.001087   Min.   :12.84
##  1st Qu.:0.011423   1st Qu.:0.014795   1st Qu.:0.002748   1st Qu.:17.63
##  Median :0.014175   Median :0.017905   Median :0.003719   Median :20.52
##  Mean   :0.015099   Mean   :0.020555   Mean   :0.003987   Mean   :21.02
##  3rd Qu.:0.017665   3rd Qu.:0.022880   3rd Qu.:0.004630   3rd Qu.:23.73
##  Max.   :0.039270   Max.   :0.060410   Max.   :0.012560   Max.   :35.13
##
##       V25             V26             V27              V28
##  Min.   :16.67   Min.   : 85.1   Min.   : 508.1   Min.   :0.08191
##  1st Qu.:26.21   1st Qu.:118.1   1st Qu.: 947.3   1st Qu.:0.12932
##  Median :30.14   Median :136.5   Median :1295.0   Median :0.14185
##  Mean   :30.14   Mean   :140.3   Mean   :1405.0   Mean   :0.14392
##  3rd Qu.:33.55   3rd Qu.:159.9   3rd Qu.:1694.2   3rd Qu.:0.15488
##  Max.   :49.54   Max.   :232.2   Max.   :3903.0   Max.   :0.22260
##
##       V29               V30               V31               V32
##  Min.   :0.05131   Min.   :0.02398   Min.   :0.02899   Min.   :0.1565
##  1st Qu.:0.24870   1st Qu.:0.32215   1st Qu.:0.15265   1st Qu.:0.2759
##  Median :0.35130   Median :0.40235   Median :0.17925   Median :0.3103
##  Mean   :0.36510   Mean   :0.43669   Mean   :0.17878   Mean   :0.3234
##  3rd Qu.:0.42368   3rd Qu.:0.54105   3rd Qu.:0.20713   3rd Qu.:0.3588
##  Max.   :1.05800   Max.   :1.17000   Max.   :0.29030   Max.   :0.6638
##
##       V33               V34              V35
##  Min.   :0.05504   Min.   : 0.400   0      :87
##  1st Qu.:0.07658   1st Qu.: 1.500   1      :35
##  Median :0.08689   Median : 2.500   2      :17
##  Mean   :0.09083   Mean   : 2.847   4      :10
##  3rd Qu.:0.10138   3rd Qu.: 3.500   13     : 6
##  Max.   :0.20750   Max.   :10.000   7      : 6
##                                     (Other):37
```

Preparing data for regression :

```
inTrain1<-createDataPartition(y=WPBC_data$V3,p=0.75,list=FALSE)
training<-WPBC_data[inTrain,]
testing<-WPBC_data[-inTrain,]
```

Fitting the model 1 with pca :

```
dat <- training[,2:34]

lmFit1 <- train(V3~., dat, method = "lm", preProcess=c("pca"),
                 trControl = trainControl(method = "cv"))
lmFit1.pred<-predict(lmFit1,testing[2:34])
sqrt(sum((lmFit1.pred-testing[,3])^2)/94)
```

```
## [1] 22.79267
```

Fitting the model 2 without pca:

```
lmFit2 <- train(V3~., dat, method = "lm",
                  trControl = trainControl(method = "cv"))
lmFit2.pred<-predict(lmFit2,testing[2:34])
sqrt(sum((lmFit2.pred-testing[,3])^2)/94)
```

```
## [1] 23.51304
```

As the data has many correlated features PCA is used to reduce the number of features for better and improved prediction but the results does not support the arguement

Cross validation was used to better generalise the error

So fit that give less standard error was selected as they give better result.