Applied Data Mining: Homework #3

Due on September 16th 2017

Instructor: Hasan Kurban

Naimesh Chaudhari

September 16, 2017

In this homework, you will work with Ionosphere Data Set to answer some questions regarding Principal Component Analysis (PCA), exploratory data analysis and k-means clustering. Here is the beginning of an R session that allows us to read this data from the web into our local R session:

```
> install.packages("data.table")
> library(data.table)
> install.packages("curl")
> mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
                   ionosphere/ionosphere.data")
```

Problem 1

For the Ionosphere Data Set, answer the following questions:

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?

This is radar data collected by Goose Bay, Labrador. 16 high frequency antennas with a total power of 6.4 kilowatts were used. These antennas were measuring the free electros in the ionosphere. Two results were determined good or bad, good radar returns are those showing evediane of some structure while bad are those that do not. The received signal was processed by an autocorrelation function which used the metrics time of pulse and the pulse number (17 different numbers). Each spulse number had two attributes hence 34 columns + the results column.

R Code

Using R, show code that answers the following questions:

1. How many entries are in the data set? There are 351 rows in this data set ...

Listing 1: Sample R Script With Highlighting

```
nrow(mydata)
```

2. How many unknown or missing data are in the data set? There are no missing values in this data set...

Listing 2: Sample R Script With Highlighting

```
table(is.na(mydata))
```

3. Create a bar plot of 1st, 2nd, 35th variables. Label the plots properly. Discuss the distribution of values *e.g.*, are uniform, skewed, normal. Place images of these bar plots into the document. Show the R code that you used below and discussion below that.

Problem 1 continued on next page...

Listing 3: Sample R Script With Highlighting

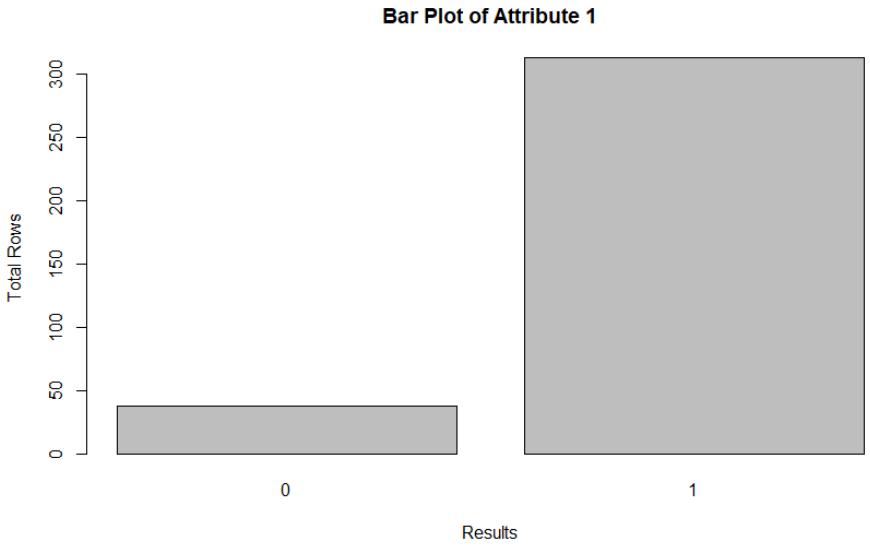
```
barplot(table(mydata$V1), main="Bar Plot of Attribute 1", xlab = "Results",
    ylab = "Total Rows" )
barplot(table(mydata$V2), main="Bar Plot of Attribute 2", xlab = "Results",
    ylab = "Total Rows" )
barplot(table(mydata$V35),main="Bar Plot of Attribute 35", xlab = "Results",
    ylab = "Total Rows" )
```

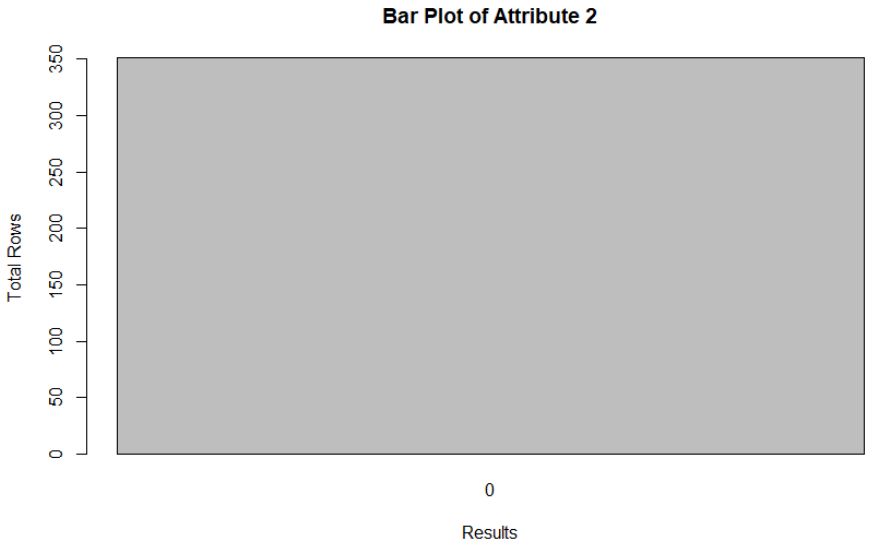
Discussion of Bar Plots

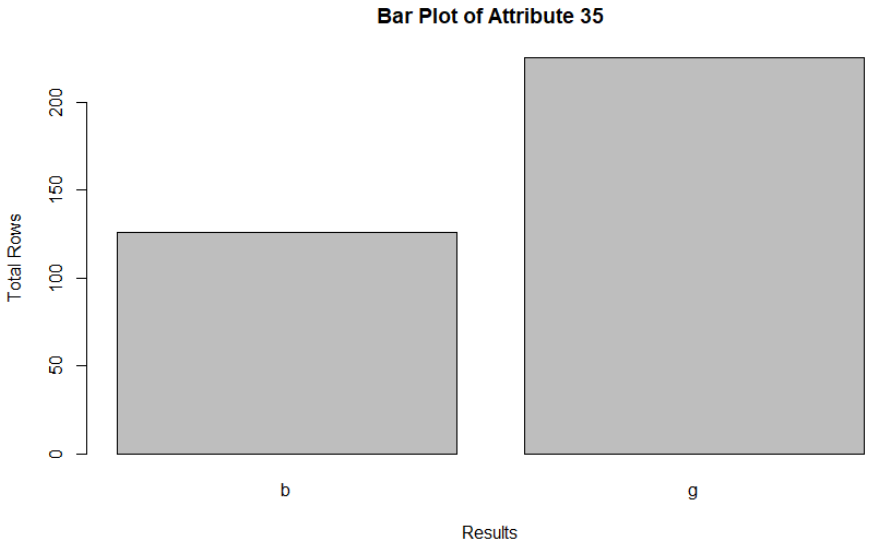
The 1s attribute is definately skewed towards the result of 1, the second is all zeros which is definately one sided, and attribute 35 is skewed towards a good result but not heavily. ...

Bar Plots

Below are the bar plots of the above discussion.







4. Make a scatter plots of $[V22, V20]$ and $[V1, V2]$ variables and color the data points with the class variable $[V35]$. Discuss the plots, i.e., do you observe any relationships between variables?

Listing 4: Sample R Script With Highlighting

```
plot(mydata$V22, mydata$V20
     ,main = "ScatterPlot V22 vs V20, Class V35", xlab = "V22", ylab = "V20"
     ,col = as.factor(mydata$V35))
legend(x = "topright", legend = levels(as.factor(mydata$V35)), col = c("black"
    , "red"), pch = 1)

plot(mydata$V1, mydata$V2
     ,main = "ScatterPlot V1 vs V2, Class V35", xlab = "V1", ylab = "V2"
     ,col = as.factor(mydata$V35))
```

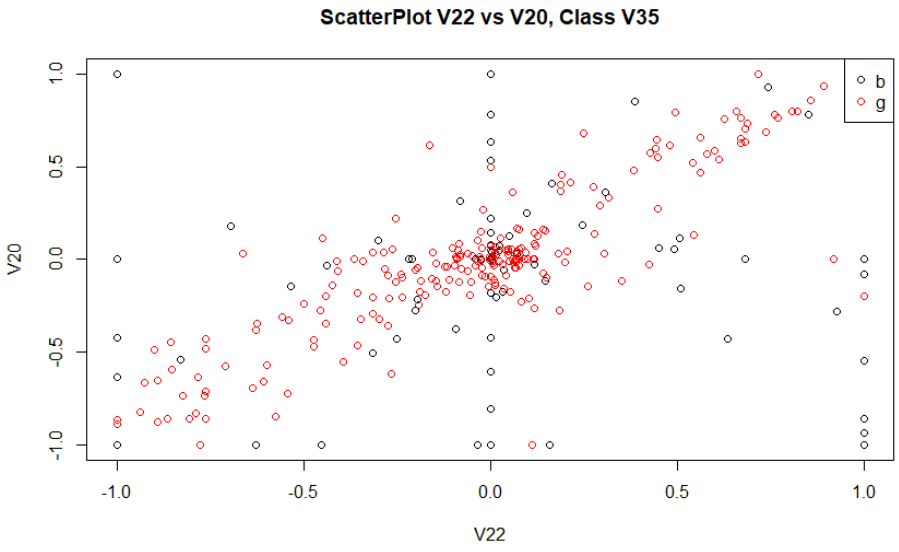
```
legend(x = "topright", legend = levels(as.factor(mydata$V35)), col = c("black"
    , "red"), pch = 1)
```

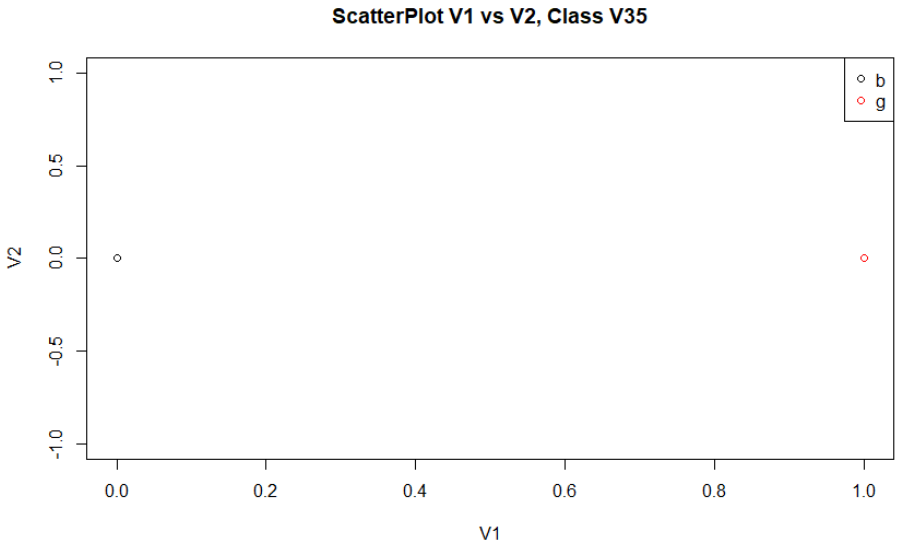
Discussion of Scatter Plots

V22 and V20 seem to have a posetive linear relationship. As V22 goes up so does V20. It seems we have good results in both ends of intersection, but when both are 0 we see a many more good results.

V2 does not seem to help the result of v1, if v1 = 0, then most likely a bad result while if v1 = 1 then most likely a good result.

Scatter Plots





Problem 2

In this question, you will run k-means clustering algorithm against Ionosphere data set. The input data for k-means is $mydata[, -35]$ – removing the class variable since this is a clustering task.

R Code

Using R, show code that answers the following questions:

1. Run "Lloyd, Forgy and Hartigan-Wong's" heuristic algorithms for k-means and report total within sum of squared error (SSE) for $k = 2$ and $nstart = 50$. Compare the results?i.e., which/why is better? Discuss $nstart$ parameter. Show the R code that you used below and discussion and results below that.

Listing 5: Sample R Script With Highlighting

```
kl <- kmeans(mydata[,-35],centers=2,nstart=50, algorithm = "Lloyd")
kf <- kmeans(mydata[,-35],centers=2,nstart=50, algorithm = "Forgy")
kh <- kmeans(mydata[,-35],centers=2,nstart=50, algorithm = "Hartigan-Wong")


kl['tot.withinss']
kf['tot.withinss']
kh['tot.withinss']
```

Total SSE

Total SSE Lloyd = 2419.365
Total SSE Forgy = 2419.365
Total SSE Hartigan-Wong's = 2419.365

Discussion of nstart and Results

It seems that all three algorithms returned the same SSE so for this test test it does not matter which one we use. Generallly we would like to pick the results set that provides the lowerst SSE.

The nstart variable picks k points randomly nstart times and tries to cluster the data points with the closest distance together. Once this is done it returns the result set with the smallest SSE.

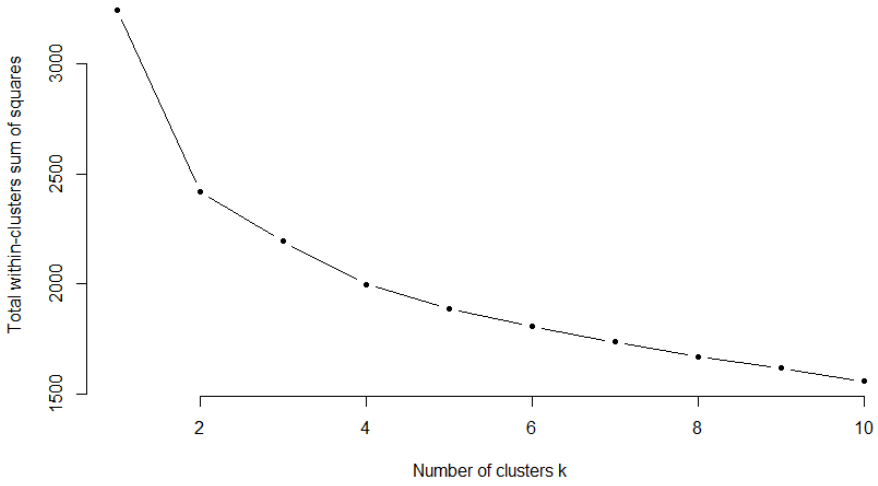
2. Elbow method is a technique used to decide optimal cluster number. The code below gives a plot of total SSE for $k = 1, \ldots, 10$. Discuss the elbow technique, i.e., what would be the optimal k based on the plot, can optimal k always be identified by elbow method?

```
> k_max <- 10
#total SSE
> tsse <- sapply(1:k_max,
+                function(k){kmeans(mydata, k, nstart=30,iter.max = 12 )
                                    $tot.withinss})
> tsse
```

```
 [1] 3243.103 2419.365 2193.320 1998.581
  1889.717 1806.150 1737.575 1668.753 1617.829 1550.105
> plot(1:k_max, tsse,
+      type="b", pch = 20, frame = FALSE,
+      xlab="Number of clusters k",
+      ylab="Total within-clusters sum of squares")
>
```

Discussion of Results

The elbow method graphs the total within clusters sum of square vs the number of cluster, you generally want to pick the location where the graph make its first elbow. Based on the plot we would choose 2 cluster as the optimal points. This technique is usefull but it cannot be always used to determind the number of clusters.



Problem 3

Use Principal Component Analysis (PCA) over Ionosphere Data Set to answer the below questions. You may want to use either "*princomp()*" or "*prcomp()*" functions in R. In this question, remove the 2nd (all 0s) and 35th variable (class variable) before using PCA.

```
> mydata <- mydata[,-35]
> mydata <- mydata[,-2]
> dim(mydata)
[1] 351  33
> mydata.pca <- prcomp(mydata, scale =TRUE)
```

R Code

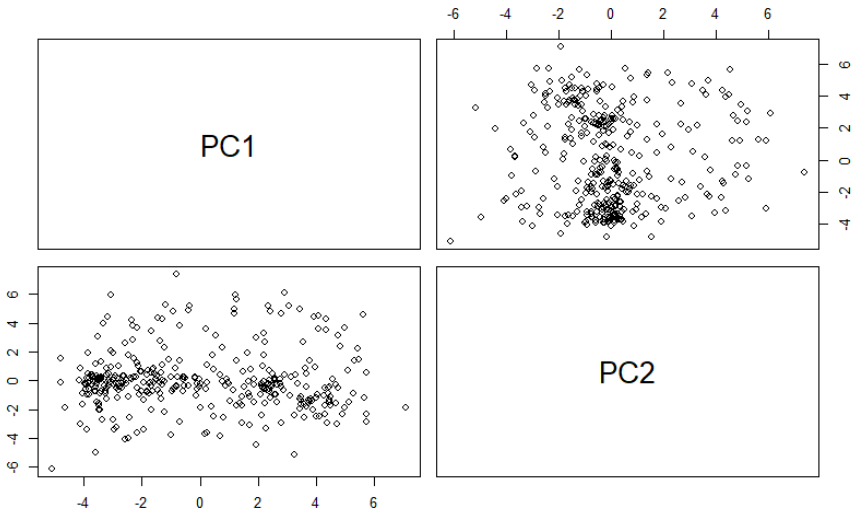
Using R, show code that answers the following questions:

1. Make a scatter plot of PC1 and PC2 (the first and second principal components). Discuss principal components? What is PC1 and PC2? Show the R code that you used below and the scatter plot and discussion below that

Listing 6: Sample R Script With Highlighting

```
pairs(mydata.pca$x[,1:2])
```

Scatter Plot



Discussion of Principal Components

PCA is a dimension reduction technique, that works by taking corellated data and doing test to get uncorallated data. In doing this we generally will reduce the number of variables we need to test.

PC1 is the first principle componenet in the new dataset, and PC2 is the second.

2. You can observe the loadings as follows (using *prcomp() function*):

```
>mydata.pca$rotation
```

Discuss loadings in PCA?i.e., how are principal components and original variables of the data (mydata) related? (loadings(mydata.pca) if *princomp()* is used)

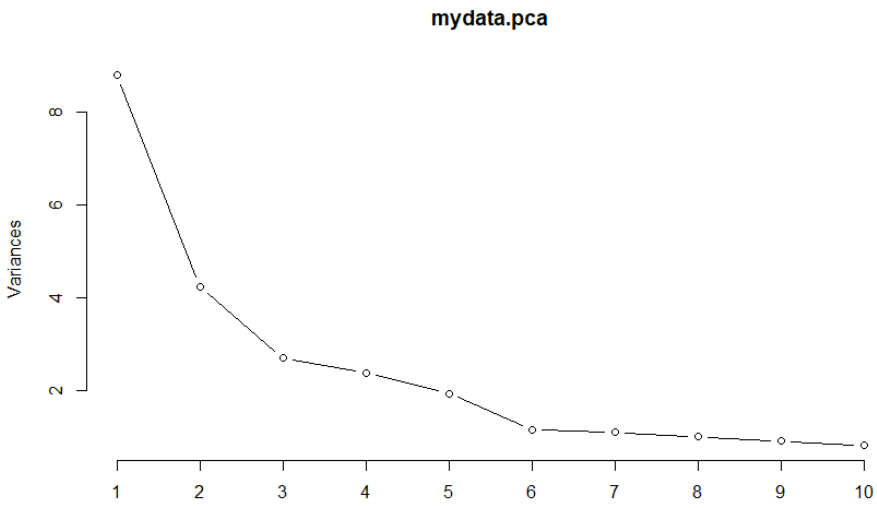
Each PC has a information from all the attributes in the table within itself. This is done so that we can minimize and pick the optimal number of principal components.

For example PC1, has info of v1 to v31

3. Scree plot is among the most popular methods to decide optimal dimension number.

```
> plot(mydata.pca, type = "l")
> screeplot(mydata.pca)
```

What is the optimal dimension number (d) for this data set? How much of the variation is kept with your optimal d? Discuss the results.



I would select 6 as the optimal d. By selecting 6 d we are capturing a majority of the variance, and lowering the total attributes from 34 to 6!