Applied Data Mining: Homework #3

Due on 9/16/2017

Instructor: Hasan Kurban

Keith Hickman

October 3, 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 dataset records observations of radar signatures of free electrons in the ionosphere. The set's purpose is to measure whether the radar signatures were detecting the electrons or passing through.

R Code

Using R, show code that answers the following questions:

1. How many entries are in the data set? A: There are 351 observations . . .

Listing 1: Sample R Script With Highlighting

```r
install.packages("Hmisc")
library(Hmisc)

mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
    ionosphere/ionosphere.data")
describe(mydata)
```

2. How many unknown or missing data are in the data set? . . . A: There are no missing variables in the dataset.

Listing 2: Sample R Script With Highlighting

```r
install.packages("Hmisc")
library(Hmisc)

mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
    ionosphere/ionosphere.data")
describe(mydata)
```

```
35  Variables      351  Observations
--------------------------------------------------------------------------------------
V1
```

Problem 1 continued on next page. . .

```
       n  missing distinct     Info      Sum     Mean      Gmd
     351        0        2     0.29      313   0.8917   0.1936

              ...
```

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.

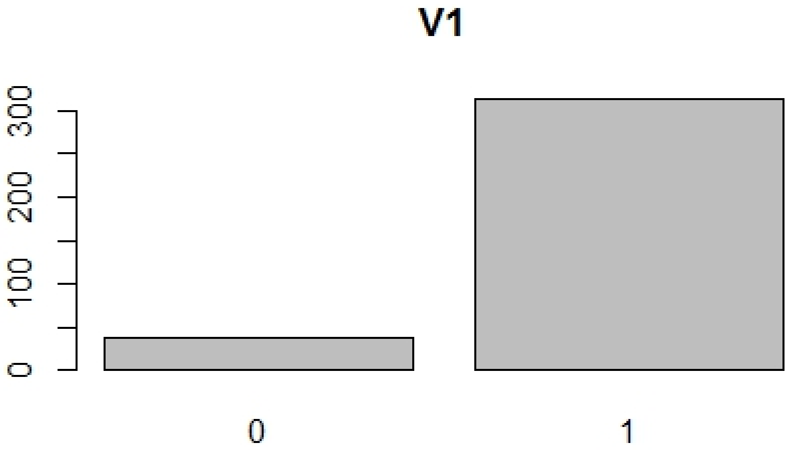
Listing 3: Sample R Script With Highlighting

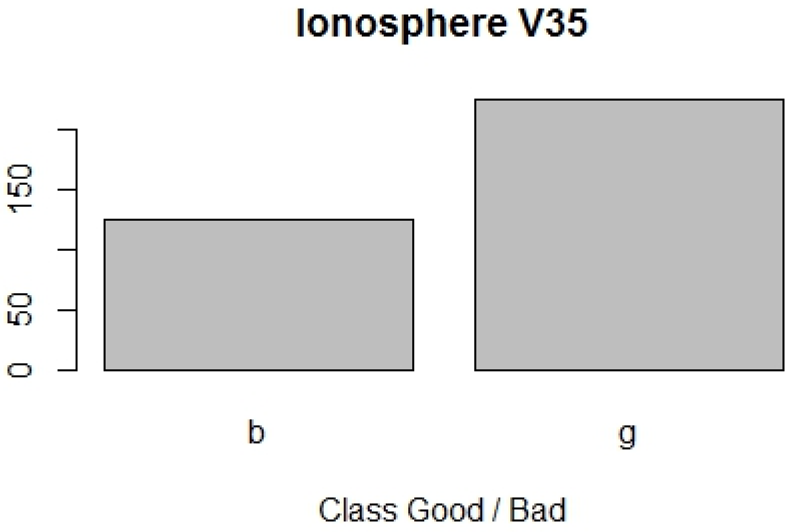
```
plot(mydata$v20, mydata$V22, col=mydata$V35,
     main="Relationship between A1 and A2 by Season",
     xlab="V22", ylab="V20")
```

Discussion of Bar Plots

A: Unfortunately, the dataset documentation is somewhat sparse. For variables V1 and V2, it seems there is a highly skewed distribution. V1 is 300 observations of 1 vs. 50 of 0. V2 is homogenous. Would that count as a uniform distribution? Either way, there is not much information contained in V2. V35 is the class or target variable, and is skewed toward toward the positive class - 1. ...

Bar Plots





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?

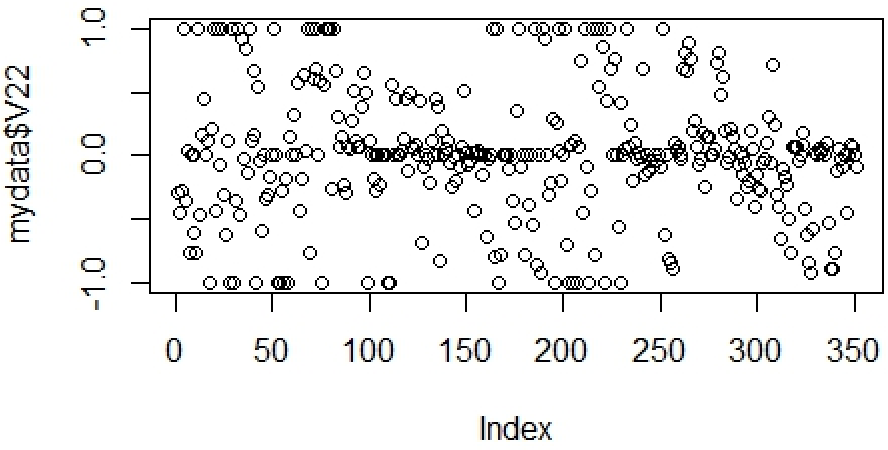
Listing 4: Sample R Script With Highlighting

```
plot(mydata$v20, mydata$V22, col=mydata$V35,
     main="Relationship between V20 and V22",
     xlab="V22", ylab="V20")
plot(mydata$v1, mydata$V2, col=mydata$V35,
     main="Relationship between V1 and V2 by Season",
     xlab="V22", ylab="V20")
```

Discussion of Scatter Plots

I was able to produce a basic scatter plot of the X, Y variables V20 and V22 respectively. I was not able to use the ggplot library's color feature to color by the target variable V35. I kept getting error messages indicating that V35 wasn't a valid color variable. However, the text states on pg. 104 that R should convert those into integers. The relationship between the two variablesV20 and V22 is linear. The relationship seems to be centered around zero . . .

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

```
mydata <- mydata[,-35]
head(mydata)
kmeans1 <- kmeans(mydata,2,nstart=50)
kmeans1$tot.withinss

```

```
k_max <- 10
#total SSE
tsse <- sapply(1:k_max, function(k){kmeans(mydata, k, nstart=30,iter.max = 12
    )$tot.withinss})
tsse
plot(1:k_max, tsse, type="b", pch = 20, frame = FALSE, xlab="Number of
    clusters k",
     ylab="Total within-clusters sum of squares")
```

Total SSE

A: The total within SSE is 2419.365. Note: I was confused by the reference to Lloyd, Forgy, and Hartigan-Wong. I couldn't find it anywhere in the text. Were they the progentors of heuristics for k-means in R?

Discussion of nstart and Results

Adjusting n-start upwards increases the number of iterations the k-means algorithm would go through, thereby potentially decreasing the SSE. . . .

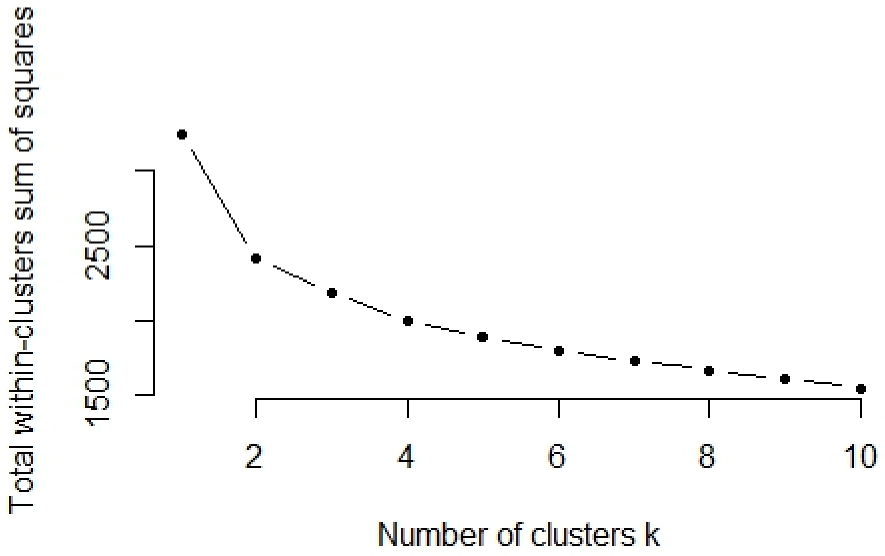
2. Elbow method is a technique used to decide optimal cluster number. The code below gives a plot of total SSE for $k = 1, \ldots, 10$. Discuss the elbow technique, i.e., what would be the optimal k based on the plot, can optimal k always be identified by elbow method?

```
> 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

I noticed that the biggest increase in SSE included within clusters is obviously from 1 to 2, then diminishes with each additional cluster. How do we determine the number of optimal clusters? I would consider the number of meaningful clusters as the main driver e.g. the number that we could act upon. If I had a customer segmentation problem, I would want to know the number of groups I could reasonably segment my customer base into.

Problem 2 continued on next page. . .



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

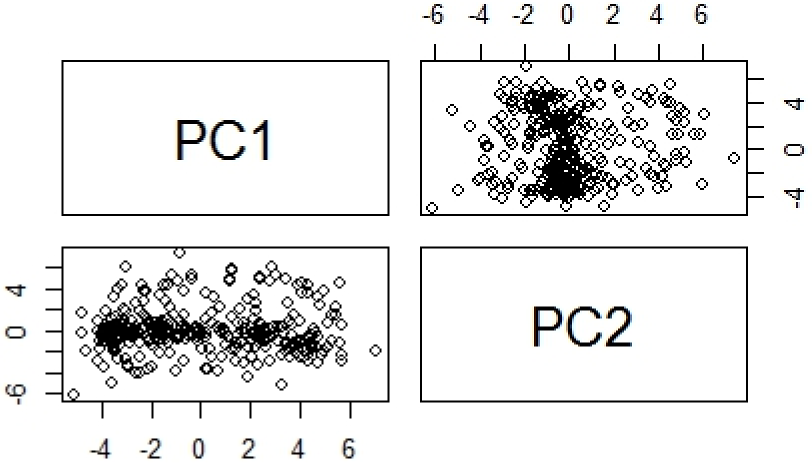
Using R, show code that answers the following questions:

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

```
mydata <- mydata[,-35]
mydata <- mydata[,-2]
head(mydata)
mydata.pca <- prcomp(mydata, scale=TRUE)
z <- mydata.pca$x[,1:2]
pairs(z)
```

Scatter Plot



Discussion of Principal Components

PCA 1 and 2 are linear combinations of the variables in the mydata dataset. They each explain a decreasing amount of the variance in the dataset. Here, the PC1 explains 26 percent of the variance, followed by 12 percent, down to 1 and 2 percent. There are roughly 11 or 12 PCs that I would include to explain most of the variance while still reducing dimensionality.

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) The original variables differ significantly in their contributions to the PCA.

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. A: If I look at the elbow plot, I would estimate that three (3) PCs would be the optimum number of dimensions, as more than three decreases the percent of variance explained by a significant amount. However, when I look at the scree plot, I would increase the number of PCs to five (5) to accomodate the max variance.