Applied Data Mining: Homework #3

Due on Fill-in this please

Instructor: Hasan Kurban

Student Name

September 4, 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?

R Code

Using R, show code that answers the following questions:

1. How many entries are in the data set? Answer here ...

Listing 1: Sample R Script With Highlighting

```
%% You provide code here %%
```

2. How many unknown or missing data are in the data set? Answer here ...

Listing 2: Sample R Script With Highlighting

```
%% You provide code here %%
```

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

```
%% You provide code here %%
```

Discussion of Bar Plots

Answer here...

Bar Plots

Place images here with suitable captions.

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

```
%% You provide code here %%
```

Discussion of Scatter Plots

Answer here...

Scatter Plots

Place images here with suitable captions.

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

```
%% You provide code here %%
```

Total SSE

Answer here...

Discussion of nstart and Results

Answer here...

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

Answer here...

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

```
%% You provide code here %%
```

Scatter Plot

Place images here with suitable captions. Answer here...

Problem 3 continued on next page...

Discussion of Principal Components

Answer here...

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)

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.