# Homework 4
# Introduction to Data Analysis and Mining
# Spring 2018
# CSCI-B 365

Instructor: Hasan Kurban

April 24, 2018

## Directions

Please follow the syllabus guidelines in turning in your homework. I am providing the LaTeX of this document too. This homework is due Friday, March 16, 2018 10:00p.m. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. All the work should be your own.

## Problem 1 [30 points]

In this question, you will first perform principal component analysis (PCA) over Ionosphere Data Set and then cluster the reduced data using your $k$-means program ($C_k$) from previous homework. You are allowed to use R packages for PCA and ignore the class variables (35th variable) while performing PCA. Answer the questions below:

**1.1)** Perform PCA over Ionosphere data set and make a scatter plot of PC1 and PC2 (the first two principal components). Are PC1 and PC2 linearly correlated?

```
adata <- read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/ionosph
inputdata <- subset(adata, select = -c(35))
pca2 = prcomp(inputdata)
plot(pca2$rotation, xlab = "PC1", ylab = "PC2")
```

**1.2)** There are three methods to pick the set of principle components: (1) In the plot where the curve bends; (2) Add the percentage variance until total 75% is reached ($70-90\%$) (3) Use the components whose variance is at least one. Show the components selected in the Ionosphere data if each of these is used.

```
pca1 = PCA(inputdata, graph = FALSE)
pca2$rotation[order(pca2$rotation[,1]),1][pca2$rotation[order(pca2$rotation[,1]),1] >=
pca1$eig[,3][pca1$eig[,3] <= 80]#keep variance at least 75
pca2$sdev[pca2$sdev^2 >= 1] #variance >= 1
```

**1.3)** Observe the loadings using prcomp() or princomp() functions in R and discuss loadings in PCA?i.e., how are principal components and original variables related?

```
pca2$rotation
```

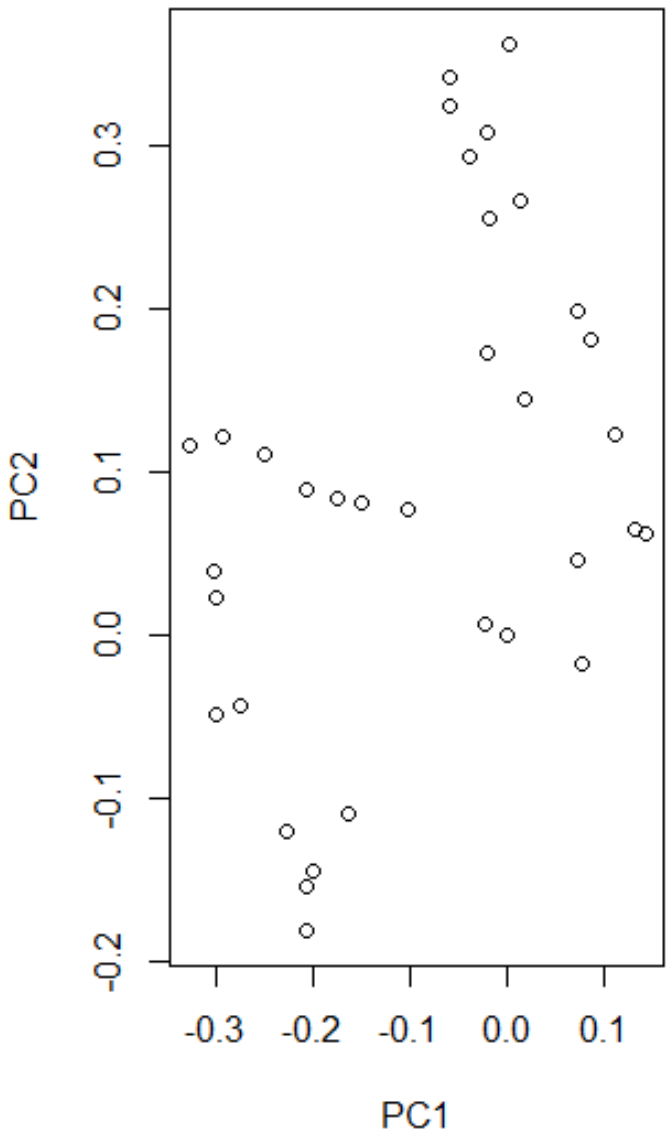



Figure 1: They are not linearly correlated

After observing the correlation between the variables and the factors. Almost all of the data has positive or negative correlation. Showing there is no correlation between each PC.

**1.4)** Perform dimensionality reduction over Ionosphere data set with PCA. Keep 90% of variance after PCA and reduce Ionosphere data set and call this data $\Delta_R$. Cluster $\Delta_R$ using your $k$-means program from previous assignment and report the total error rates for $k = 2, \ldots, 5$ for 20 runs each. Plots are generally a good way to convey complex ideas quickly, i.e., box plots, whisker plots. Discuss your results, i.e how did PCA affect performance of $k$-means clustering.

```
len = length(pca1$eig[,3][pca1$eig[,3] <= 90]) + 1
k1 = kmeans(pca1$eig[c(1:len),], 2, iter.max = 20, algorithm = "Lloyd", trace=FALSE)$tot.w
k2 = kmeans(pca1$eig[c(1:len),], 3, iter.max = 20, algorithm = "Lloyd", trace=FALSE)$tot.w
k3 = kmeans(pca1$eig[c(1:len),], 4, iter.max = 20, algorithm = "Lloyd", trace=FALSE)$tot.w
k4 = kmeans(pca1$eig[c(1:len),], 5, iter.max = 20, algorithm = "Lloyd", trace=FALSE)$tot.w
boxplot(k1,k2,k3,k4)
```

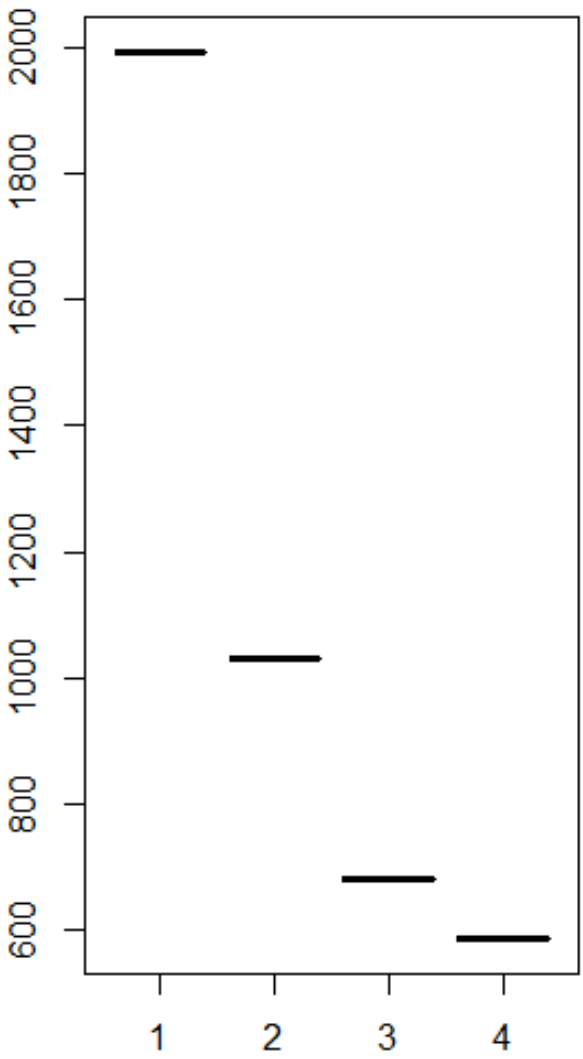



Figure 2: Total error rate decrease exponentially as k increase

# Problem 2 [30 points]

Randomly choose 50 points from Ionosphere data set (call this data set $I_{50}$) and perform hierarchical clustering. You are allowed to use R packages for this question. (Ignore the class variable while performing hierarchical clustering.)

**2.1)** Using hierarchical clustering with complete linkage and Euclidean distance cluster $I_{50}$. Give the dendrogram.

```
randomRows = function(df,n){
  return(df[sample(nrow(df),n),])
}
I50 = randomRows(inputdata, 50)
test = hclust(dist(I50))
plot(test)
```

**2.2)** Cut the dendrogram at a height that results in two distinct clusters. Calculate the error-rate.

```
cut1 = cutree(test, 2)
rate = errorRate(cut1,adata)
[1] 0.4219269
```

**2.3)** First, perform PCA on $I_{50}$ (Keep 90% of variance ). Then hierarchically cluster the reduced data using complete linkage and Euclidean distance. Report the dendrogram.

```
pca3 = PCA(I50, graph = FALSE)
reduced = pca3$eig[,3][pca3$eig[,3] <= 90]
test2 = hclust(dist(reduced))
plot(test2)
```

**2.4)** Cut the dendrogram at a height that results in two distinct clusters. Give the error-rate. Discuss your findings, i.e., how did PCA affect hierarchical clustering results?

```
cut2 = cutree(test3, 2)
rate2 = errorRate(cut2,adata)
[1] 1.1
```

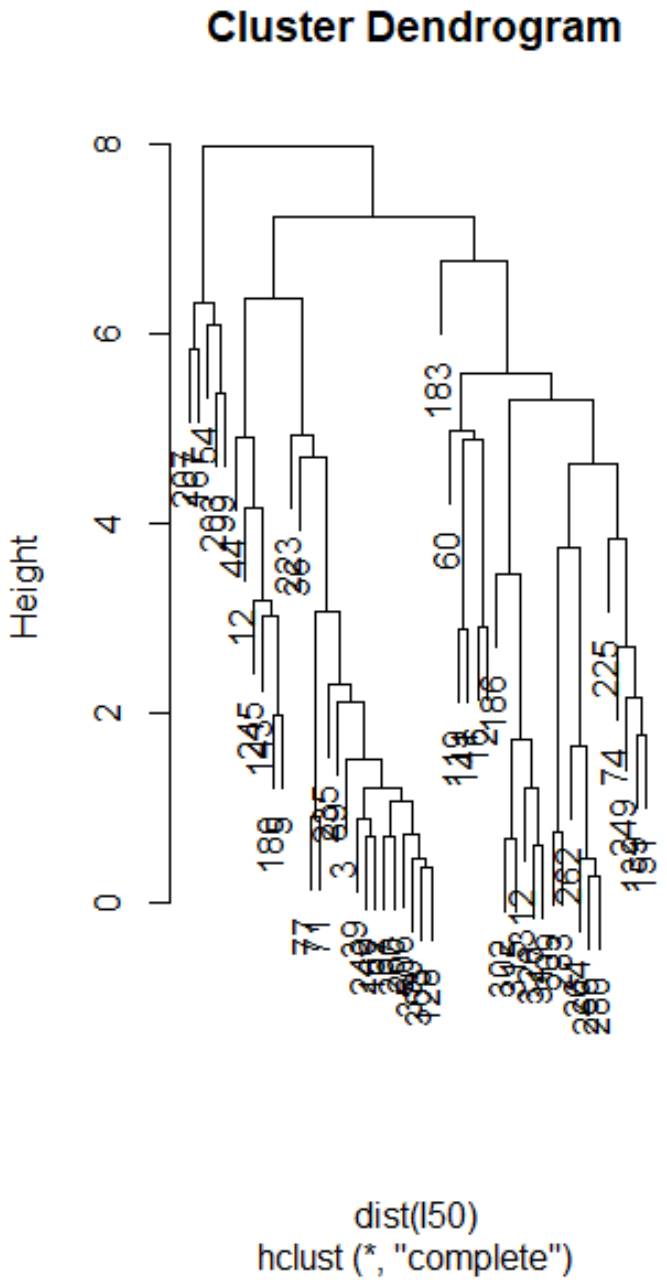
After taking the PCA, the component used are significantly decrease, and making the dendrogram easily view. We also ignored some data by keeping a certain variance.

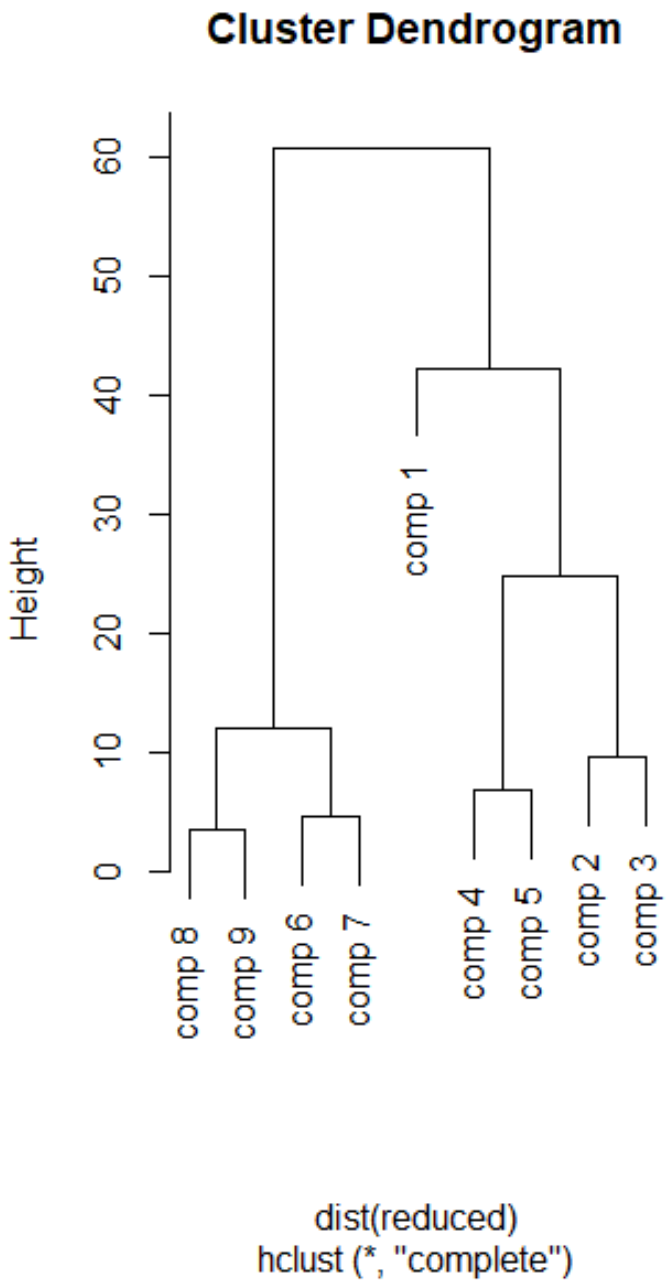
# Problem 3 [20 points]

From textbook, Chapter 8 exercises 16, 18 and 30 (Pages 563-566)

**16** Graphs are shown below. R code is provided in compressed file named "Q16.R"

**18** For both Ward's method and bisecting K-means have no refinement step.
Therefore, Ward's method, and besecting K-means represents a global minimum without removing unwanted elements. Ordinary K-means produces local minimum.


Cluster Dendrogram
Height
0
2
4
6
8
183
60
225
74
dist(l50)
hclust (*, "complete")


Cluster Dendrogram
Height
0
10
20
30
40
50
60
comp 1
comp 8
comp 9
comp 6
comp 7
comp 4
comp 5
comp 2
comp 3
dist(reduced)
hclust (*, "complete")

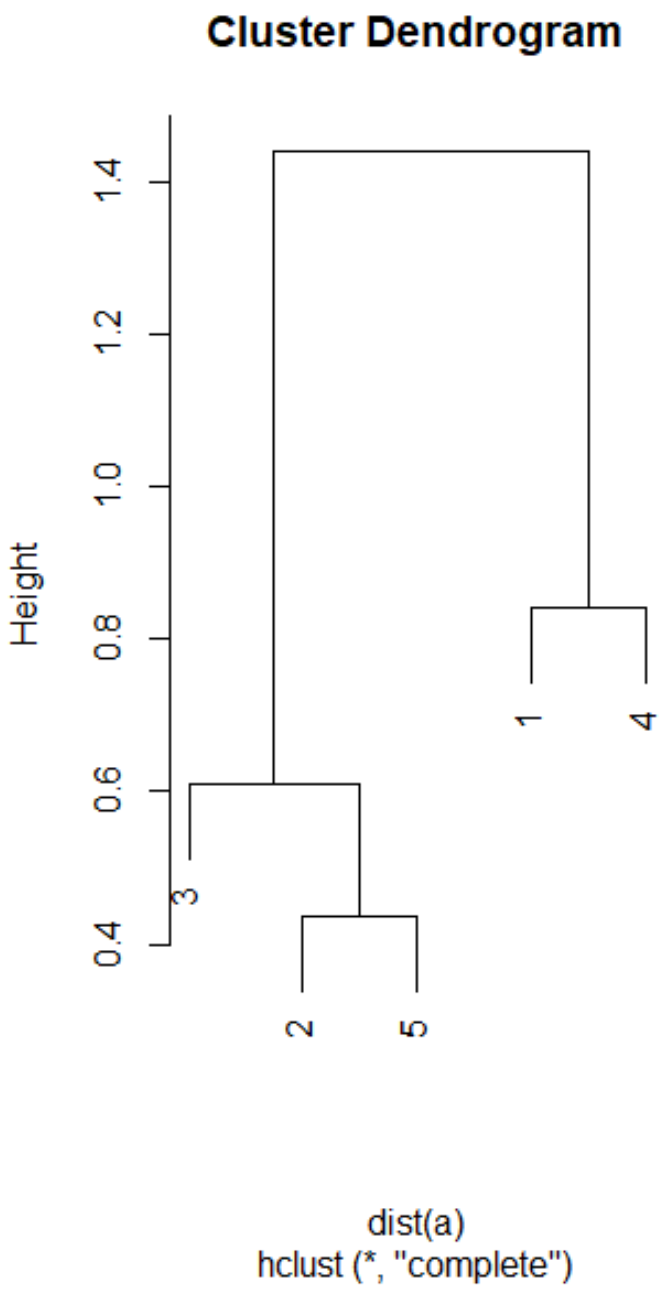



Figure 3: Question16 complete linkage

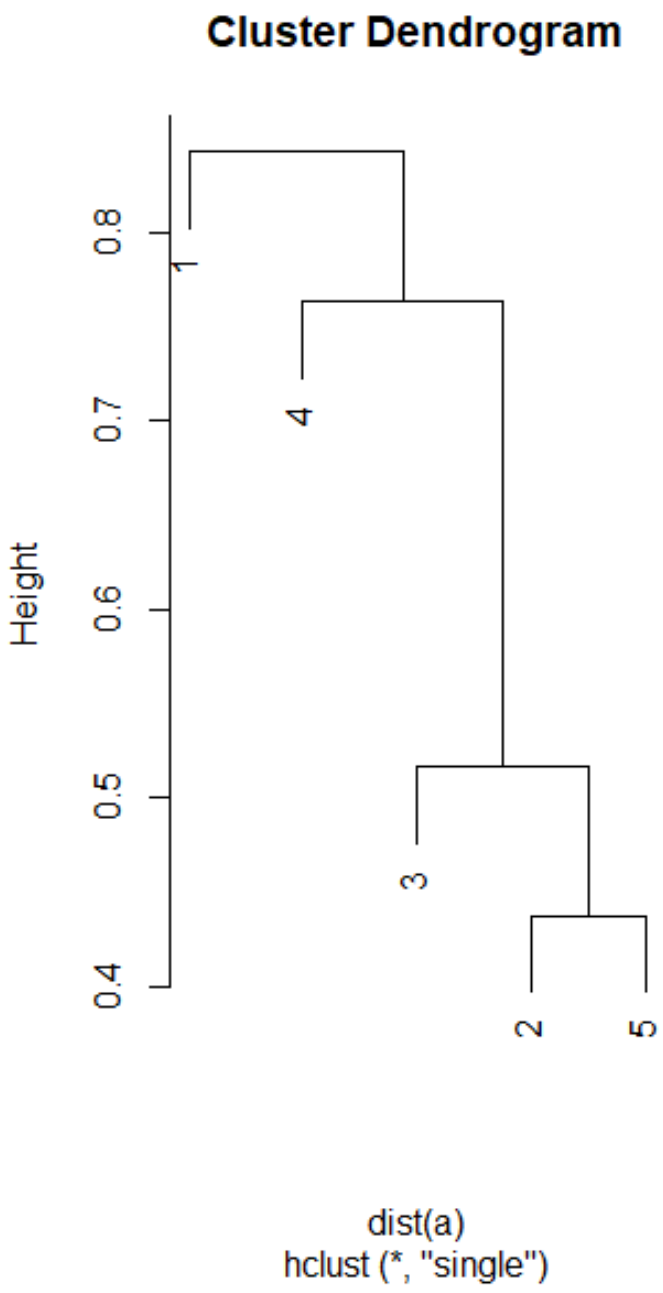



Figure 4: Question16 single linkage

**30a** For the K means clustering of terms, it will cover all the terms, and the result will be not overlapping. However, the top terms clustering is likely to have many terms might not appear in the cluster as the threshold is specified.

**30b** Taking the top documents for the term cluster.

## Extra Credit [10 points]

From textbook, Chapter 8 exercise 12 (Page 562).

**12a** The advantage of the leader algorithm is that it only needs one scan of the data, so this algorithm is more efficient.
The disadvantage is that leader algorithm cannot set the "k" cluster, and this algorithm tends to have worse sum of squared errors than the K-means algorithm.

**12b** Setting the cluster value in a different way that based on the distribution of the ordered data, as we can visualize the percentile on the box plot. Then we can assign data points to its according cluster. (i.e 25th percentile, 50th percentile, outliers etc.)

## What to Turn-in

Submit a .zip file that includes the files below. Name the .zip file as "usename-section number", i.e., hakurban-B365.

- The *tex and *pdf of the written answers to this document.
- *`R`files for:
  - R code for problem 1 ("pca1.R").
  - R code for problem 2 ("hierarchical2.R").