Math 7553 – Spring 2018

HW #4 (Hand In)

Shuddha Chowdhury

Submission Date: 05/01/2018

Chapter 10: Exercise 2

Suppose that we have four observations, for which we compute a

Dissimilarity matrix, given by

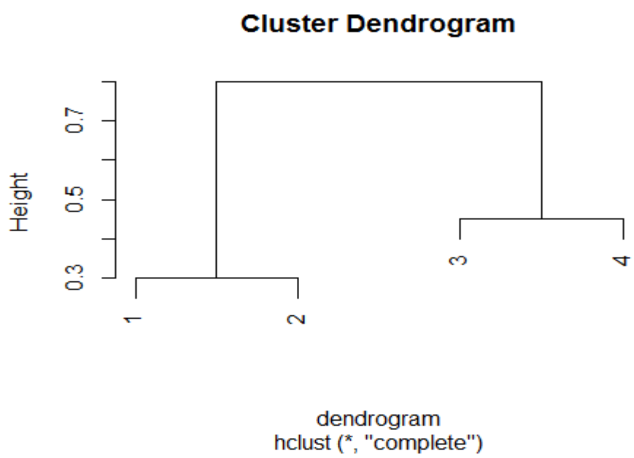
$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

Ans:

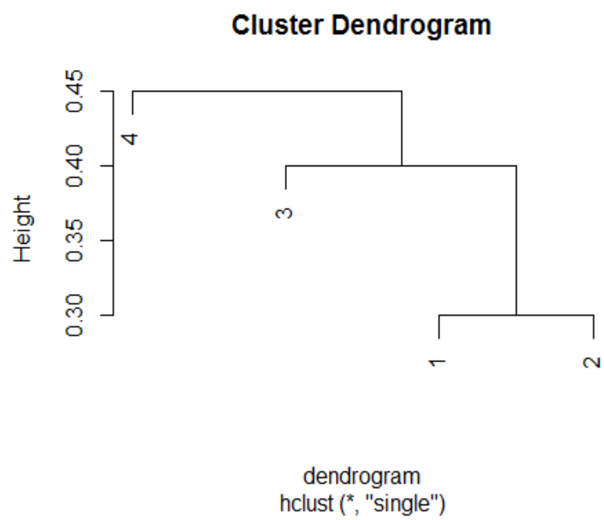
```
dendrogram = as.dist(matrix(c(0, 0.3, 0.4, 0.7,
                              0.3, 0, 0.5, 0.8,
                              0.4, 0.5, 0.0, 0.45,
                              0.7, 0.8, 0.45, 0.0), nrow=4))
plot(hclust(dendrogram, method="complete"))
```



(b) Repeat (a), this time using single linkage clustering.

Ans:

```
plot(hclust(dendrogram, method="single"))
```



(c) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster?

Ans: (1,2), (3,4)

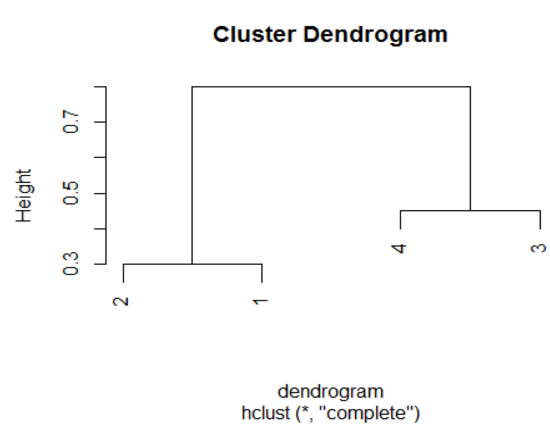
(d) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?

Ans: (1, 2, 3), (4)

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

Ans:

```
plot(hclust(dendrogram, method="complete"), labels=c(2,1,4,3))
```



Reference: Beautiful dendrogram visualizations in R
http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning

Chapter 10: Exercise 3

In this problem, you will perform K-means clustering manually, with K = 2, on a small example with n = 6 observations and p = 2 features. The observations are as follows.

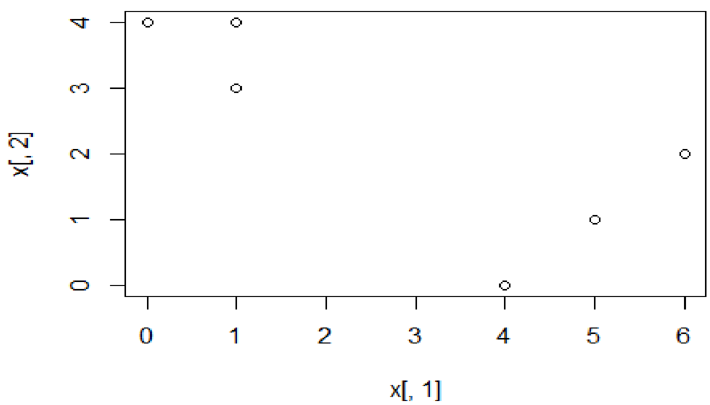
Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

```
set.seed(1)
x = cbind(c(1, 1, 0, 5, 6, 4), c(4, 3, 4, 1, 2, 0))
x

##      [,1] [,2]
## [1,]    1    4
## [2,]    1    3
## [3,]    0    4
## [4,]    5    1
## [5,]    6    2
## [6,]    4    0
```

(a) Plot the observations.

Ans: **plot(x[,1], x[,2])**



(b) Randomly assign a cluster label to each observation. You can use the sample () command in R to do this. Report the cluster labels for each observation.

Ans:

```
labels = sample(2, nrow(x), replace=T)
labels

## [1] 1 1 2 2 1 2
```

(c) Compute the centroid for each cluster.

Ans:

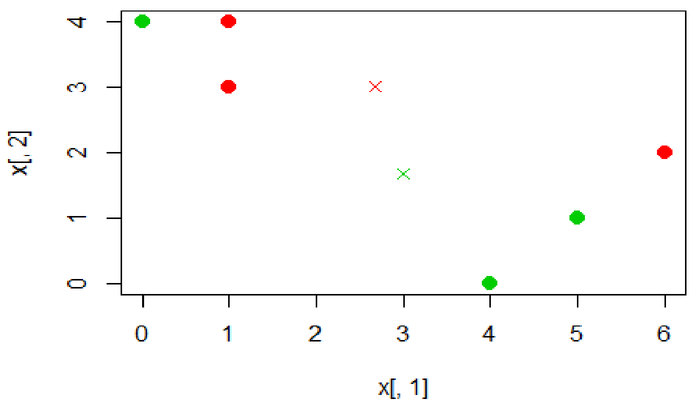
```
centroid_1st = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))
centroid_2nd = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))
centroid_1st

## [1] 2.666667 3.000000

centroid_2nd

## [1] 3.000000 1.666667

plot(x[,1], x[,2], col=(labels+1), pch=20, cex=2)
points(centroid_1st[1], centroid_1st[2], col=2, pch=4)
points(centroid_2nd[1], centroid_2nd[2], col=3, pch=4)
```



(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

Ans:

```
euclid = function(a, b) {
  return(sqrt((a[1] - b[1])^2 + (a[2]-b[2])^2))
}
assign_labels = function(x, centroid_1st, centroid_2nd) {
  labels = rep(NA, nrow(x))
  for (i in 1:nrow(x)) {
    if (euclid(x[i,], centroid_1st) < euclid(x[i,], centroid_2nd)) {
      labels[i] = 1
    } else {
      labels[i] = 2
    }
  }
  return(labels)
}
labels = assign_labels(x, centroid_1st, centroid_2nd)
labels

## [1] 1 1 1 2 2 2
```

(e) Repeat (c) and (d) until the answers obtained stop changing.

Ans:

```
last_labels = rep(-1, 6)
while (!all(last_labels == labels)) {
  last_labels = labels
  centroid_1st = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))
  centroid_2nd = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))
  print(centroid_1st)
  print(centroid_2nd)
  labels = assign_labels(x, centroid_1st, centroid_2nd)
}

## [1] 0.6666667 3.6666667
## [1] 5 1
```

```
labels
```

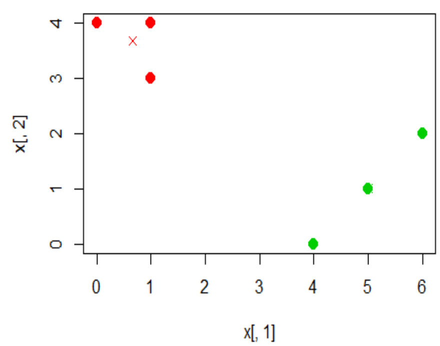
```
## [1] 1 1 1 2 2 2
```

(f) In your plot from (a), color the observations according to the

Cluster labels obtained.

Ans:

```
plot(x[,1], x[,2], col=(labels+1), pch=20, cex=2)
points(centroid_1st[1], centroid_1st[2], col=2, pch=4)
points(centroid_2nd[1], centroid_2nd[2], col=3, pch=4)
```



Chapter 10: Exercise 7

In the chapter, we mentioned the use of correlation-based distance and Euclidean distance as dissimilarity measures for hierarchical clustering. It turns out that these two measures are almost equivalent: if each observation has been centered to have mean zero and standard deviation one, and if we let r_{ij} denote the correlation between the ith and jth observations, then the quantity 1-rij is proportional to the squared Euclidean distance between the ith and jth observations. On the USArrests data, show that this proportionality holds. Hint: The Euclidean distance can be calculated using the dist() function, and correlations can be calculated using the cor() function.

Ans:

```
library(ISLR)
set.seed(1)

distance = scale(USArrests)
a = dist(distance)^2
b = as.dist(1 - cor(t(distance)))
summary(b/a)
```

```
##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000086 0.069135 0.133943 0.234193 0.262589 4.887686
```

Chapter 10: Exercise 9

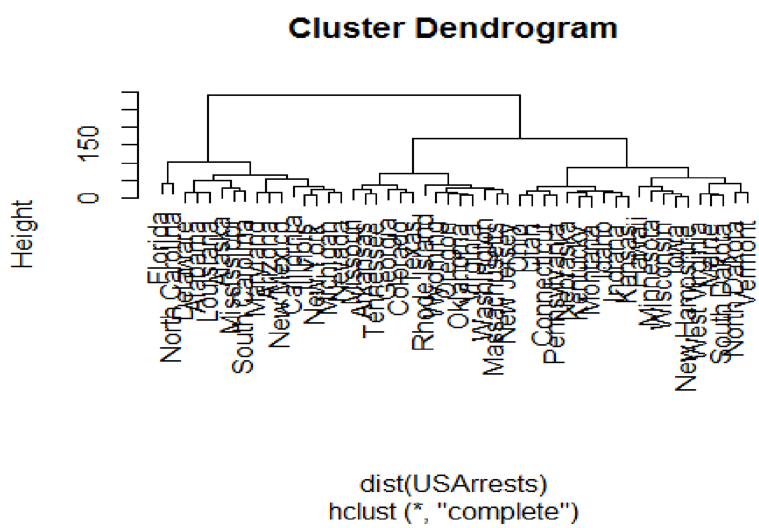
Consider the USArrests data. We will now perform hierarchical clustering on the states.

```r
library(ISLR)
set.seed(2)
```

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

Ans:

```r
hierarchical_clustering.complete = hclust(dist(USArrests),
method="complete")
plot(hierarchical_clustering.complete)
```



(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

Ans:

```r
cutree(hierarchical_clustering.complete, 3)

##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
```

```
##             2             2             3             2             1
##  South Dakota     Tennessee         Texas          Utah       Vermont
##             3             2             2             3             3
##      Virginia    Washington West Virginia     Wisconsin       Wyoming
##             2             2             3             3             2
```

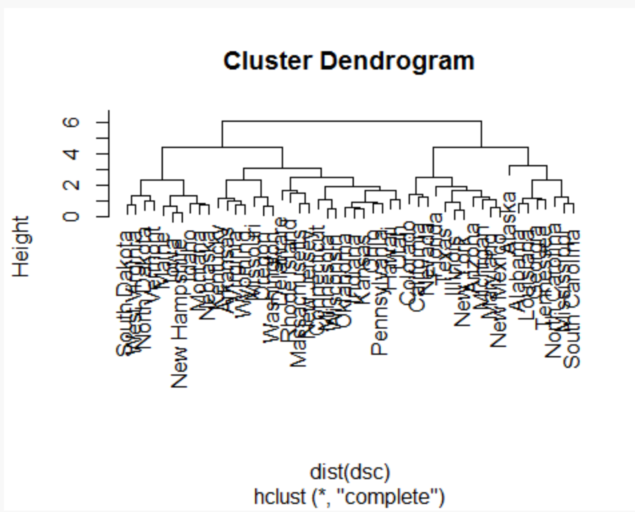
```r
table(cutree(hierarchical_clustering.complete, 3))
```

```
##
##  1  2  3
## 16 14 20
```

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

Ans:

```r
dsc = scale(USArrests)
hierarchical_clustering.s.complete = hclust(dist(dsc), method="complete")
plot(hierarchical_clustering.s.complete)
```



(d) What effect does scaling the variables have on the hierarchical clusterin g obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your an swer.

Ans:

```r
cutree(hierarchical_clustering.s.complete, 3)
```

```
##         Alabama          Alaska         Arizona        Arkansas      California 
##               1               1               2               3               2 
##        Colorado     Connecticut        Delaware         Florida         Georgia 
##               2               3               3               2               1 
##          Hawaii           Idaho        Illinois         Indiana            Iowa 
##               3               3               2               3               3 
##          Kansas        Kentucky       Louisiana           Maine        Maryland 
##               3               3               1               3               2 
##   Massachusetts        Michigan       Minnesota     Mississippi        Missouri 
##               3               2               3               1               3 
##         Montana        Nebraska          Nevada   New Hampshire      New Jersey 
##               3               3               2               3               3 
##      New Mexico        New York  North Carolina    North Dakota            Ohio 
##               2               2               1               3               3 
##        Oklahoma          Oregon    Pennsylvania    Rhode Island  South Carolina 
##               3               3               3               3               1 
##    South Dakota       Tennessee           Texas            Utah         Vermont 
##               3               1               2               3               3 
##        Virginia      Washington   West Virginia       Wisconsin         Wyoming 
##               3               3               3               3               3
```

```
table(cutree(hierarchical_clustering.s.complete, 3))
```

```
## 
##  1  2  3 
##  8 11 31
```

```
table(cutree(hierarchical_clustering.s.complete, 3), cutree(hierarchical_clus
tering.complete, 3))
```

```
##    
##      1  2  3
##   1  6  2  0
##   2  9  2  0
##   3  1 10 20
```

If we scale the variables then it will have an effect on the maximum height of the dendogram which we obtain from the hierarchical clustering. Though it doesn’t have an effect on the bushiness of the tree but it has an effect on the cluster which we obtain after cutting the dendogram into 3 clusters. I think the data set needs to be standardized because the data which we measured has different units. (UrbanPop compared to the three other columns)

Question 5:

Select a medium to high dimensional data set that is available online. This can be from any source, ideally it would come from a domain of application that has some interest to you. It could come from a repository such as the UC Irvine Machine Learning Repository (see URL next page), a data set from some text book on machine learning or statistical learning, an article from a journal article, or even your own research project. The goal is to formulate research questions regarding this data and propose a statistical learning methodology, covered during this course, that would/might address the question(s) posed. This is a proposal NOT an analysis itself.

Ans:

a) **Describe the location, content, and context of the data set (for example, the "glass" dataset form the UCI repository on the next page).**

Welcome to the UC Irvine Machine Learning Repository!

https://archive.ics.uci.edu/ml/datasets

https://archive.ics.uci.edu/ml/datasets/Adult

This data set is also known as "Census Income" dataset.

Source:

Donor:

Ronny Kohavi and Barry Becker

Data Mining and Visualization

Silicon Graphics.

Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:

>50K, <=50K.

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm,

Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

b) Pose one or more research questions that might be addressed by analyzing this data set using one or more of the methods discussed in this course. What is the scientific merit of the proposed project?

Ans:

Research Question:

a) Which of the variables (age, occupation, sex, etc.) are most decisive for determining the income of a person?

b) Can we build a machine learning model which can predict if a person will make $50000 per year given data like education, gender and marital status?

To answer question a) we can build a decision tree classifier with the training data set and then build a good classification model to answer the question.

To answer question b) we can build a machine learning model (a logistic regression) which can predict if a person will make more than $50000 per year.

The scientific merit of the research project is social scientists or state high officials can find out meaningful research question answer based on this research and then it can be applied in important decision making such as education budget allocation, medical insurance policy decision, elderly benefits etc.

c) Describe in some detail the process of down loading the data, preprocessing the data – if required, the analysis steps, and the general form of expected results of the analysis. DO NOT carry out the analysis.

Ans:

Process Detail:

The process for the analysis is at first we need to load the dataset and read the text data as csv format for our analysis. Then using various plot like histogram we can plot the distribution of each feature so that we can have a better understanding of our data. We can view age, workclass, education, marital status, Occupation, Race, and Relationship in the plot.

Then we can build a classifier which tries to predict what will the income of given person given the features in our dataset such as education, age marital status. Then we can apply logistic regression technique to answer the above question. Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both. The categorical variable y, in general, can assume different values. In the simplest case scenario y is binary meaning that it can assume either the value 1 or 0. The general form of expected results of the analysis are some variables which impact the income of a person positively in the analysis might be.

Capital Gain, Age, Hours per week and some of the variables may impact the result negatively are Never Married, Gender etc. The above process will help us to predict whether a person will make $50000 or not.