Homework 5

Joshua Oswari - A14751270

5/10/2019

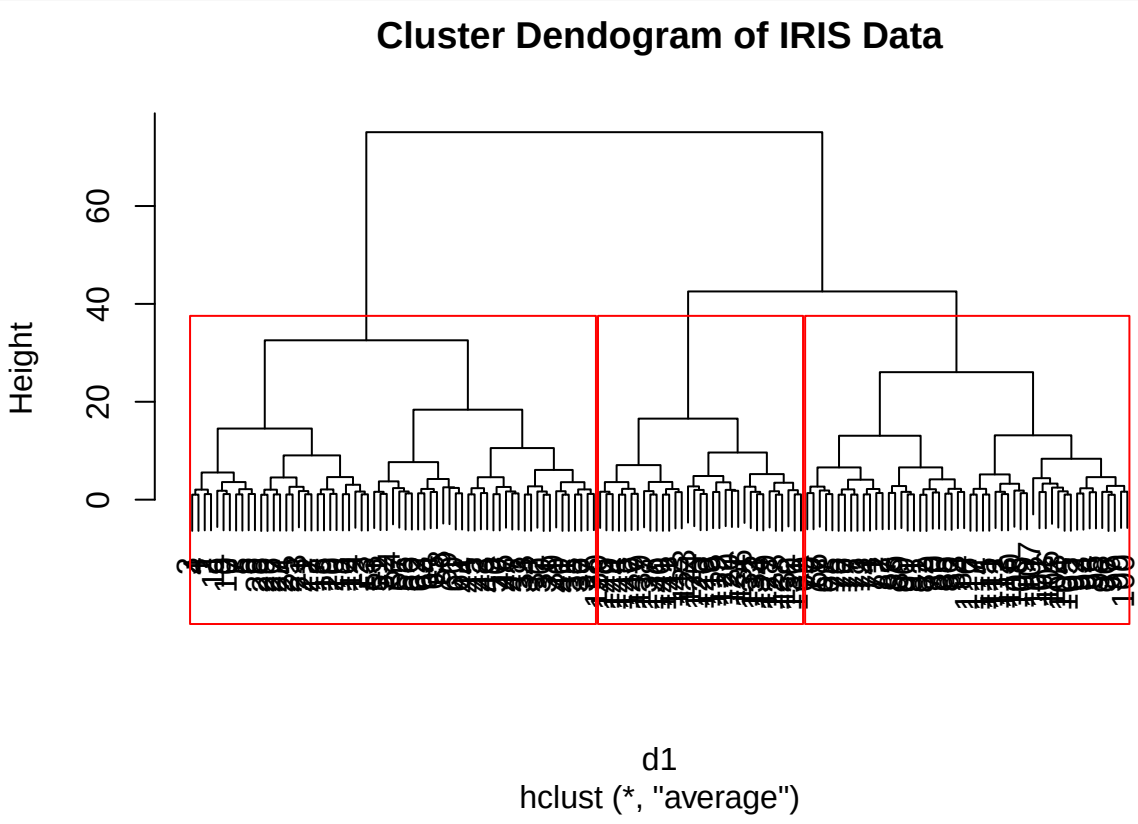
Problem 1

========== Part 1 ==========

```
library(stats)
data = read.csv(file = "~/Documents/Math189/iris.csv", header = TRUE, fill = TRUE)
iris = data[,2:6]

#subset of data that is purely integer
pureIris = data[,1:4]

d1 = dist(pureIris, method = "euclidean")
hc1 = hclust(d1, method = "average")
clusterCut1 = cutree(hc1, k = 3)
plot(hc1, main = "Cluster Dendogram of IRIS Data")
rect.hclust(hc1, k=3, border="red")
```

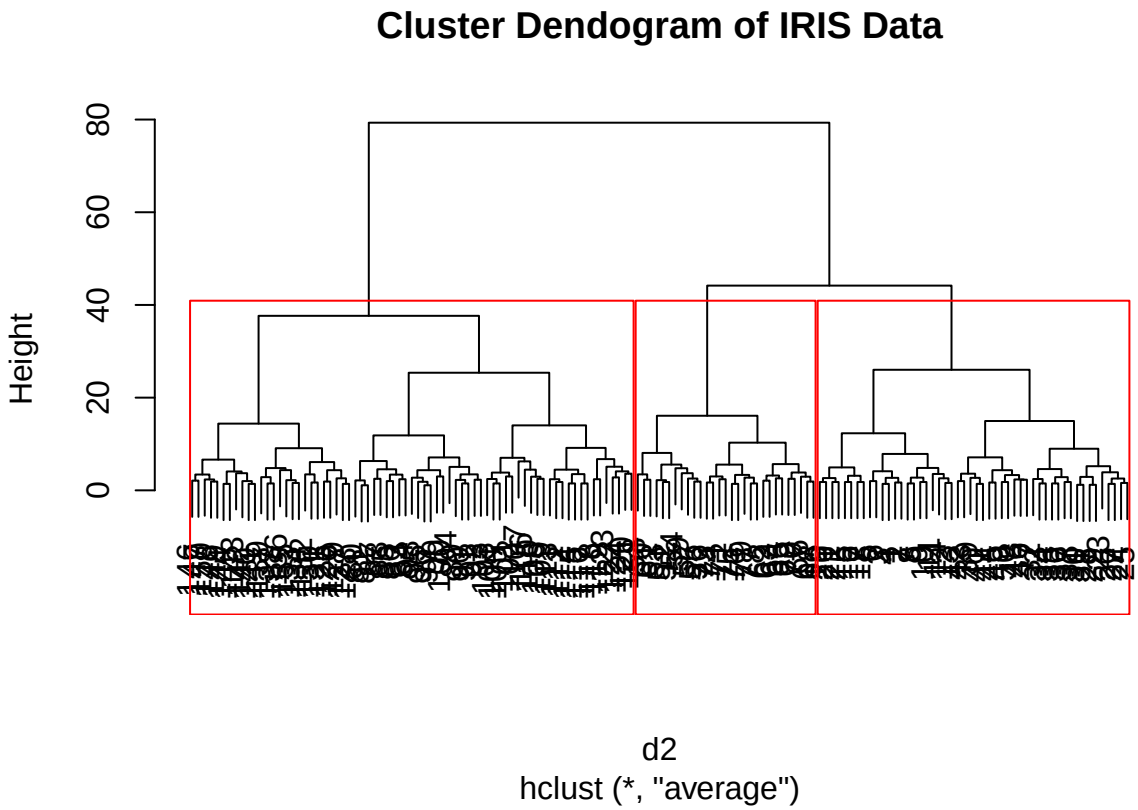



```
#Correct clustering rate is
table1 = table(clusterCut1,iris$Species)
paste("Correct rate is ", sum(diag(table1))/150*100 , "%")
```

```
## [1] "Correct rate is  78.6666666666667 %"
```

========== Part 2 ==========

```
#######------ Part (1) -------######
#(1) Retaining all the settings in part 1, except using the Taxicab
#distance as the similarity measure.
d2 = dist(pureIris, method = "manhattan")
hc2 = hclust(d2, method = "average")
clusterCut2 = cutree(hc2, k = 3)
plot(hc2, main = "Cluster Dendogram of IRIS Data")
rect.hclust(hc2, k=3, border="red")
```




```
table(clusterCut2,iris$Species)
```

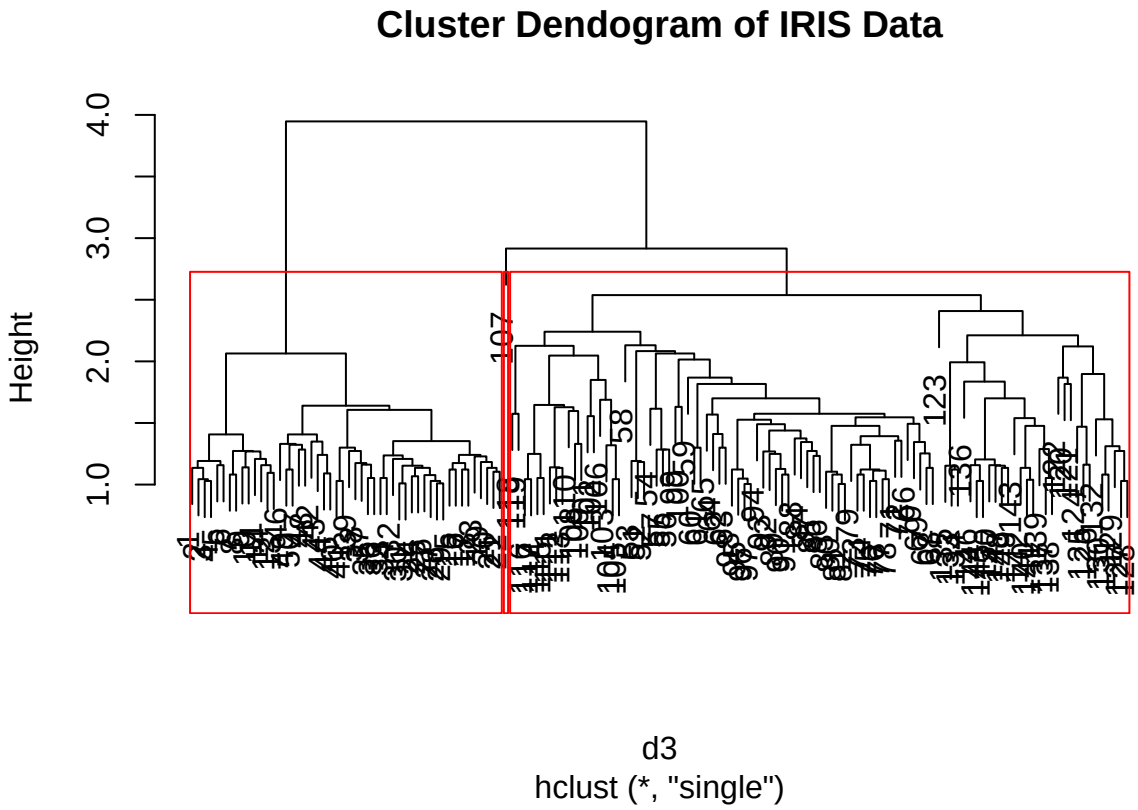
```
##
## clusterCut2 setosa versicolor virginica
##           1     50          0         0
##           2      0         29         0
##           3      0         21        50
```

```
#Correct clustering rate is
table2 = table(clusterCut2,iris$Species)
paste("Correct rate is ", sum(diag(table2))/150*100 , "%")
```

```
## [1] "Correct rate is  86 %"
```

```
#######------ Part (2) -------######
#(2) Retaining all the settings in part 1, except using the single
#linkage as the linkage function.
d3 = dist(pureIris, method = "euclidean")
hc3 = hclust(d3, method = "single")
```

```
clusterCut3 = cutree(hc3, k = 3)
plot(hc3, main = "Cluster Dendogram of IRIS Data")
rect.hclust(hc3, k=3, border="red")
```




```
table(clusterCut3,iris$Species)
```

```
##
## clusterCut3 setosa versicolor virginica
##           1     50          0         0
##           2      0         50        49
##           3      0          0         1
```

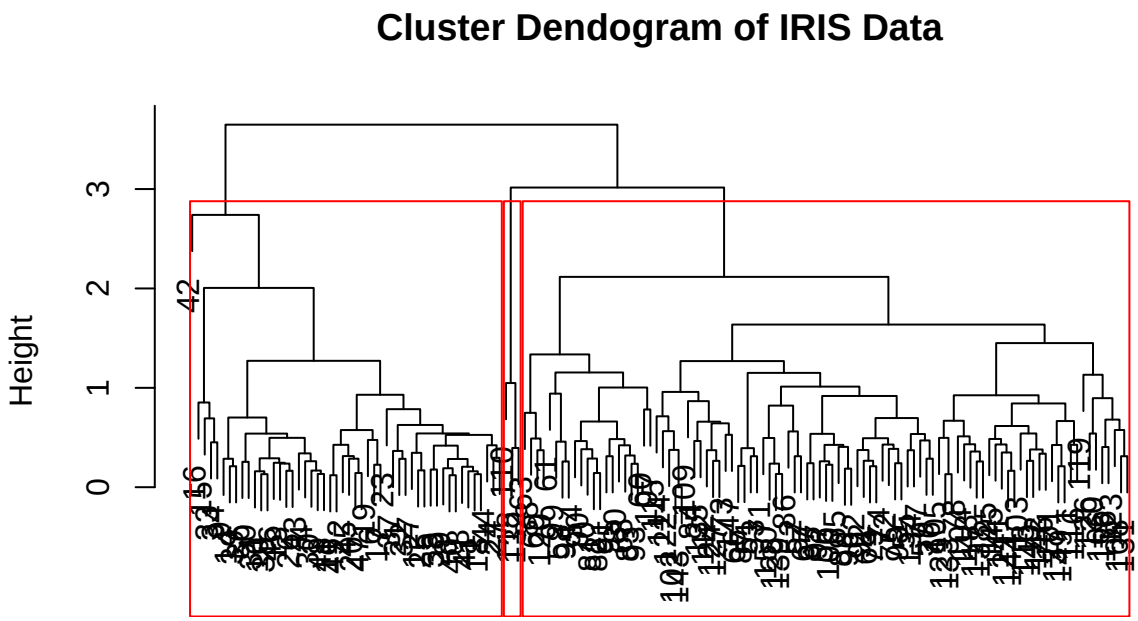
```
#Correct clustering rate is
table3 = table(clusterCut3,iris$Species)
paste("Correct rate is ", sum(diag(table3))/150*100 , "%")
```

```
## [1] "Correct rate is  67.3333333333333 %"
```

```
#######------ Part (3) -------######
#(3) Retaining all the settings in part 1, except standardizing
#each variable before applying the clustering. (Standardize each
#column in the dataset to have zero mean and unit variance).

#standardize variables
standardizIris = scale(iris[,1:4])
d4 = dist(standardizIris, method = "euclidean")
hc4 = hclust(d4, method = "average")
clusterCut4 = cutree(hc4, k = 3)
plot(hc4, main = "Cluster Dendogram of IRIS Data")
```

```
rect.hclust(hc4, k=3, border="red")
```



d4
hclust (*, "average")

```
table(clusterCut4,iris$Species)
```

```
##
## clusterCut4 setosa versicolor virginica
##           1     50          0         0
##           2      0         50        47
##           3      0          0         3
```

```
#Correct clustering rate is
table4 = table(clusterCut4,iris$Species)
paste("Correct rate is ", sum(diag(table4))/150*100 , "%")
```

```
## [1] "Correct rate is  68.6666666666667 %"
```