Q43

John Bute

2024-11-29

R Markdown

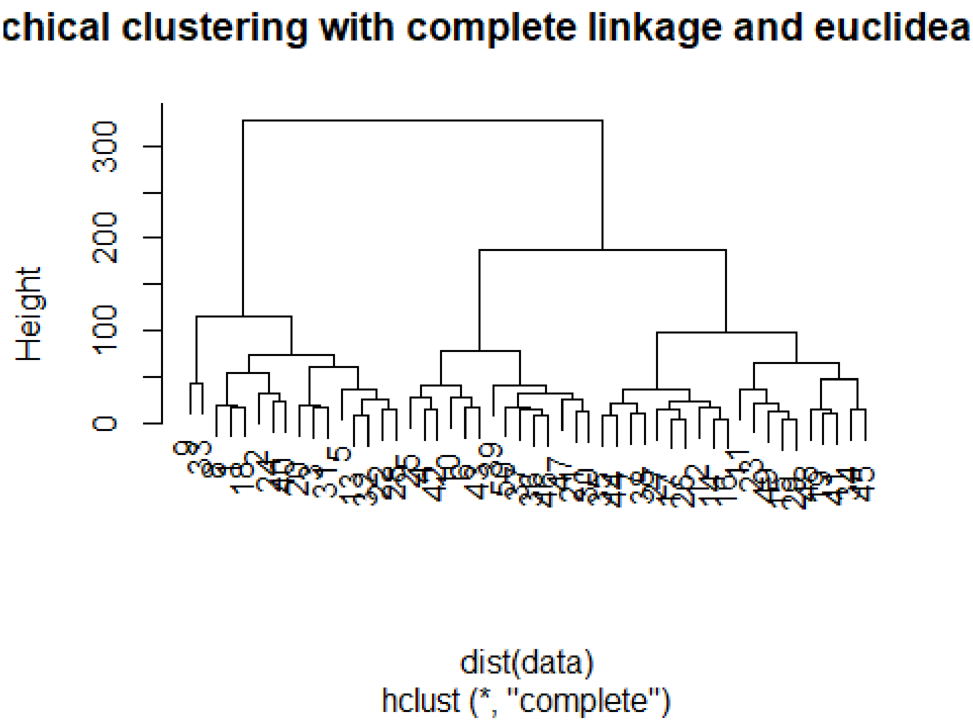
Consider the USArrests data. We perform hierarchical clustering on the states. 1. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states (don't scale the data at first).

```
data <- read.csv("C:/Users/johnb/Desktop/Machine
Learning/data/USArrests.csv")

hc_clustering <- hclust(dist(data), method = "complete")

## Warning in dist(data): NAs introduced by coercion

plot(hc_clustering, main = "Hierarchical clustering with complete linkage and
euclidean distance")
```



For the next question, it seems that a value of slightly higher than 100 cuts the cluster into three distinct ones 2. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```r
clusters <- cutree(hc_clustering, k = 3)
cluster_membership <- data.frame(State = rownames(data), Cluster = clusters)
print(cluster_membership)
```

```
##    State Cluster
## 1      1       1
## 2      2       1
## 3      3       1
## 4      4       2
## 5      5       1
## 6      6       2
## 7      7       3
## 8      8       1
## 9      9       1
## 10    10       2
## 11    11       3
## 12    12       3
## 13    13       1
## 14    14       3
## 15    15       3
## 16    16       3
## 17    17       3
## 18    18       1
## 19    19       3
## 20    20       1
## 21    21       2
## 22    22       1
## 23    23       3
## 24    24       1
## 25    25       2
## 26    26       3
## 27    27       3
## 28    28       1
## 29    29       3
## 30    30       2
## 31    31       1
## 32    32       1
## 33    33       1
## 34    34       3
## 35    35       3
## 36    36       2
## 37    37       2
## 38    38       3
## 39    39       2
## 40    40       1
## 41    41       3
## 42    42       2
## 43    43       2
## 44    44       3
## 45    45       3
## 46    46       2
## 47    47       2
## 48    48       3
## 49    49       3
## 50    50       2
```

```r
split(cluster_membership$State, cluster_membership$Cluster)
```

```
## $`1`
##  [1] "1"  "2"  "3"  "5"  "8"  "9"  "13" "18" "20" "22" "24" "28" "31" "32"
"33"
## [16] "40"
##
## $`2`
##  [1] "4"  "6"  "10" "21" "25" "30" "36" "37" "39" "42" "43" "46" "47" "50"
##
## $`3`
##  [1] "7"  "11" "12" "14" "15" "16" "17" "19" "23" "26" "27" "29" "34" "35"
"38"
## [16] "41" "44" "45" "48" "49"
```

```r
cluster_counts <- as.data.frame(table(clusters))

colnames(cluster_counts) <- c("Cluster", "Number of States")

print(cluster_counts)
```

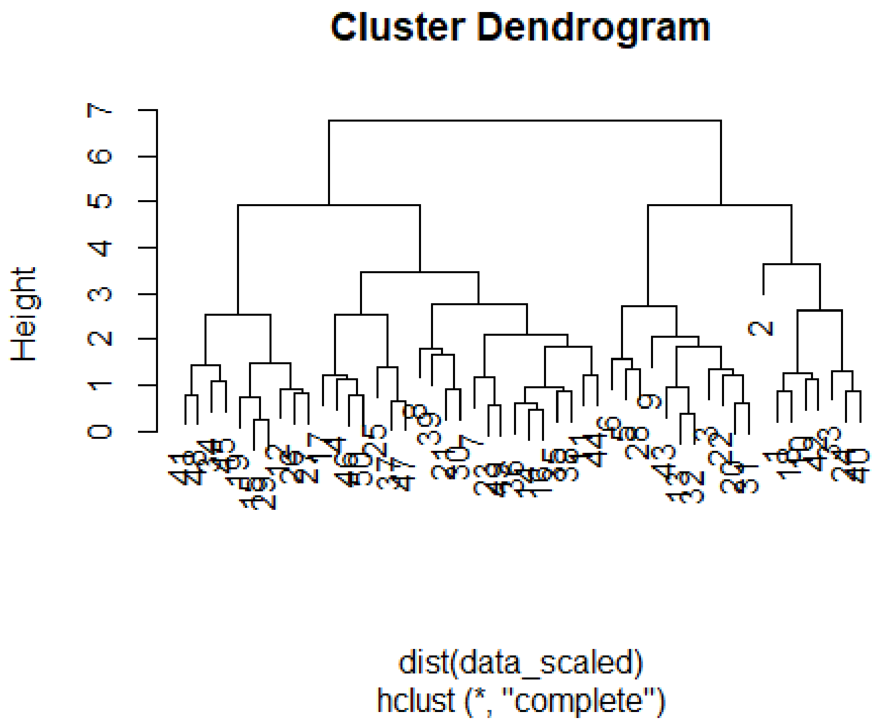
```
##   Cluster Number of States
## 1       1               16
## 2       2               14
## 3       3               20
```

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

```r
data_scaled <- data
data_scaled[, c(2:5)] <- scale(data[, c(2:5)])
hierarical_scaled_clustering <- hclust(dist(data_scaled), method="complete")
```

```
## Warning in dist(data_scaled): NAs introduced by coercion
```

```r
plot(hierarical_scaled_clustering)
```



```
clusters <- cutree(hierarical_scaled_clustering, k = 3)
cluster_membership <- data.frame(State = rownames(data), Cluster = clusters)
print(cluster_membership)
```

```
##    State Cluster
## 1      1       1
## 2      2       1
## 3      3       2
## 4      4       3
## 5      5       2
## 6      6       2
## 7      7       3
## 8      8       3
## 9      9       2
## 10    10       1
## 11    11       3
## 12    12       3
## 13    13       2
## 14    14       3
## 15    15       3
## 16    16       3
## 17    17       3
## 18    18       1
## 19    19       3
## 20    20       2
## 21    21       3
## 22    22       2
## 23    23       3
## 24    24       1
## 25    25       3
## 26    26       3
## 27    27       3
## 28    28       2
## 29    29       3
## 30    30       3
## 31    31       2
## 32    32       2
## 33    33       1
## 34    34       3
## 35    35       3
## 36    36       3
## 37    37       3
## 38    38       3
## 39    39       3
## 40    40       1
## 41    41       3
## 42    42       1
## 43    43       2
## 44    44       3
## 45    45       3
## 46    46       3
## 47    47       3
## 48    48       3
## 49    49       3
## 50    50       3
```

```
print(split(cluster_membership$State, cluster_membership$Cluster))
```

```
## $`1`
## [1] "1"  "2"  "10" "18" "24" "33" "40" "42"
## 
## $`2`
##  [1] "3"  "5"  "6"  "9"  "13" "20" "22" "28" "31" "32" "43"
## 
## $`3`
##  [1] "4"  "7"  "8"  "11" "12" "14" "15" "16" "17" "19" "21" "23" "25" "26"
"27"
## [16] "29" "30" "34" "35" "36" "37" "38" "39" "41" "44" "45" "46" "47" "48"
"49"
## [31] "50"
```

```
cluster_counts <- as.data.frame(table(clusters))

colnames(cluster_counts) <- c("Cluster", "Number of States")

print(cluster_counts)
```

```
##   Cluster Number of States
## 1       1                8
## 2       2               11
## 3       3               31
```

4. What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer

Firstly, the height of the dendogram is smaller when scaled than when it is not. Furthermore, the clusters, when cut into 3, are affected. Furthermore, I think that the variables should be scaled because the four variables in the dataset have different scales.