p8106_hw6_yg2625

Yue Gu

5/7/2019

Cluster analysis

We perform hierarchical clustering on the states using the USArrests data in the ISLR package. For each of the 50 states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. The data set also contains the percent of the population in each state living in urban areas, UrbanPop. The four variables will be used as features for clustering.

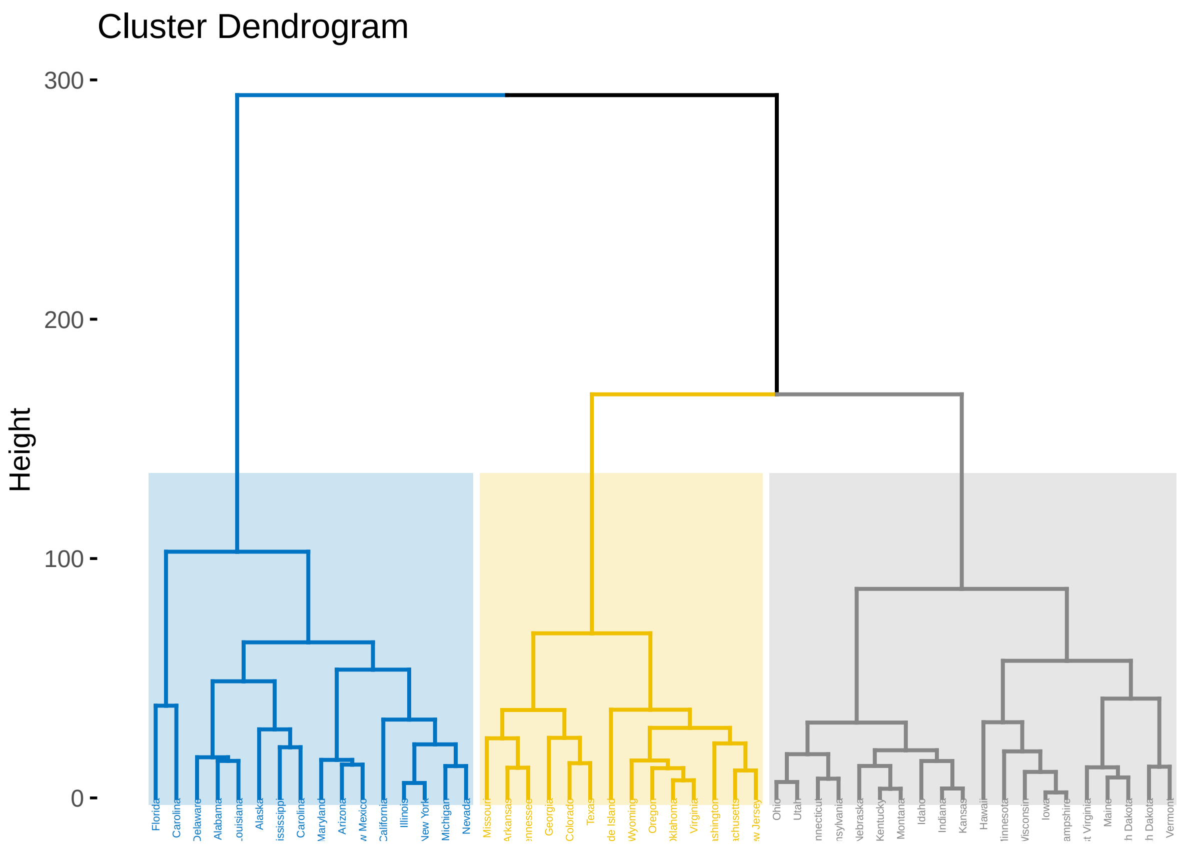
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

```r
# load data
data(USArrests)
arr_data = USArrests %>%
  janitor::clean_names()

# fit hierarchical cluster
hc.complete <- hclust(dist(arr_data), method = "complete")
```

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

```r
# visualize dendrogram in 3 distinct clusters
fviz_dend(hc.complete, k = 3,
          cex = 0.3,
          palette = "jco",
          color_labels_by_k = TRUE,
          rect = TRUE, rect_fill = TRUE, rect_border = "jco",
          labels_track_height = 2.5)
```



```
# show the states in clusters
ind3.complete <- cutree(hc.complete, 3)
# name of states in cluster 1
arr_data[ind3.complete == 1,] %>% row.names()
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "California"
##  [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
##  [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
## [13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
```

```
# name of states in cluster 2
arr_data[ind3.complete == 2,] %>% row.names()
```

```
##  [1] "Arkansas"      "Colorado"      "Georgia"       "Massachusetts"
##  [5] "Missouri"      "New Jersey"    "Oklahoma"      "Oregon"
##  [9] "Rhode Island"  "Tennessee"     "Texas"         "Virginia"
## [13] "Washington"    "Wyoming"
```

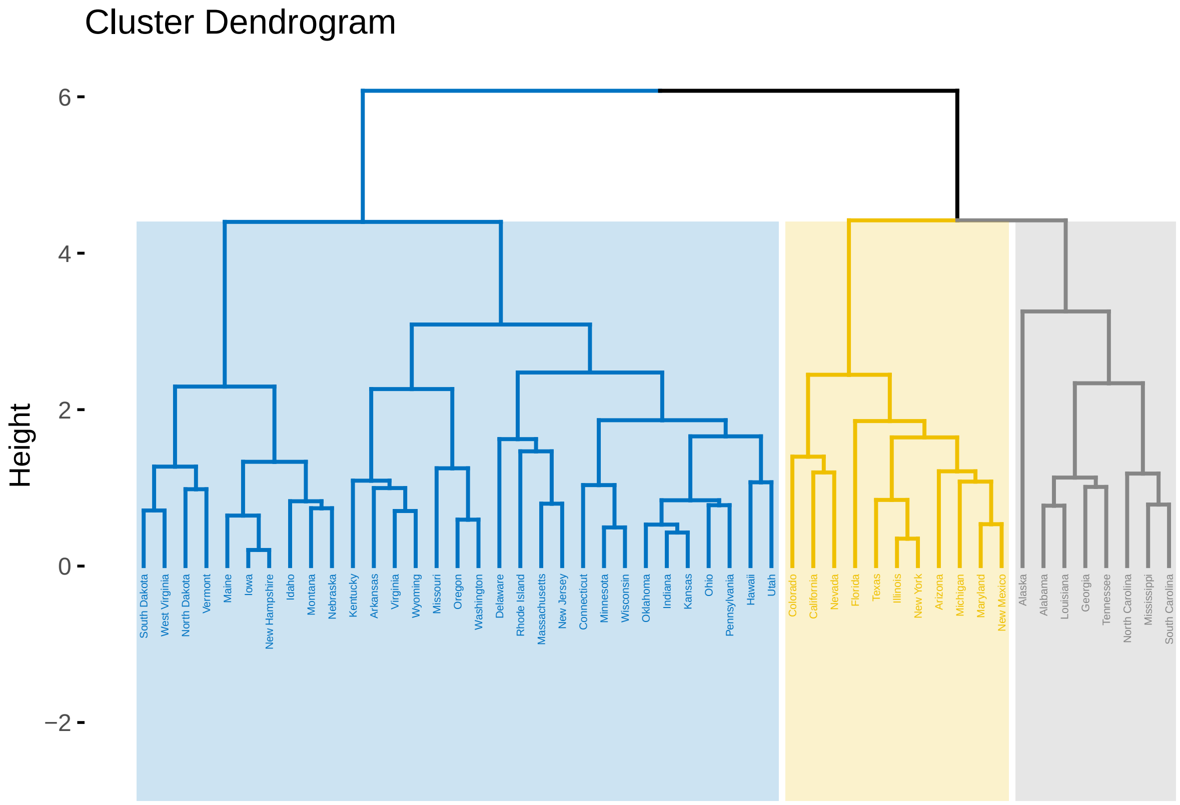
```
# name of states in cluster 3
arr_data[ind3.complete == 3,] %>% row.names()
```

```
##  [1] "Connecticut"   "Hawaii"        "Idaho"         "Indiana"
##  [5] "Iowa"          "Kansas"        "Kentucky"      "Maine"
##  [9] "Minnesota"     "Montana"       "Nebraska"      "New Hampshire"
## [13] "North Dakota"  "Ohio"          "Pennsylvania"  "South Dakota"
## [17] "Utah"          "Vermont"       "West Virginia" "Wisconsin"
```

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

```
# fit hierarchical cluster with standardized variables
hc.complete_std <- hclust(dist(scale(arr_data)), method = "complete")

# visualize dendrogram
fviz_dend(hc.complete_std, k = 3,
          cex = 0.3,
          palette = "jco",
          color_labels_by_k = TRUE,
          rect = TRUE, rect_fill = TRUE, rect_border = "jco",
          labels_track_height = 2.5)
```



```
# show the states in clusters
ind3.complete_std <- cutree(hc.complete_std, 3)
# name of states in cluster 1
arr_data[ind3.complete_std == 1,] %>% row.names()
```

```
## [1] "Alabama"        "Alaska"         "Georgia"        "Louisiana"
## [5] "Mississippi"    "North Carolina" "South Carolina" "Tennessee"
```

```
# name of states in cluster 2
arr_data[ind3.complete_std == 2,] %>% row.names()
```

```
##  [1] "Arizona"    "California" "Colorado"   "Florida"    "Illinois"
##  [6] "Maryland"   "Michigan"   "Nevada"     "New Mexico" "New York"
## [11] "Texas"
```

```
# name of states in cluster 3
arr_data[ind3.complete_std == 3,] %>% row.names()
```

```
##  [1] "Arkansas"      "Connecticut"   "Delaware"      "Hawaii"       
##  [5] "Idaho"         "Indiana"       "Iowa"          "Kansas"       
##  [9] "Kentucky"      "Maine"         "Massachusetts" "Minnesota"    
## [13] "Missouri"      "Montana"       "Nebraska"      "New Hampshire"
## [17] "New Jersey"    "North Dakota"  "Ohio"          "Oklahoma"     
## [21] "Oregon"        "Pennsylvania"  "Rhode Island"  "South Dakota" 
## [25] "Utah"          "Vermont"       "Virginia"      "Washington"   
## [29] "West Virginia" "Wisconsin"     "Wyoming"
```

(d) 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?

Scaling the variables generates different hierarchical results and more states fall into the 1st cluster. I belive we should scale the variables before the inter-observation dissimilarities are computed, for the reason that the magnitudes and units for some variables in the orignial arrest data is different(e.g. murder and urban_pop), and scaling the variables could generate variables with same unit to ensure the variables are clustered under the same measurements.