Chapter 10 Problem 9

Andira Putri

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

Background on Hierarchical Clustering

In unsupervised learning, clustering methods aim to find subgroups (or clusters) in the data in hopes of discovering some structure. The partitions in the data set are created by assigning similar observations to one cluster and different observations in other subgroups. What it means for observations to be "similar" or "different" is domain specific– you need to have a good understanding of the data's context. There are two main clustering methods; K-means and hierarchical clustering. K-means partitions a data set into K distinct, non-overlapping clusters. On the contrary, in hierarchical clustering, there is no particular value of K, and the model can be represented by tree-like diagrams called dendrograms.

Understanding Dendrogram Diagrams

Starting from the bottom of the tree…

1. Leaves = observations
2. Leaves fuse to branches = observations that are similar
3. Branches fuse with other leaves or brances = observations not as similar as earlier fusions

Rank of similarity –> tree height. Leaves and branches fusing at lower heights are more similar than the fusions above.

Hierarchical Clustering Algorithm

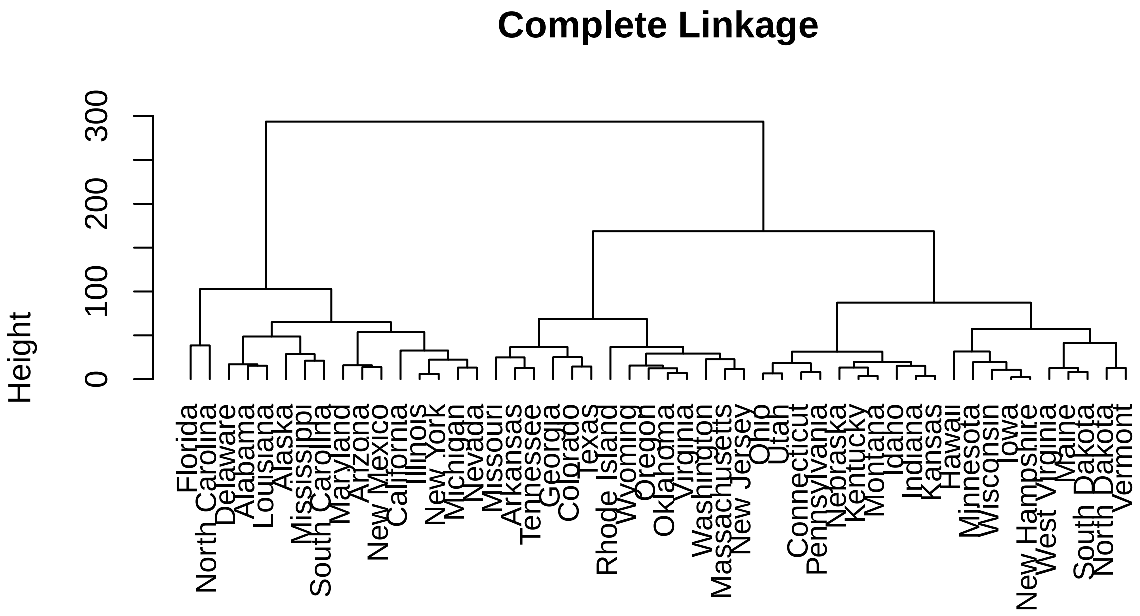
1. Begin with `n` observations and a measure, most commonly Euclidean distance, of all $n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.
2. For i=n,n-1,…,2:

- Examine all pairwise inter-cluster dissimilarities among the i clusters, and identify the pair of clusters that are most similar. Fuse them. Dissimilarity determines height of the fuse.
- Compute pairwise inter-cluster dissimilarities among the i-1 remaining clusters.

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

Complete linkage involves maximal intercluster dissimilarity. We compute pairwise dissimilarites between clusters A and B, and we record the largest dissimilarity.

```r
df=USArrests
#hierarchical clustering
hc.comp=hclust(dist(df),method="complete")
#dendrogram
plot(hc.comp,main="Complete Linkage",xlab="",sub="",cex=0.9,hang=-1)
```



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

```r
#cut tree
cutree(hc.comp,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
```

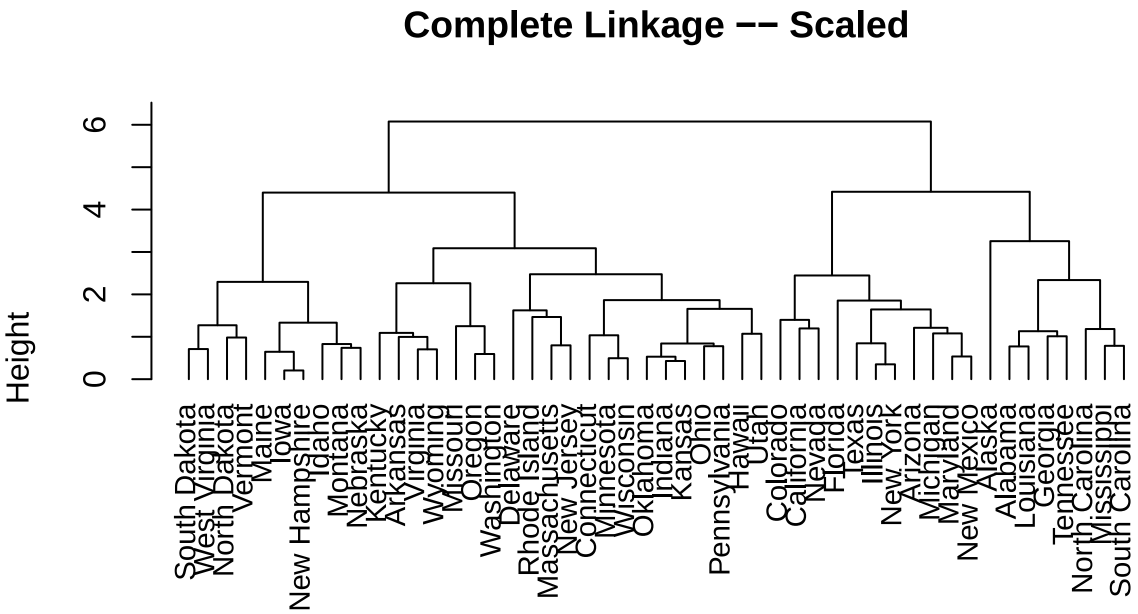
```
#create frequency table
table(cutree(hc.comp,3))
```

```
##
##  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.

For any unsupervised learning method like PCA, K-means, and hierarchical clustering, it is important to considering scaling especially when all measurements may not be in the same units.

```
#scale variables to have mean=0, std dev=1
df.scaled=scale(df)
#repeat hierarchical clustering on new scale
hc.comp.sc=hclust(dist(df.scaled),method="complete")
#dendrogram
plot(hc.comp.sc,main="Complete Linkage -- Scaled",xlab="",sub="",cex=0.9,hang=-1)
```



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? Provide a justification for your answer?

There is a difference in maximum height for both dendrogram plots. The scaled variables make dendrogram plots easier to interpret because the max height is around 7, which is much smaller than the previous dendrogram. In addition, the scaled variables are better because not all variables in the `USArrests` set were measured in the same units.

```r
#cut tree
cutree(hc.comp.sc,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
```

```r
#dendrogram
table(cutree(hc.comp.sc,3))
```

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