Project 5 Part 3

Kaushal Khatiwada

2024-05-13

Use NCI60 data of ISLR2 package

A) Scale the nci.data as sd.data object

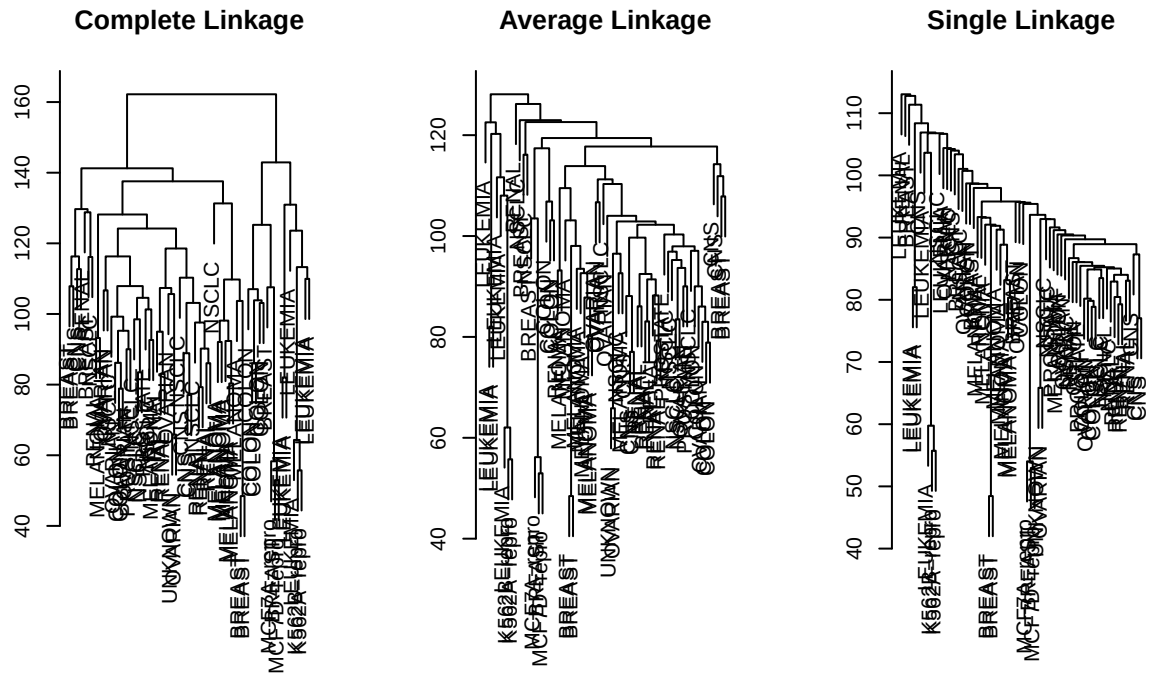
```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.3.3
```

```
nci.data <- NCI60$data
nci.labs <- NCI60$labs
sd.data <- scale(nci.data)
```

B) Fit hierarchical cluster analysis on the sd.data using complete, average and single linkage methods, show the results with dendrogram and interpret them carefully

```
par(mfrow = c(1,3))
data.dist <- dist(sd.data)
plot(hclust(data.dist), xlab = "", sub = "", ylab = "",
     labels = nci.labs, main = "Complete Linkage")
plot(hclust(data.dist, method = "average"), xlab = "", sub = "", ylab = "",
     labels = nci.labs, main = "Average Linkage")
plot(hclust(data.dist,method = "single"), xlab = "", sub = "", ylab = "",
     labels = nci.labs, main = "Single Linkage")
```



C Find the best number for clusters using "cutree" function with best distance value

```
clusters <- hclust(dist(sd.data))
trees <- cutree(clusters, 3)
table(trees, nci.labs)
```

```
##      nci.labs
## trees BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro MCF7D-repro
##     1      5   5     2           0           0        0           0           0
##     2      0   0     0           1           1        6           0           0
##     3      2   0     5           0           0        0           1           1
##      nci.labs
## trees MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##     1        8     9       6        2     9       1
##     2        0     0       0        0     0       0
##     3        0     0       0        0     0       0
```

D) Use your roll number as set.seed and perform k-means clustering on sd.data with the best number of clusters/distance value with nstart=20

```
set.seed(13)
km <- kmeans(sd.data, center=3, nstart = 20)
km.clusters <- km$cluster
```

E) Get summary of the k-means clustering and interpret them carefully

```
summary(km.clusters)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.594   2.000   3.000
```

F) Plot this k-means results using base r plot and cluster package and interpret them carefully

```
par(mfrow = c(1,1))
plot(nci.data,col=km.clusters)
points(km$centers)
```