Sardana_Module10-HW

Prachi Sardana

2023-04-03

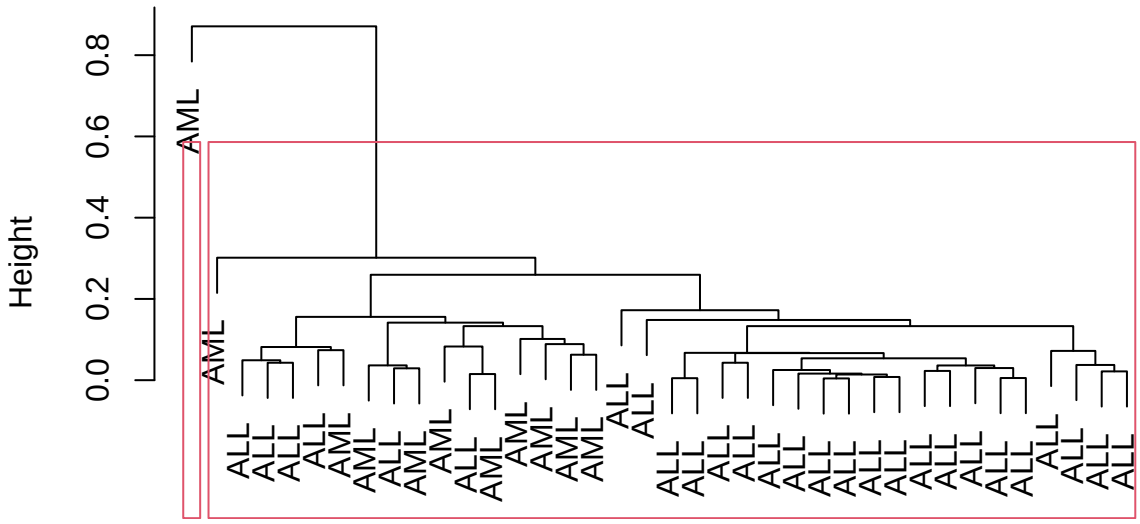
Question 1 a) To conduct hierarchical clustering using single linkage and Ward linkage and Plot the cluster dendrogram for both fit. Get two clusters from each of the methods. Use function table() to compare the clusters with the two patient groups ALL/AML

The ward linkage function works better than the single linkage as it is segregating the data into easy certain clusters

```
# Loaded the data golub
data(golub, package="multtest")

# Factored golub data
gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

grep("CCND3 Cyclin D3",golub.gnames[,2])
```

```
## [1] 1042
```

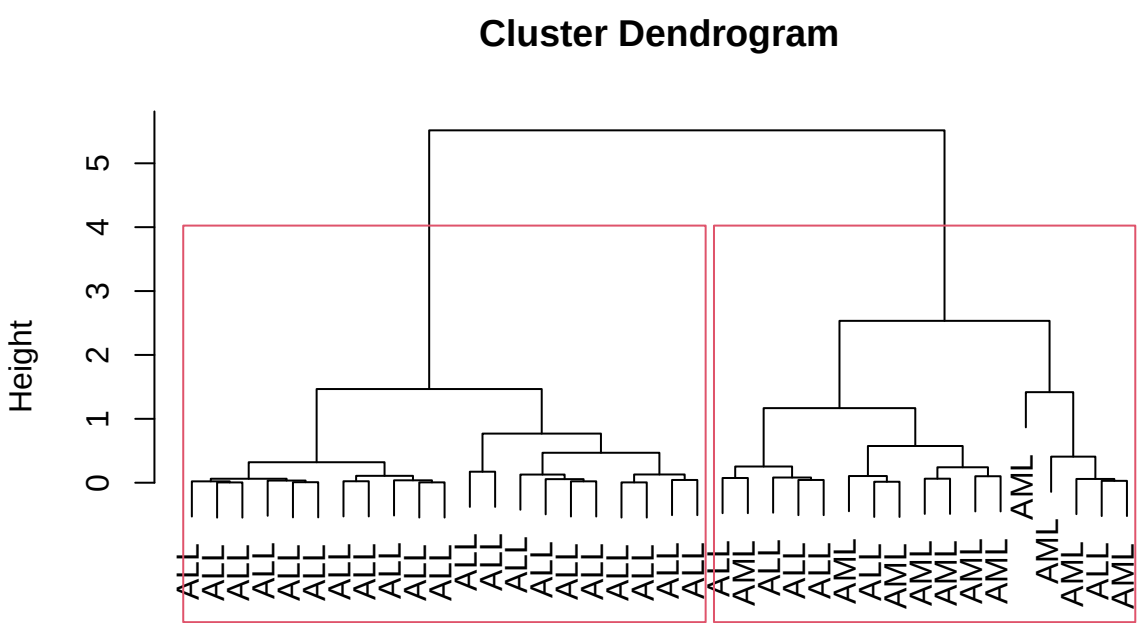
```
clustdata <- data.frame(golub[1042,])

#Single clustering
hcALL.sing<-hclust(dist(clustdata, method="euclidian"), method="single")
plot(hcALL.sing, labels=gol.fac)


hcALL.ward<-hclust(dist(clustdata, method="euclidian"), method="ward.D2")
rect.hclust(hcALL.sing,k=2)
```




```
single_method<-cutree(hcALL.sing, k=2)
table(single_method,gol.fac)
```

```
##              gol.fac
## single_method ALL AML
##             1  27  10
##             2   0   1
```

```
#Ward clustering
plot(hcALL.ward, labels=gol.fac)
rect.hclust(hcALL.ward,k=2)
```




```
ward_method<-cutree(hcALL.ward, k=2)
table(ward_method,gol.fac)
```

```
##             gol.fac
## ward_method ALL AML
##           1  21   0
##           2   6  11
```

b) To use k-means cluster analysis to get two clusters. Use table() to compare the two clusters with the two patient groups ALL/AML.

```
colnames(clustdata) <- c("CCND3 Cyclin D3")
k_meansfit <- kmeans(clustdata, centers = 2)
table(k_meansfit$cluster, gol.fac)
```

```
##    gol.fac
##     ALL AML
##   1  22   1
##   2   5  10
```

c)

```
k_meansfit$centers
```

```
##   CCND3 Cyclin D3
## 1        2.045689
## 2        0.738366
```

d) It is evident that the two clusters maintain their stability even after re sampling. The bootstrap means are nearly identical to the 2-means obtained from the initial clustering, indicating that the estimation bias is minimal. The estimates of cluster means are quite accurate, as demonstrated by the relatively small 95% bootstrap confidence intervals.

```
initial <-k_meansfit$centers
n <- dim(clustdata)[1]; nboot<-1000
boot.CI <- matrix(NA,nrow=nboot,ncol = 2)
for (i in 1:nboot){
  data.star <- clustdata[sample(1:n,replace=TRUE),]
  CI <- kmeans(data.star, initial, nstart = 2)
  boot.CI[i,] <- c(CI$centers[1],CI$centers[2])
}
apply(boot.CI,2,mean)
```

```
## [1] 2.0295892 0.6927931
```

```
quantile(boot.CI[,1],c(0.025,0.975))
```

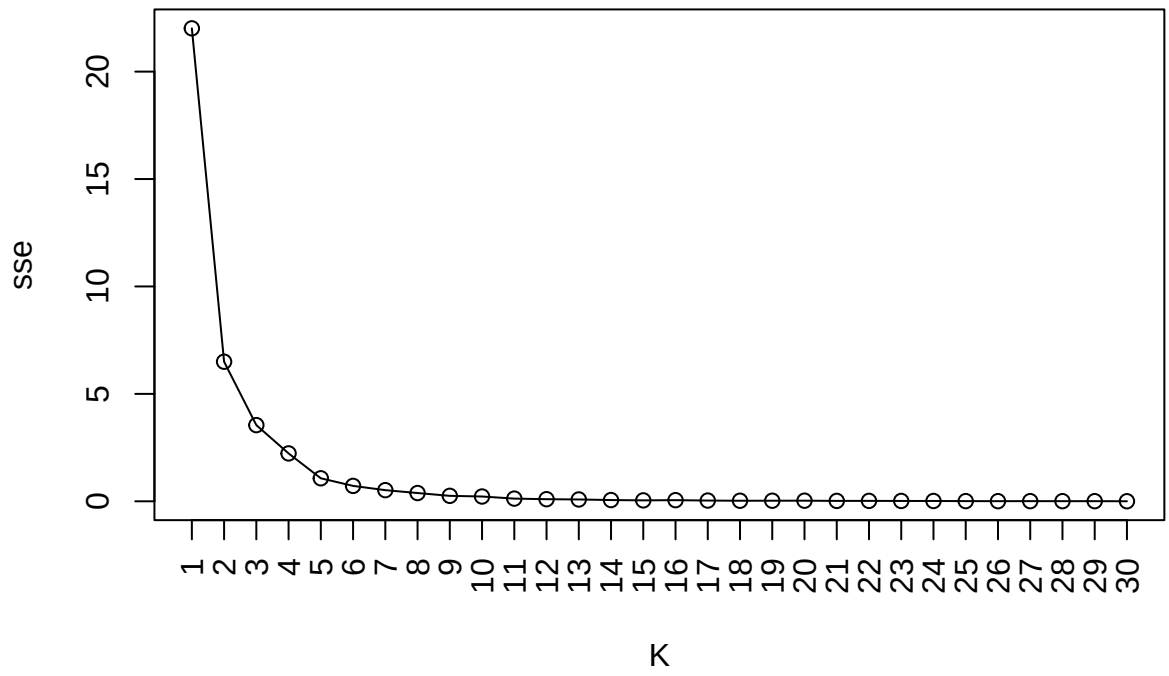
```
##     2.5%    97.5%
## 1.833507 2.194923
```

```
quantile(boot.CI[,2],c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.1989498 1.0521656
```

e) Plot of K versus SSE, for K=1, . . . , 30. A big drop off is observed in SSE from K=1 to K=2. There is a further drop off of SSE to K = 3. Thereafter, decrease in SSE starts to level off. So two or three clusters seems best in the data.

```
K<-(1:30); sse<-rep(NA,length(K))
for (k in K){
  sse[k] <- kmeans(clustdata, centers = k, nstart = 10)$tot.withinss
}
plot(K, sse, type='o', xaxt='n'); axis(1, at = K, las=2)
```



Question 2

a) Selected the oncogenes and antigens from the Golub data using grep function.

```
data(golub, package = "multtest")
gol.fac <- factor(golub.cl, levels = 0:1,labels = c("ALL", "AML"))
oncogenes <- grep("oncogene", golub.gnames[,2])
oncogenes
```

```
##  [1]  501  502  503  587  758  766  775  805  817  819  938 1067 1090 1111 1211
## [16] 1268 1542 1596 1615 1735 1747 1750 1788 1818 1820 1837 1839 2004 2291 2302
## [31] 2488 2517 2661 2681 2692 2703 2714 2715 2892 2981 2990 2993
```

```
antigens <- grep("antigen", golub.gnames[,2])
antigens
```

```
##  [1]  166  313  388  497  504  514  527  540  548  614  646  664  685  763  808
## [16]  826  832  833  834  872  885  890  892  893  926  936  947 1008 1010 1075
## [31] 1087 1208 1258 1279 1287 1412 1422 1467 1531 1616 1645 1719 1748 1752 1756
## [46] 1760 1781 1789 1798 1806 1808 1827 1852 1863 1882 1893 1908 1911 1964 2007
## [61] 2170 2171 2231 2371 2546 2581 2613 2653 2672 2749 2761 2855 2989 3026 3047
```

(b) On the selected data, do clustering analysis for the genes (not for the patients). Using K-means and K-medoids with K=2 to cluster the genes. Use

b)

```
clusdata <- rbind(golub[oncogenes,], golub[antigens,])

clusters.2km <- kmeans(clusdata, centers=2)
g.name<-rep(c("oncogenes","antigens"), c(length(oncogenes), length(antigens)))

# kmeans clustering
table(clusters.2km$cluster, g.name)
```

```
##    g.name
##     antigens oncogenes
##   1       41        22
##   2       34        20
```

```
# k-mediod clustering
library(cluster)
k.pam <- pam(clusdata, k=2)
table(k.pam$clustering, g.name)
```

```
##    g.name
##     antigens oncogenes
##   1       49        29
##   2       26        13
```

c) Fisher test is used to test the marginal independence in two tables.

Hierarchical clustering method provides cluster related to two gene groups.

```
k_means_table <- table(clusters.2km$cluster, g.name)

k_mediod_table <- table(k.pam$clustering, g.name)

fisher.test(k_means_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  k_means_table
## p-value = 0.8484
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4790382 2.5005814
## sample estimates:
## odds ratio
##   1.095394
```

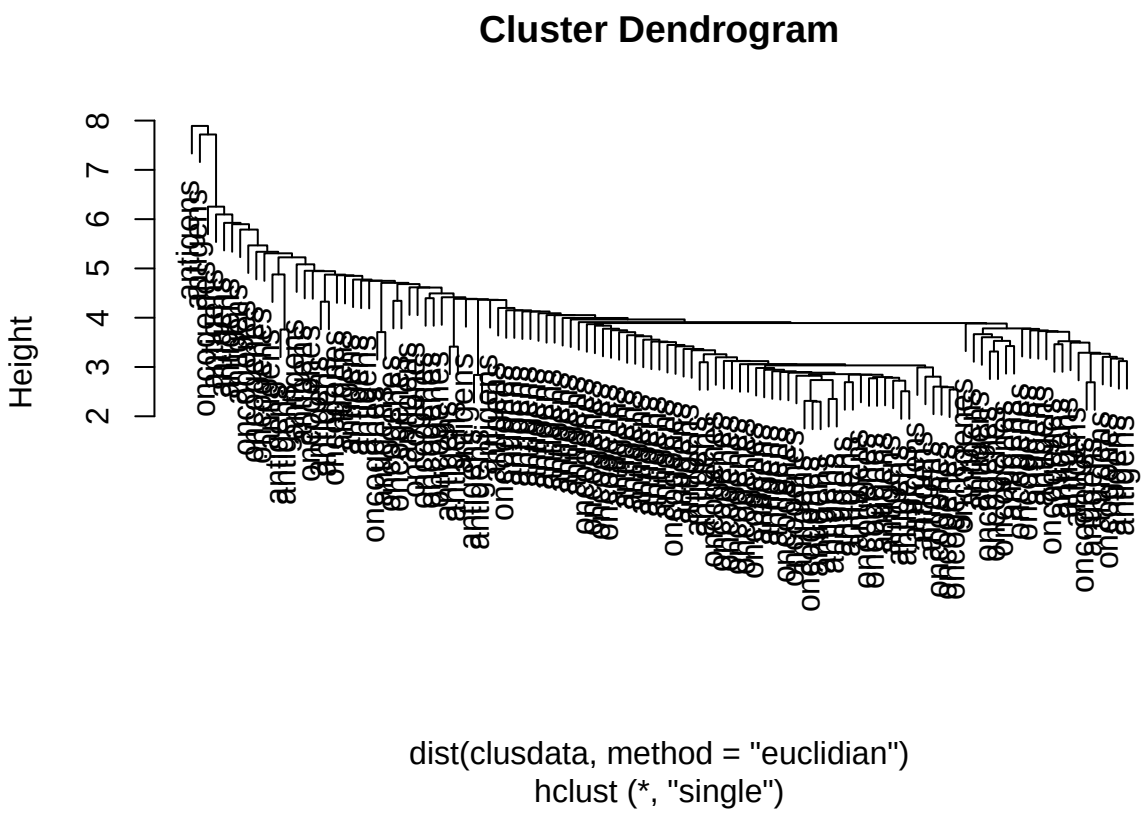
```
fisher.test(k_mediod_table)
```

```
##
##  Fisher's Exact Test for Count Data
```

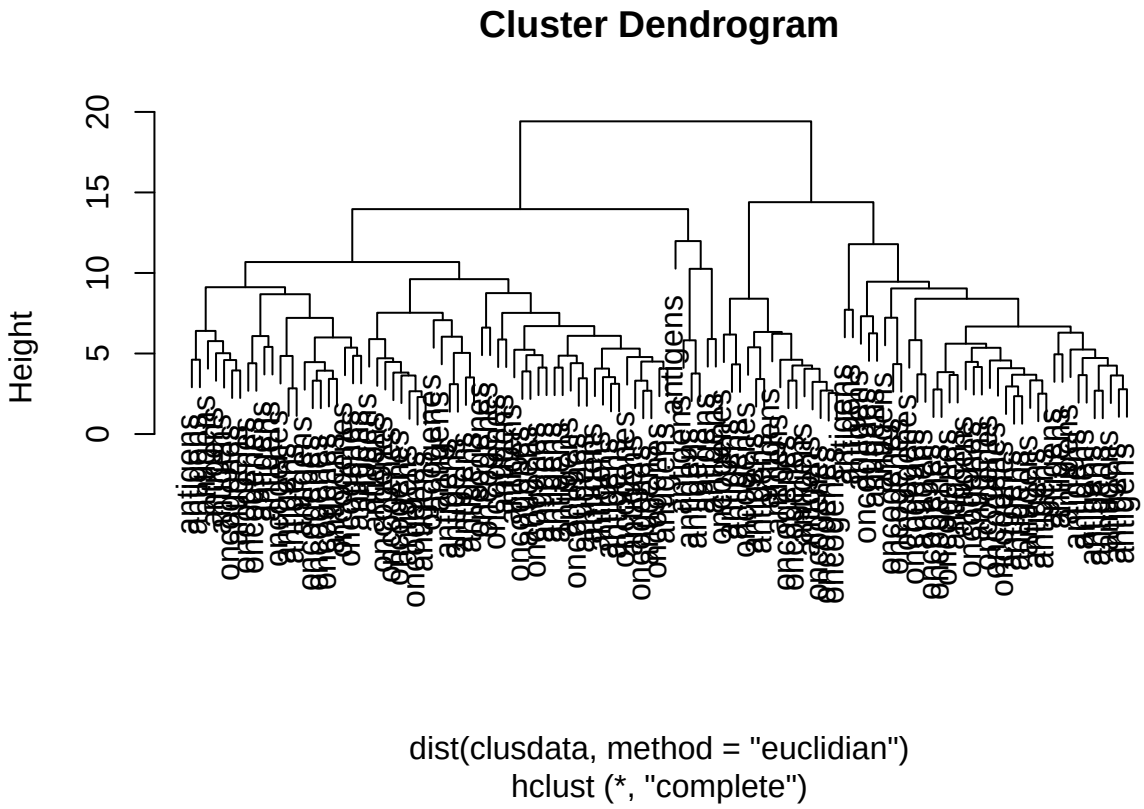
```
##
## data:  k_mediod_table
## p-value = 0.8383
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3426507 2.0276730
## sample estimates:
## odds ratio
##   0.846038
```

d) To plot the cluster dendrograms for this part of golub data with single linkage and complete linkage, using Euclidean distance.

```
clusdata <- rbind(golub[oncogenes,], golub[antigens,])
# Single linkage
hcALL.sing<-hclust(dist(clusdata,method="euclidian"),method="single")
g.name<-rep(c("oncogenes","antigens"), c(length(oncogenes), length(antigens)))

# Cluster dendogram  of single linkage
plot(hcALL.sing, labels=g.name)
```




```
# Complete linkage
hcALL.complete <-hclust(dist(clusdata,method="euclidian"),method="complete")

# Cluster dendogram of complete linkage
plot(hcALL.complete, labels = g.name)
```



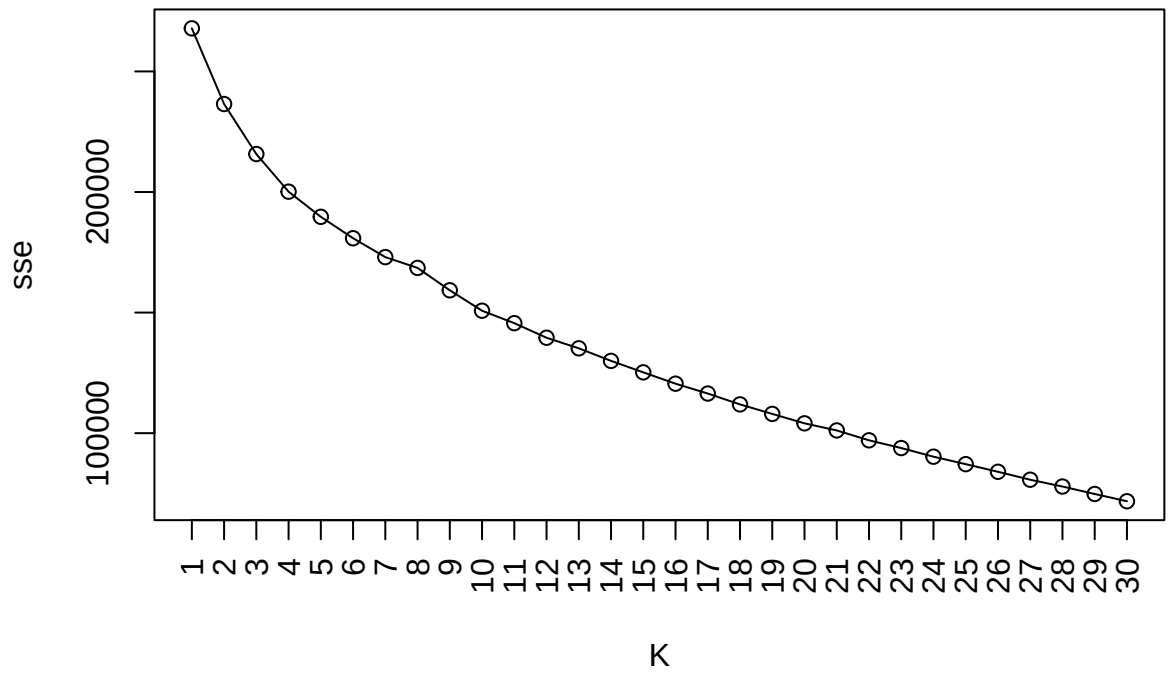
Question 3

a) Using k-means clustering, produce a plot of K versus SSE, for K=1,. . . , 30 clustering starts drop after k = 7 .

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.2.3
```

```
# Loading NCIdata
ncidata <- NCI60$data
# Loading cancer cell lines contained in ncilabs
ncilabs <- NCI60$labs

# k means clustering
K<-(1:30); sse<-rep(NA,length(K))
for (k in K){
  sse[k] <- kmeans(ncidata, centers = k, nstart = 10)$tot.withinss
}
plot(K, sse, type='o', xaxt='n'); axis(1, at = K, las=2)
```



b) K-medoids clustering (K=7) with 1-correlation as the dissimilarity measure

```
k_medoid<-pam(as.dist(1-cor(t(ncidata))),k=7)
table(ncilabs,k_medoid$clustering)
```

```
##
## ncilabs       1 2 3 4 5 6 7
##   BREAST      0 3 0 0 2 0 2
##   CNS         1 4 0 0 0 0 0
##   COLON       0 0 0 7 0 0 0
##   K562A-repro 0 0 0 0 0 1 0
##   K562B-repro 0 0 0 0 0 1 0
##   LEUKEMIA    0 0 0 0 0 6 0
##   MCF7A-repro 0 0 0 0 1 0 0
##   MCF7D-repro 0 0 0 0 1 0 0
##   MELANOMA    0 1 0 0 0 0 7
##   NSCLC       2 2 0 3 1 1 0
##   OVARIAN     2 0 1 2 1 0 0
##   PROSTATE    0 0 1 1 0 0 0
##   RENAL       7 1 1 0 0 0 0
##   UNKNOWN     0 0 1 0 0 0 0
```

Through the table , breast cancer and prostate cancer share very less similarity in clustering pattern while colon cancer and leukemia are defined into single cluster. The small cell lung cancer (NSCLC) have a large clustering which is similar to clustering pattern of ovarian cancer.