Math7340 HW10

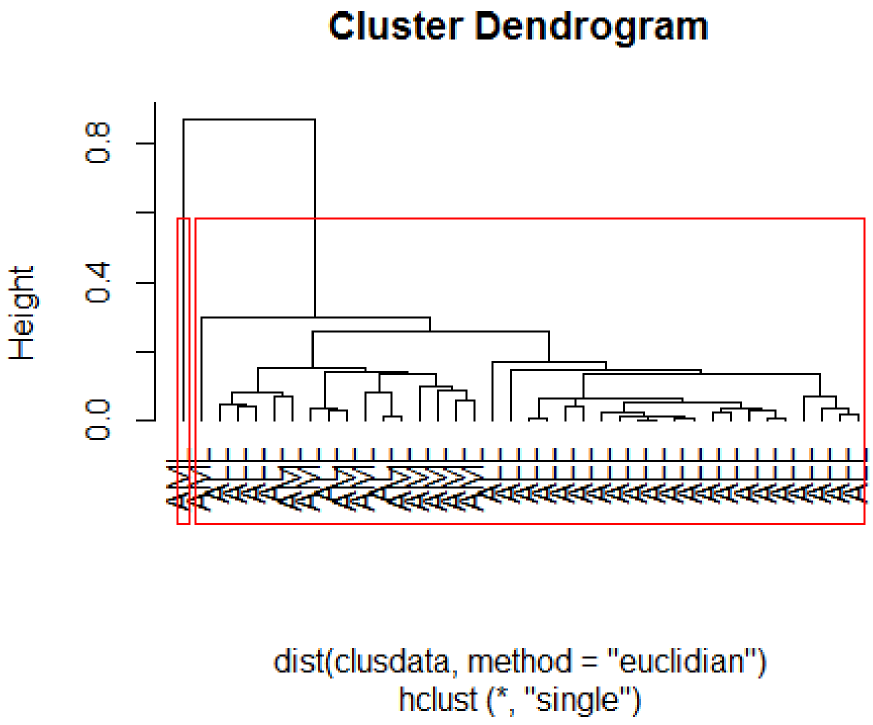
Chengbo Gu

Problem 1: (40 points)

Clustering analysis on the "CCND3 Cyclin D3" gene expression values of the Golub et al. (1999) data.

(a) Conduct hierarchical clustering using single linkage and Ward linkage. 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. Which linkage function seems to work better here?

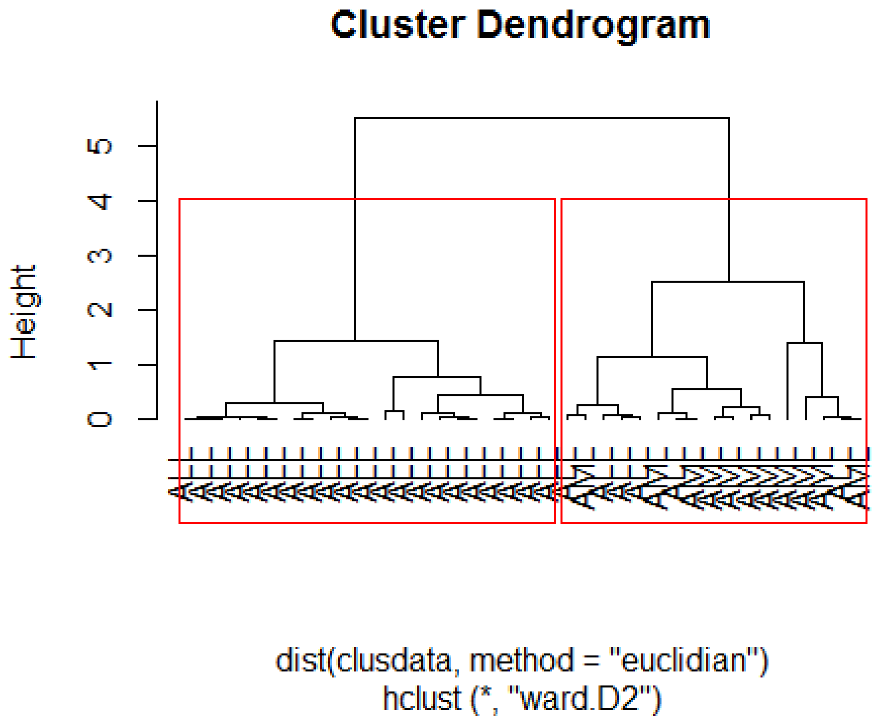
```
data(golub, package="multtest")
clusdata <- data.frame(golub[1042,])
colnames(clusdata)<-c("CCND3 Cyclin D3")
gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

hc.single <- hclust(dist(clusdata, method="euclidian"), method="single")
plot(hc.single, hang=-1, labels=gol.fac)
rect.hclust(hc.single, k=2)
```



```
groups.single <- cutree(hc.single, k=2)
table(gol.fac, groups.single)
```

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

```
hc.ward <- hclust(dist(clusdata,method="euclidian"),method="ward.D2")
plot(hc.ward, hang=-1, labels=gol.fac)
rect.hclust(hc.ward, k=2)
```



```
groups.ward <- cutree(hc.ward, k=2)
table(gol.fac, groups.ward)
```

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

Ward linkage works better here.

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

```
cl.2mean <- kmeans(clusdata, centers=2, nstart = 10)
table(gol.fac, cl.2mean$cluster)
```

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

(c) Which clustering approach (hierarchical versus k-means) produce the best matches to the two diagnose groups ALL/AML?

They are equally good.

(d) Find the two cluster means from the k-means cluster analysis. Perform a bootstrap on the cluster means. Do the confidence intervals for the cluster means overlap? Which of these two cluster means is estimated more accurately?

```
initial <-cl.2mean$centers
n <- dim(clusdata)[1]
nboot<-1000
boot.cl <- matrix(NA,nrow=nboot,ncol = 2)
for (i in 1:nboot){
  dat.star <- clusdata[sample(1:n,replace=TRUE),]
  cl <- kmeans(dat.star, initial, nstart = 10)
  boot.cl[i,] <- c(cl$centers[,1])
}
apply(boot.cl,2,mean)
```

```
## [1] 0.6929059 2.0316467
```

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

```
##       2.5%      97.5%
## 0.07525656 1.05858157
```

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

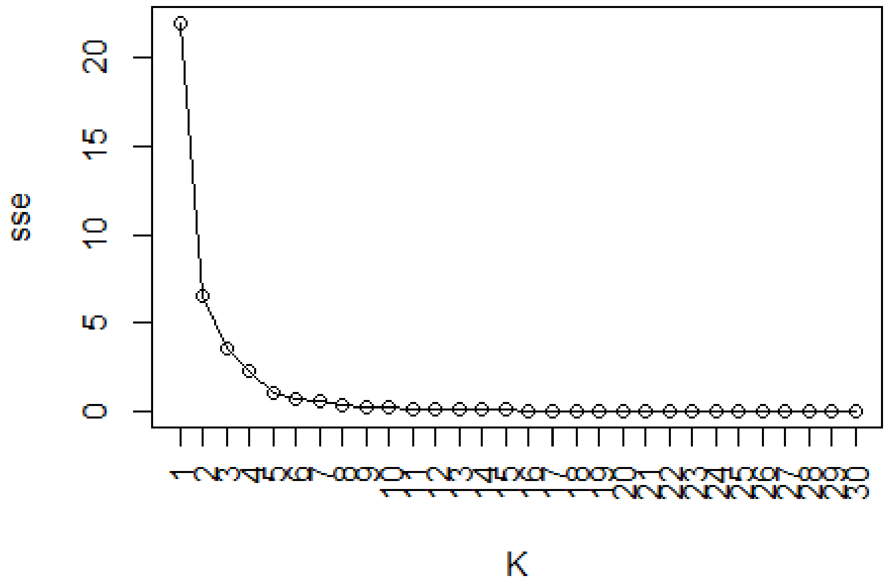
```
##     2.5%    97.5%
## 1.839615 2.203065
```

CIs don't overlap.

Cluster mean `2.0316467` is estimated more accurately because of shorter CI.

(e) Produce a plot of K versus SSE, for K=1, …, 30. How many clusters does this plot suggest?

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



Seems that 2-4 clusters are appropriate.

Problem 2 (30 points):

Cluster analysis on part of Golub data.

(a) Select the oncogenes and antigens from the Golub data. (Hint: Use grep()).

```
data(golub, package="multtest")

oncoIndex <- grep("oncogene", golub.gnames[,2])
antiIndex <- grep("antigen", golub.gnames[,2])
tag <- c(rep(0, length(oncoIndex)), rep(1, length(antiIndex)))
clusdata <- rbind(golub[oncoIndex,], golub[antiIndex,])
gol.fac <- factor(tag,levels=0:1, labels= c("oncogene","antigen"))
```

(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 table() to compare the resulting two clusters with the two gene groups oncogenes and antigens for each of the two clustering analysis.

```
library(cluster)

cl.2mean <- kmeans(clusdata, centers=2, nstart=10)
meanTable <- table(gol.fac, cl.2mean$cluster)
meanTable
```

```
##
## gol.fac     1  2
##   oncogene 20 22
##   antigen  34 41
```

```
cl.2medoid <- pam(dist(clusdata, method='eucl'), k=2)
medoidTable <- table(gol.fac, cl.2medoid$cluster)
medoidTable
```

```
##
## gol.fac     1  2
##   oncogene 29 13
##   antigen  49 26
```

(c) Use appropriate tests (from previous modules) to test the marginal independence in the two by two tables in (b). Which clustering method provides clusters related to the two gene groups?

```
fisher.test(meanTable)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  meanTable
## 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
```

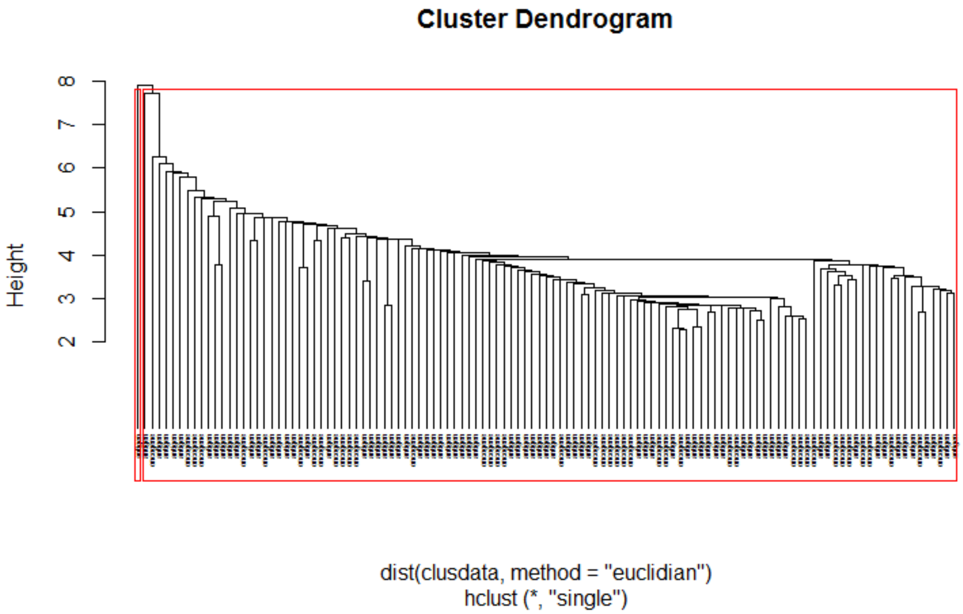
```
fisher.test(medoidTable)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  medoidTable
## p-value = 0.8383
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4931762 2.9184238
## sample estimates:
## odds ratio
##    1.18198
```

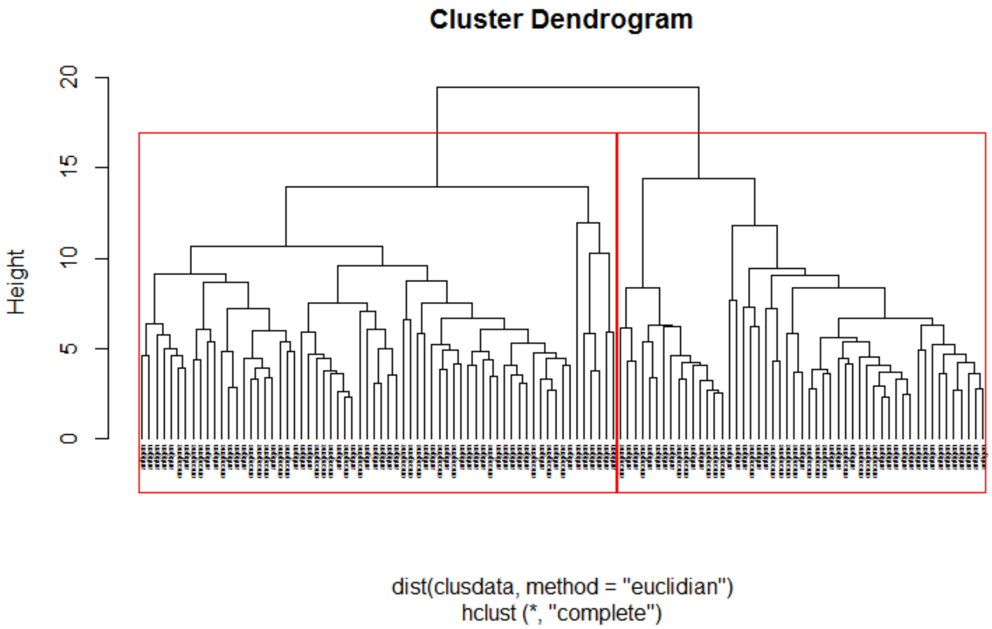
Both clustering methods have p-values that are greater than 0.05. So none of them provides clusters related to the two gene groups.

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

```
hc.single <- hclust(dist(clusdata, method="euclidian"), method="single")
plot(hc.single, hang=-1, labels=gol.fac, cex=0.38)
rect.hclust(hc.single, k=2)
```



```
hc.complete <- hclust(dist(clusdata, method="euclidian"), method="compl
ete")
plot(hc.complete, hang=-1, labels=gol.fac, cex=0.38)
rect.hclust(hc.complete, k=2)
```



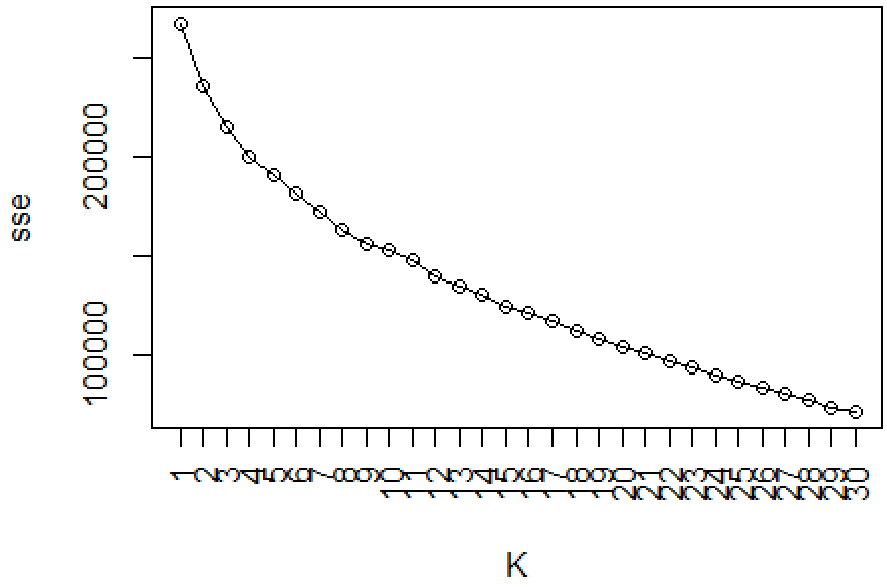
Problem 3 (30 points):

Clustering analysis on NCI60 cancer cell line microarray data (Ross et al. 2000)

The ncidata (64 by 6830 matrix) contains 6830 gene expression measurements on 64 cancer cell lines. The cancer cell lines labels are contained in ncilabs. We do clustering analysis on the 64 cell lines (the rows).

(a) Using k-means clustering, produce a plot of K versus SSE, for K=1,..., 30. How many clusters appears to be there?

```r
library(ISLR)
ncidata <- NCI60$data
ncilabs <- NCI60$labs

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)
```



7~9 seems to be good.

(b) Do K-medoids clustering (K=7) with 1-correlation as the dissimilarity measure on the data. Compare the clusters with the cell lines. Which types of cancer are well identified in a cluster? Which types of cancer are not grouped into a cluster? According to the clustering results, which types of cancer are most similar to ovarian cancer?

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

```
##
## 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
```

Well identified: CNS, COLON, LEUKEMIA, MELANOMA, RENAL

Not grouped: BREAST, NSCLC, OVARIAN, PROSTATE

NSCLC is the most similar to ovarian cancer.