ComBio HomeWork2

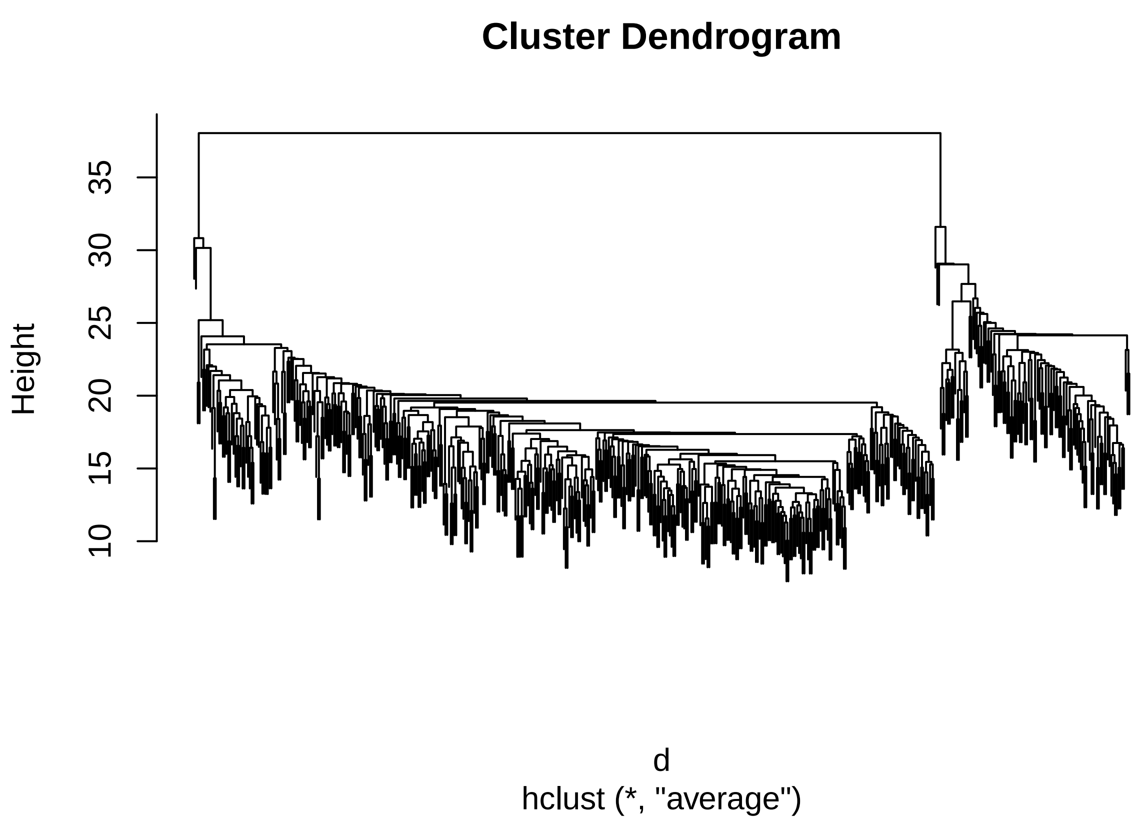
Jiaqi Ma

Probelm 1

Hierarchical Clustering on Original Data

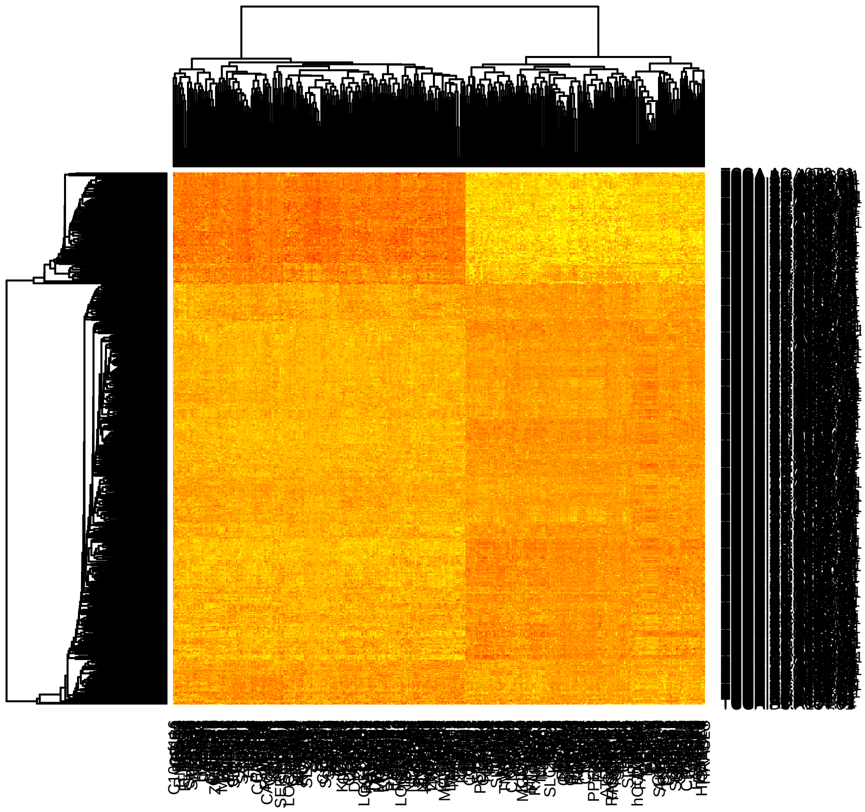
Load training data and plot hierarchical clustring result:

```
dataset<-read.table("GeneMatrix.txt")
m_matrix <- data.matrix(dataset)
d<-dist(t(dataset))
cl<-hclust(d,method="average")
dendcl <- as.dendrogram(cl)
plot(cl,labels=FALSE)
```



Heatmap:

```
heatmap(t(dataset), Rowv=dendcl, Colv=F, scale="none")
```



Load clinical data and caculate accuracy:

```
pre_data<-read.delim("clinical_data.txt")
gnd_SID<-gsub("-",".",pre_data$sampleID)
gnd_label<-pre_data$ER_Status_nature2012
pre_label=cutree(cl,k=2)
cnt=0
for (i in 1:length(pre_label))
  {
  if(pre_label[i]==1 && as.character(gnd_label[which(gnd_SID==names(pre_label)[i])])=="Positive" || pre_
    {
    cnt=cnt+1
    }
  }
```

Accurate numbers:

```
print(cnt)
```

```
## [1] 489
```

All numbers:

```
print(length(pre_label))
```

```
## [1] 522
```

Accuracy:

```
print(cnt/length(pre_label))
```

```
## [1] 0.9367816
```

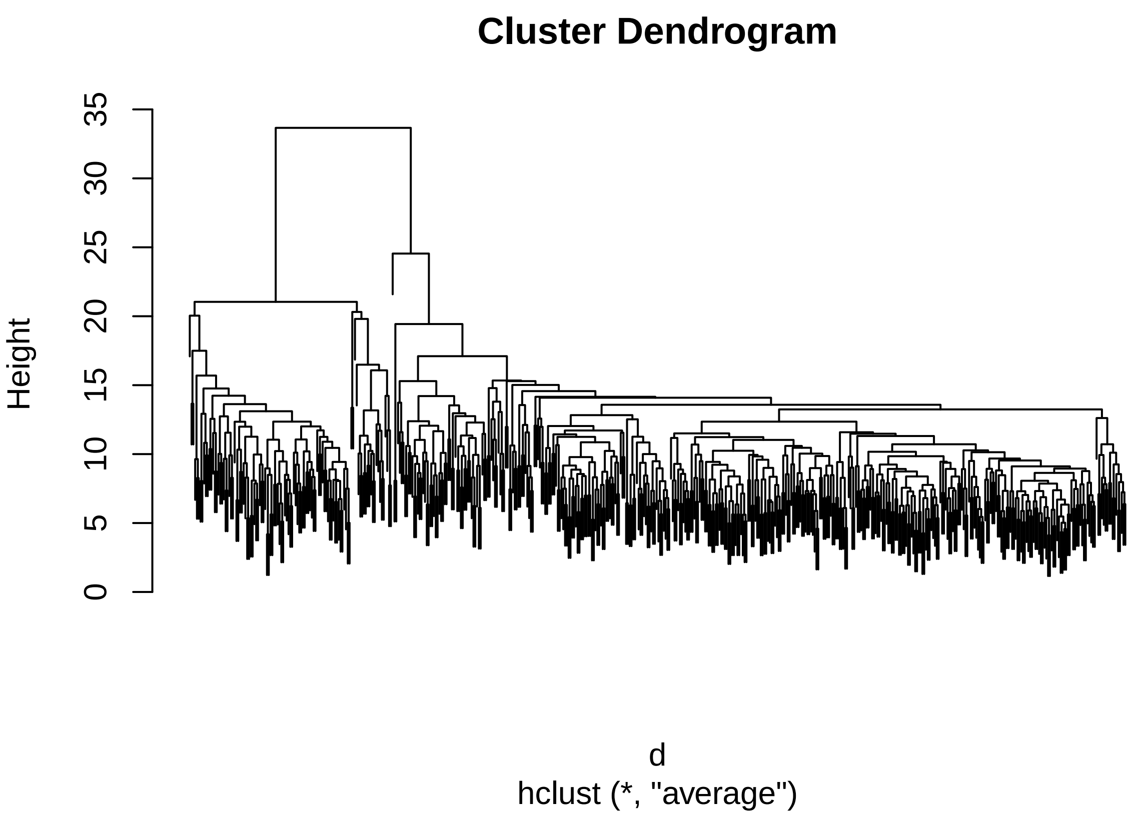
Hierarchical Clustering on PCA Data

Exact PCA from the original data :

```
ev<-eigen((m_matrix)%*%t(m_matrix))
pca<-t(m_matrix)%*%ev$vectors[,1:20]
```

Hierarchical clustering on PCA data:

```
d<-dist(pca)
cl<-hclust(d,method="average")
dendcl <- as.dendrogram(cl)
plot(cl,labels=FALSE)
```



Caculate the accuracy:

```
pre_label=cutree(cl,k=2)
cnt=0
for (i in 1:length(pre_label))
  {
```

```
  if(pre_label[i]==1 && as.character(gnd_label[which(gnd_SID==names(pre_label)[i])])=="Positive" || pre_
    {
    cnt=cnt+1
    }
  }
```

Accurate numbers:

```
print(cnt)
```

```
## [1] 487
```

All numbers:

```
print(length(pre_label))
```

```
## [1] 522
```

Accuracy:

```
print(cnt/length(pre_label))
```

```
## [1] 0.9329502
```

Problem 2

PCA in EIGENSTRAT.

Caculate the eigen vectors of covariance matrix of the data.

```
data = matrix( c(1,0,2,0,2,0,2,
                 1,1,1,0,1,0,2,
                 1,2,1,1,1,1,1,
                 0,1,0,2,0,1,1,
                 0,2,1,2,0,1,0), nrow=7, ncol = 5)
ev<-eigen(t(data)%*%data)
```

The first principal component:

```
ev$vectors[,1]
```

```
## [1] -0.4875976 -0.4361288 -0.5386884 -0.3374341 -0.4098698
```

The result of hierarchical clustering on PCA is similar to that on the original data thus shows that PCA could significantly compress the data without much information loss.

Problem 3

Maximum Variance Formulation

Given a dataset $\{x_n\}$ where $n = 1, 2, ..., n$ and x_n is a D dimensional vector, the goal is to project the data onto a $M < D$ dimensional space while maximizing the variance of the projected data.

Let u_1 be a D dimentional unit vector. The mean of the sample set $\bar{x}$ is

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

then the variance of the data projected on u_1 is

$$\frac{1}{N} \sum_{n=1}^{N} u_1^T x_n - u_1^T \bar{x}^2 = u_1^T S u_1$$

where S is the data covariance matrix

$$S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$$

Note that our goal is

$$\max_{u_1} u_1^T S u_1$$

s.t.

$$u_1^T u_1 = 1$$

The solution of this constrained convex optimization problem is the first eigenvector of S.

Similarly, when asking for $M > 1$ dimention of PCs, the solution will be the first M eigenvectors.

Minimum-Error Formulation

Given a dataset $\{x_n\}$ where $n = 1, 2, ..., n$ and x_n is a D dimensional vector, the goal is to project the data onto a $M < D$ dimensional space while minimizing the error between the projected data and the original data.

Given a complete orthonormal set of D dimensional basis vectors $\{u_i\}$, where

$$u_i^T u_j = \delta_{ij}, i, j = 1, 2, ..., D$$

then the oringinal data can be represented by

$$x_n = \sum_{i=1}^{D} \alpha_{ni} u_i = \sum_{i=1}^{D} (x_n^T u_i) u_i$$

The projected data can be represented by

$$\tilde{x}_n = \sum_{i}^{M} z_{ni} u_i + \sum_{i=M+1}^{D} b_i u_i$$

and the error can be represented by

$$J = \frac{1}{N} \sum_{n=1}^{N} ||x_n - \tilde{x}_n||^2$$

To minimize J w.r.t. $\{z_{ni}\}$ and $\{b_i\}$, setting the derivatives to zero and we obtain

$$z_{ni} = x_n^T u_i, i = 1, ..., M$$

and

$$b_i = \bar{x}^T u_i, i = M + 1, ..., D$$

If we substitute z_{ni}, b_i in J and we can obtain

$$J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} (x_n^T u_i - \bar{x}^T u_i)^2 = \sum_{i=M+1}^{D} u_i^T S u_i$$

Hence the goal is to solve the optimization problem

$$\min_u J$$

s.t.

$$u^T u = I$$

The solution is that $\{u_i\}$,$i = M + 1, ..., D$ should be the smallest $(D - M)$ eigenvectors of S and thus the PCs should be the largest M eigenvectors.