Homework 2

Fei Xia

December 19, 2015

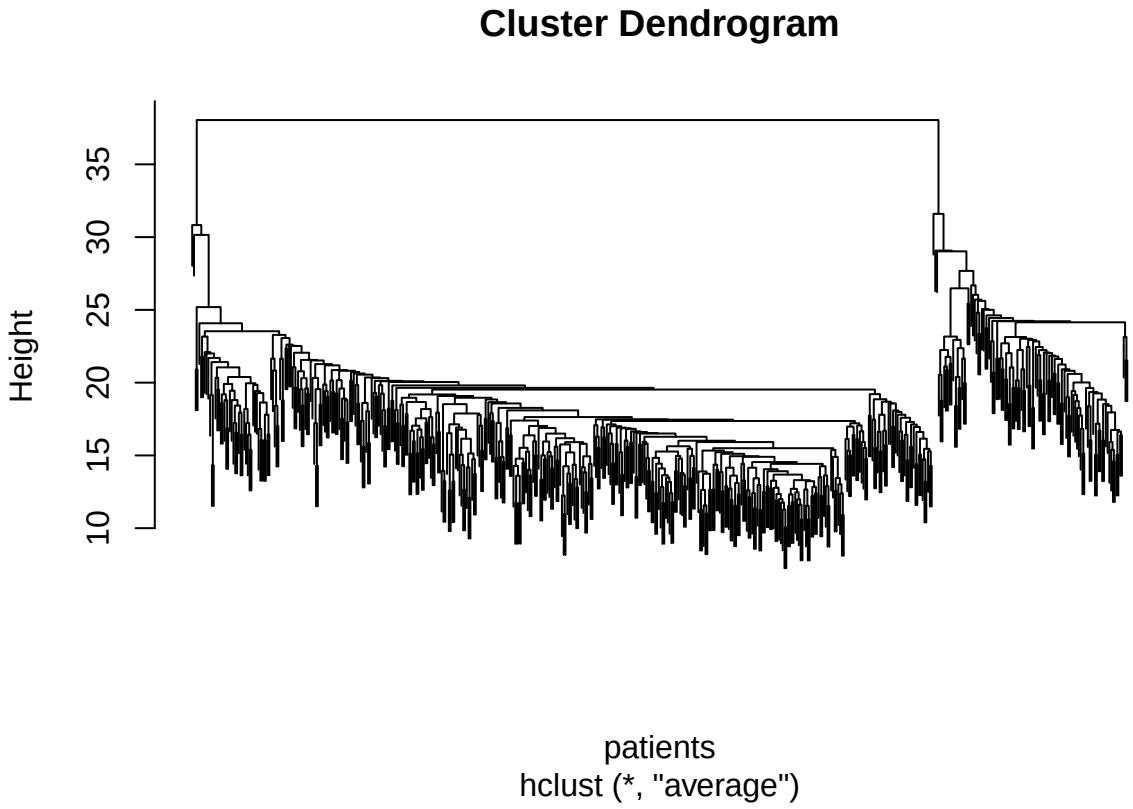
Question 1

Read in the file, compute distance

```
mat<-t(read.table("GeneMatrix.txt", header=TRUE, sep = "\t",row.names = 1,as.is=TRUE))
di <- dist(mat)
hc <- hclust(di, method = "average")
```

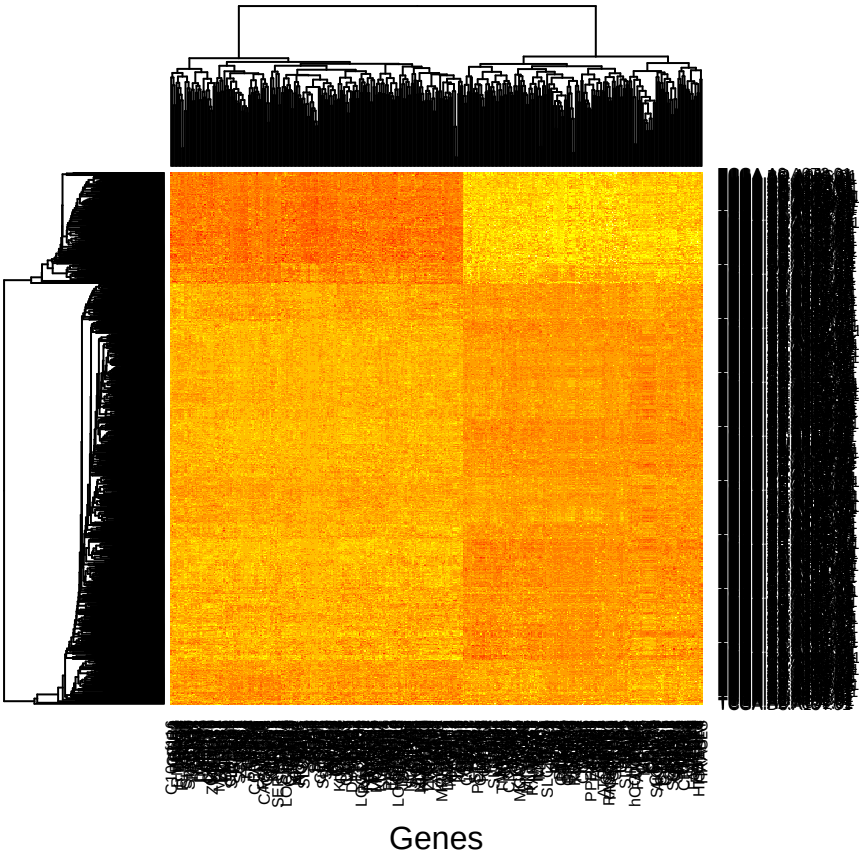
Hierarchical clustering of patients (too many labels, so labels are not plotted):

```
plot(hc, xlab="patients", labels = FALSE)
```



Heatmap:

```
heatmap(mat, Colv=F, Rowv = as.dendrogram(hc), scale='none', xlab = "Genes", ylab = "Patients")
```



Read the clinical data (ground truth):

```
g<-read.table("clinical_data", header=TRUE, sep = "\t",row.names = 1,as.is=TRUE)
labels <- g[,"ER_Status_nature2012"]
v <- rownames(g)
```

This is a 3-classification problem, use accuracy(true/all) as metric.

Get the clustering results:

```
groups <- cutree(hc, k=3)
vg <- rownames(mat)
```

Calculate accuracy:

```
nz <- which(labels!="")
labels <- data.frame(labels[nz])
v <- v[nz]
v <- gsub("-", ".", v)
rownames(labels) <- v
inter <- intersect(v,vg)
prediction <- groups[inter]
ground <- labels[inter,]
result <- data.frame(labels[inter,])
result[,"prediction"] = groups[inter]
colnames(result) <- c("ground_truth", "prediction")
```

```
rownames(result) <- inter
result$ground_truth <- gsub("Positive", 1, result$ground_truth)
result$ground_truth <- gsub("Negative", 2, result$ground_truth)
result$ground_truth <- gsub("Indeterminate", 3, result$ground_truth)
#result
```

Show confusion matrix:

```
tbl<-table(Ground_Truth = result$ground_truth, Predicted = result$prediction)
tbl
```

```
##              Predicted
## Ground_Truth   1   2   3
##            1 332   9   1
##            2  16  82   0
```

Get accuracy:

```
acc = (tbl[1,1] + tbl[2,2])/length(inter)
acc
```

```
## [1] 0.9409091
```

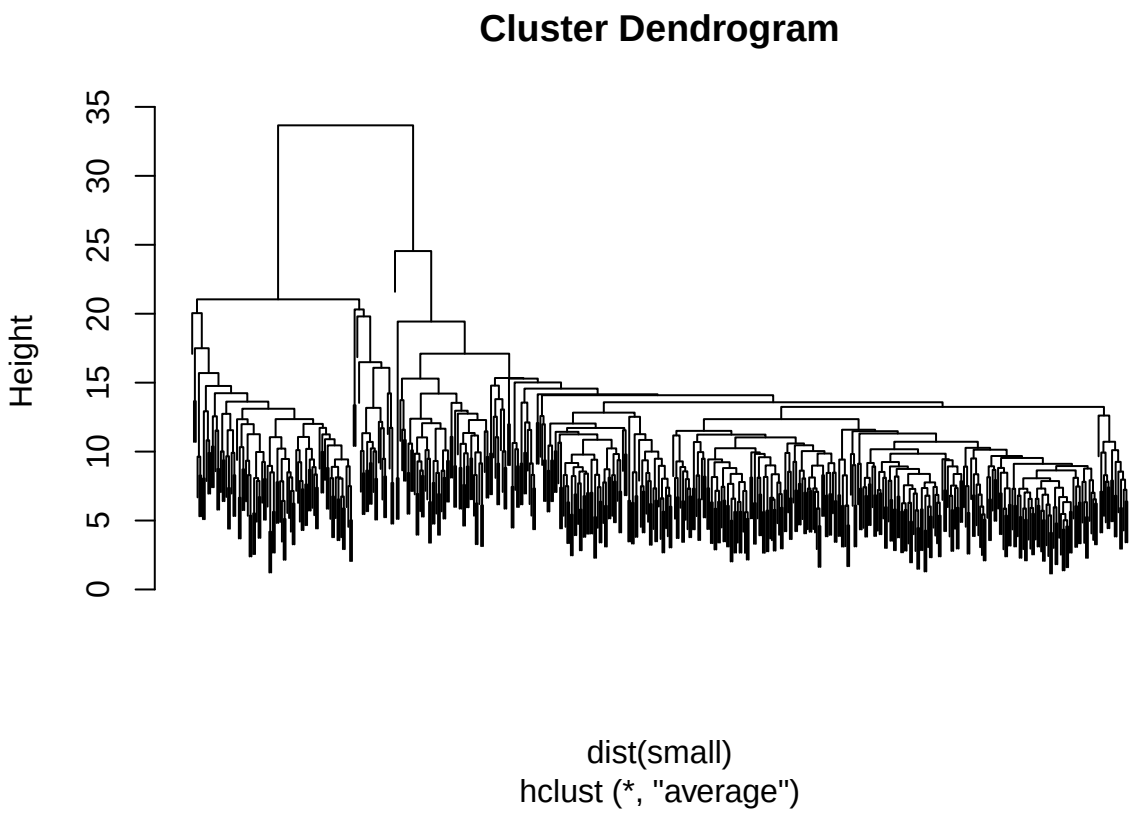
PCA implementation

Use eigenvector decomposition to implement PCA, choose number of principle components = 20.

```
x = cov(mat)
r<-eigen(x)
small <- as.matrix(mat) %*% r$vectors[,1:20]
```

Plot clustering:

```
hc = hclust(dist(small), method = "average")
plot(hc , labels = FALSE)
```



Get the clustering results:

```
groups <- cutree(hc, k=3)
vg <- rownames(mat)
```

Calculate accuracy:

```
nz <- which(labels!="")
prediction <- groups[inter]
result <- data.frame(labels[inter,])
result[,"prediction"] = groups[inter]
colnames(result) <- c("ground_truth", "prediction")
rownames(result) <- inter
result$ground_truth <- gsub("Positive", 1, result$ground_truth)
result$ground_truth <- gsub("Negative", 2, result$ground_truth)
result$ground_truth <- gsub("Indeterminate", 3, result$ground_truth)
#result
```

Show confusion matrix:

```
tbl<-table(Ground_Truth = result$ground_truth, Predicted = result$prediction)
tbl
```

```
##             Predicted
## Ground_Truth   1   2
##            1 329  13
##            2  15  83
```

Get accuracy:

```
acc = (tbl[1,1] + tbl[2,2])/length(inter)
acc
```

```
## [1] 0.9363636
```

The accuracy is similar to that without using PCA, but we reduced the dimension to 20.

Question 2

Principal components analysis corrects for stratification in genome-wide association studies.

```
mat = 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)
x = cov(mat)
r<-eigen(x)
```

Axis of variation:

```
r$vectors[,1]
```

```
## [1] -0.6383335 -0.3494259  0.1082564  0.4002980  0.5463277
```

This is the same (up to scales) to the one in the paper.

Question 3

Minimum-error formulation

Question: find a set of D-dimensional complete basis $\{u_i\}$, that satisfy

$$u_i^T u_j = \delta_{ij}$$

such that approximation

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

has the minimum error.

The error is:

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

minimize w.r.t z_{ni} gives

$$z_{nj} = x_n^T u_j$$

minimize w.r.t b_j gives:

$$b_j = \bar{x}^T u_j$$

Maximum variance formulation

Sample mean is given by

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

Variance of projected data is given by

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

where S is the data covariance matrix.

Maximize the variance get:

$$Su_1 = \lambda_1 u_1$$

So set u_1 to be the the first eigen vector. The rest can be easily shown using proof by induction.