Clustering

K-means

We'll use the tissues gene expression file to look at clustering. There are expressions for 100 genes in 1816 samples. These samples come from 5 different tissues.

```
library(rafalib)
e <- read.delim("data/expr.txt", row.names=1)
tab <- read.table("data/class_labels.txt", header=T)
head(tab)
```

```
##                  sample_id tissue
## 1 GTEX-111CU-1826-SM-5GZYN      3
## 2 GTEX-111FC-0226-SM-5N9B8      3
## 3 GTEX-111VG-2326-SM-5N9BK      3
## 4 GTEX-111YS-2426-SM-5GZZQ      3
## 5 GTEX-1122O-2026-SM-5NQ91      3
## 6 GTEX-1128S-2126-SM-5H12U      3
```

```
head(e)[,1:2]
```

```
##                    GTEX.111CU.1826.SM.5GZYN GTEX.111FC.0226.SM.5N9B8
## ENSG00000104879.4                -0.4651989               -0.2794045
## ENSG00000143632.10               -0.4107903               -0.1409032
## ENSG00000244734.2                -1.0309454               -0.7426210
## ENSG00000188536.8                -0.9517054               -0.8217659
## ENSG00000206172.4                -0.9177072               -0.8854275
## ENSG00000111245.10               -0.6066508                0.0150647
```

We train the in-built `kmeans` model with 5 clusters on the first 5 samples, for a maximum of 10 iterations.

```
set.seed(1)
km <- kmeans(t(e[1:5,]), centers=5, iter.max=10)
cbind(cluster=c(1,2,3,4,5), samples=km$size)
```

```
##      cluster samples
## [1,]       1     422
## [2,]       2     237
## [3,]       3     196
## [4,]       4     400
## [5,]       5     561
```

```
table(tissue=tissue, cluster=km$cluster)
```

```
##       cluster
## tissue   1   2   3   4   5
##      0   0 237 193   0   0
##      1 217   0   0   5  98
##      2 137   0   1   1 184
##      3  68   0   2   1 279
##      4   0   0   0 393   0
```

Bayesian Information Criterion

We'll now optimize the number of clusters (k) using Bayesian Information Criterion,

$$BIC = -2 \log \hat{L} + m \log n$$

Visualizing our output

```
d <- dist(t(e)) # distance between sample points
km <- kmeans(t(e), centers=5)
mds <- cmdscale(d)

mypar(1,2)
```

```
plot(mds[,1], mds[,2], col=km$cluster, pch=16)
```