Hierarchical Clustering of a Breast Cancer Dataset

Jeremy Haakenson

2024-06-11

The data can be found at:

https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

Load packages.

```
library(dplyr)
library(ggplot2)
library(ggdendro)
```

Load data.

```
bc1 = read.csv('Breast_Cancer.csv')
```

Update column names.

```
colnames(bc1)[c(6, 14)] = c('stage', 'Regional.Node.Positive')
```

Convert Grade to numeric and stage to categorical.

```
bc2 = bc1 %>%
  mutate(Grade = as.numeric(as.factor(Grade)),
         stage = as.factor(stage))

bc2$Grade = bc2$Grade - 1
```

Check for NAs.

```
sum(is.na(bc2))
```

```
## [1] 0
```

There are no NAs.

Pull out the numeric columns and stage.

```
bc3 = bc2[c(1, 8, 10, 13:15, 6)]
```

See which stages are represented in the data.

```
levels(bc3$stage)
```

```
## [1] "IIA"  "IIB"  "IIIA" "IIIB" "IIIC"
```

I will use the mean values for each stage.

```
s2a = bc3 %>%
  filter(stage == 'IIA') %>%
  summarize(age.mean = mean(Age),
            grade.mean = mean(Grade),
            tumor.mean = mean(Tumor.Size),
            exam.mean = mean(Regional.Node.Examined),
            pos.mean = mean(Regional.Node.Positive),
            surv.mean = mean(Survival.Months))

s2b = bc3 %>%
  filter(stage == 'IIB') %>%
  summarize(age.mean = mean(Age),
            grade.mean = mean(Grade),
            tumor.mean = mean(Tumor.Size),
            exam.mean = mean(Regional.Node.Examined),
            pos.mean = mean(Regional.Node.Positive),
            surv.mean = mean(Survival.Months))

s3a = bc3 %>%
  filter(stage == 'IIIA') %>%
  summarize(age.mean = mean(Age),
            grade.mean = mean(Grade),
            tumor.mean = mean(Tumor.Size),
            exam.mean = mean(Regional.Node.Examined),
            pos.mean = mean(Regional.Node.Positive),
            surv.mean = mean(Survival.Months))

s3b = bc3 %>%
  filter(stage == 'IIIB') %>%
  summarize(age.mean = mean(Age),
            grade.mean = mean(Grade),
            tumor.mean = mean(Tumor.Size),
            exam.mean = mean(Regional.Node.Examined),
            pos.mean = mean(Regional.Node.Positive),
            surv.mean = mean(Survival.Months))

s3c = bc3 %>%
  filter(stage == 'IIIC') %>%
  summarize(age.mean = mean(Age),
            grade.mean = mean(Grade),
            tumor.mean = mean(Tumor.Size),
            exam.mean = mean(Regional.Node.Examined),
            pos.mean = mean(Regional.Node.Positive),
            surv.mean = mean(Survival.Months))
```

Combine mean values to make a new dataframe.

```
bc4 = rbind.data.frame(s2a, s2b, s3a, s3b, s3c)
rownames(bc4) = c('IIA', 'IIB', 'IIIA', 'IIIB', 'IIIC')
```

Make a distance matrix.

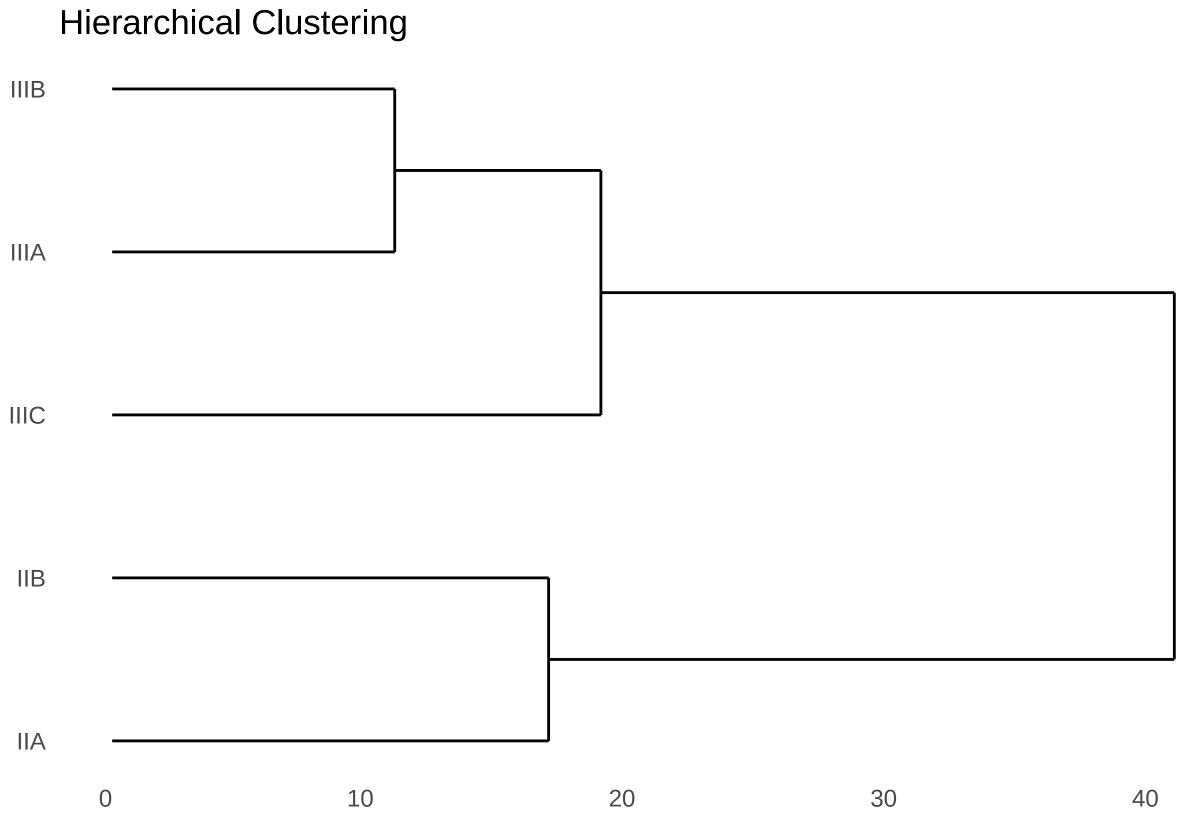
```
dist1 = dist(bc4, method = 'euclidean')
```

Perform hierarchical clustering analysis.

```
clust1 = hclust(dist1)
```

Plot with ggdendro.

```
ggdendrogram(clust1, rotate = T, size = 2) +
  labs(title = 'Hierarchical Clustering')
```



As expected, Stage III clusters together, as does Stage II.