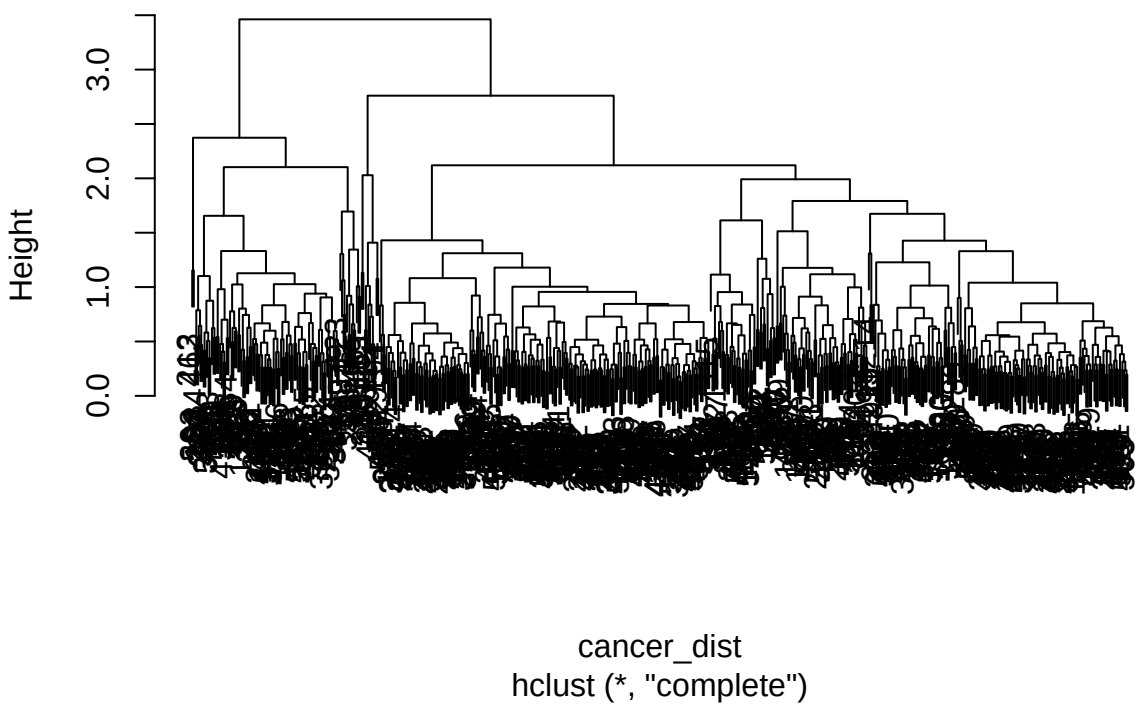
HW8_cluster.R

lukemcevoy

2021-11-22

```
# clear the environment
rm(list=ls())

# select the data
filename<-'/Users/lukemcevoy/Develop/stevens/f21/dataMining/week10/hw7/wisc_bc_ContinuousVar.csv'
cancer<-read.csv(filename)
View(cancer)
# cancer_df<-data.frame(lapply(na.omit(cancer),as.numeric))
cancer_df<-data.frame(cancer)
cancer_df<-cancer_df[-1]
cancer_df$diagnosis <- ifelse(cancer_df$diagnosis == 'M', 1, 0)
View(cancer_df)

normalized_cancer_df<-as.data.frame(apply(cancer_df[,1:ncol(cancer_df)], 2, function(x) (x-min(x))/(max
View(normalized_cancer_df)

# We want to cluster with all features BUT diagnosis, we remove here
normalized_cancer_df<-normalized_cancer_df[-1]

# split data
index<-sort(sample(nrow(normalized_cancer_df), round(.3*nrow(normalized_cancer_df))))
training<-normalized_cancer_df[-index,]
test<-normalized_cancer_df[index,]

cancer_dist<-dist(normalized_cancer_df[,-ncol(normalized_cancer_df)])
hclust_resutls<-hclust(cancer_dist)
plot(hclust_resutls)
```




```
dev.off()
```

```
## null device
##           1
```

```
hclust_2<-cutree(hclust_resutls,2)
table(hclust_2,cancer_df$diagnosis)
```

```
##
## hclust_2   0   1
##        1   0 103
##        2 357 109
```