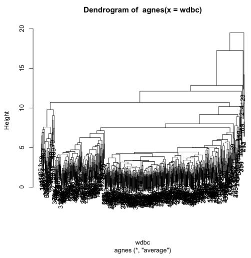
Math 5364

Data Mining 2

Homework 20

Mary Barker

1. (a) Perform agglomerative hierarchical clustering on the `wdbc` data set, and plot the dendrogram.



   (b) Cut the dendrogram to produce a clustering wtih $k = 2$ clusters.

   (c) Test whether the cluster labels obtained in this way are independent of diagnoses. How effective is this clustering at predicting diagnoses?

   The clustering is extremely ineffective at predicting diagnoses.

   A table of the cluster labels and diagnosis is shown below. Regardless of diagnosis, the majority of rows of data were clustered into one cluster.

	cluster 1	cluster 2
B	357	0
M	209	3

Source Code

```
#Data Mining hw 20
library(cluster)
library(mixtools)

source(’˜/Dropbox/Tarleton/data_mining/generic_functions/dataset_ops.R’)
wdbc <- read.csv(’˜/Dropbox/Tarleton/data_mining/dfiles/wdbc.data’,header=F,sep=’,’)
wdbc <- wdbc[,-1]
nr   <- nrow(wdbc)
nc   <- ncol(wdbc)

wdbc <- standardize(wdbc, 2:nc)
#(a) Perform agglomerative hierarchical clustering on the wdbc data set, and plot
#    the dendrogram.
wdbc.agnes = agnes(wdbc)
pltree(wdbc.agnes)

#(b) Cut the dendrogram to produce a clustering with k = 2 clusters.
wdbc.cluster = cutree(as.hclust(wdbc.agnes), k = 2)


#(c) Test whether the cluster labels obtained in this way are independent of diagnosis.
#    How effective is this clustering at predicting diagnoses?

table(wdbc$V2, wdbc.cluster)

wdbc.gauss = mvnormalmixEM(wdbc[,2:nc], k = 2)
wdbc.gauss$sigma #covariance
wdbc.gauss$mu #mean
wdbc.guass$lambda # prior probabilities
wdbc.gauss$posterior
```