HW 7

Ronald Barry

3/10/2022

Problem One

Read the data in the appendix, into R. You will examine the data in 3d (you can use the mouse to rotate the scatterplot, if you wish):

```
library(car)
library(car)
scatter3d(x1~x2+x3, data=dat, neg.res.col="white",
          pos.res.col="white", surface.alpha=0.0)
```

a. Use kmeans to determine the optimal number of clusters (how do you know how many clusters are 'optimal'?) and then plot the 3d plot. HINT: if 'tmp' is the output of kmeans, use

```
scatter3d(x1~x2+x3,point.col=tmp$cluster)
```

b. Now use a hierarchical method. Try both complete and single linkage. Do you get the same clustering in the same order for both linkages?

c. What IS complete linkage? What is single linkage?

Problem Two:

The dataset WoodySpecies.pdf is on Blackboard with this assignment. It consists of plant densities obtained from many 10 m^2 plots:

> Bruce W. Hoagland, Lisa R. Sorrels and Susan M. Glenn1 (1996) Woody Species Composition of Floodplain Forests of the Little River, McCurtain and LeFlore Counties, Oklahoma **Proc. Okla. Acad. Sci.** Vol. 76, pp. 23 - 29.

a. You should get this dataset into R, then transpose it with code similar to this:

```
dat <- read.csv(f, header=TRUE)
dat_t <- t(dat[,-1])
colnames(dat_t) <- dat[,1]
dat
```

b. Now use kmeans to determine a 'reasonable' number of clusters. You should justify your choice of cluster numbers.

c. Briefly, how does kmeans() work?

d. Now cluster the data using any hierarchical clustering method you wish. Don't forget to think about a reasonable distance measure. Do you get similar clusters to when you used kmeans?

e. Use the cut() function to cut the tree at some level and plot the part of the dendrogram above that cut.

Example code:

```
tmp <- cut(as.dendrogram(MyhclustOutput),h=wherecut)
plot(tmp$upper)
plot(tmp$lower[[1]])
```

Problem Three

Continue with the analysis from the last problem (WoodySpecies data). Here you will look at cophenetic correlation and tanglegrams.

a. What IS a cophenetic correlation? What is it used for?

b. Using the dendrogram from part (d), problem two, find the correltion between the cophenetic distance matrix and the regular distance matrix.

c. You should now compute another hierarchical clustering using a different method from what you used in the last problem and save the dendrogram. Using the tanglegram() function in dendextend package, examine whether this new dendrogram gives similar clusters to the dendrogram from the last problem.

Problem Four

a. What is a divisive hierarchical technique, as opposed to an agglomerative hierarchical technique?

b. Use code similar to the code below to run DIANA (DIvisive ANAlysis) on the Woody Species data.

```
library(cluster)
res.diana <- diana(distmat, stand = TRUE)

# Plot the dendrogram
library(factoextra)
fviz_dend(res.diana, cex = 0.5,
          k = 4, # Cut in four groups
          palette = "jco" # Color palette
          )
```

Problem Five

Somewhere, find a dataset that you are interested in that compares multiple sites, countries, etc. Choose your favorite (?) clustering method, then use it to cluster the dataset. How many clusters did you get? Why did you pick that number of clusters? Do you think your approach did a good job of clustering?

APPENDIX:

```
dat <-
structure(c(216, 99, 100, 172, 102, 126, 179, 111, 113, 210, 103, 78, 171, 105, 110, 197, 87, 109, 191, 94, 99, 2
23, 109, 91, 185, 93, 122, 177, 107, 110, 191, 104, 90, 232, 79, 86, 171, 107, 104, 207, 83, 109, 165, 114, 123,
204, 111, 90, 210, 111, 86, 160, 112, 117, 208, 90, 96, 167, 103, 117, 317, 112, 96, 261, 115, 107, 279, 108, 103
, 287, 121, 116, 330, 84, 99, 261, 104, 123, 328, 83, 85, 284, 99, 101, 298, 95, 100, 276, 119, 101, 267, 115, 94
, 324, 90, 94, 291, 95, 103, 277, 105, 116, 329, 93, 93, 321, 88, 93, 278, 104, 109, 335, 97, 101, 294, 92, 102,
308, 95, 120, 291, 95, 112, 288, 103, 101, 306, 95, 106, 291, 111, 100, 297, 100, 114, 306, 103, 106, 320, 87, 85
, 288, 103, 101, 302, 108, 91, 296, 110, 96, 400, 107, 110, 389, 110, 108, 375, 99, 104, 390, 91, 107, 375, 102,
100, 391, 87, 104, 387, 94, 83, 411, 100, 102, 353, 106, 118, 400, 104, 89), .Dim = c(60L,
3L), .Dimnames = list(NULL, c("x1", "x2", "x3")))
dat <- as.data.frame(dat)
```