Homework 5 (4570, 4880)

Due Friday May 3 by 2 pm

★ Turn in a typed hardcopy of your explanations, relevant output, and R code by the due date and time.

Flower Data. This data set contains **8** measurements on **18** flowers. Perform unsupervised learning on this data based only upon the variables `V5 soil, V6 preference, V7 height, V8 distance.`

```
V1 binary    winters - whether plant may be left in garden when it freezes
V2 binary    shadow - whether plant needs to stand in the shadow
V3 (a)binary tubers - plants with tubers and plants that grow any other way
V4 nominal   color - flower color (1=white,2=yellow,3=pink,4=red,5=blue)
V5 ordered   soil - whether plant grows in dry(1), normal(2), wet(3) soil
V6 ordered   preference - personal preference ranking from 1 to 18
V7 interval  height - plant height (cm)
V8 interval  distance - distance (cm) that should be left between plants
```

H.5.1 Perform PCA with `scale=TRUE.`

(a) Give the 4 sets of PC loadings for all 4 variables. Interpret the first two PCs based upon the magnitude and sign of the loadings using a cutoff of 0.5.

(b) Give the PCA summaries (standard deviation, PVE, cumulative PVE). Explain how many components should be retained based upon the average eigenvalue and PVE.

(c) Plot the first two PC scores with the labels. Explain whether or not there are PC1 or PC2 outliers.

H.5.2 Perform CA with the scaled data (dat.s).

(a) Fit Hierarchial CA using `method="average"`. Give the dendrogram with the gap criterion line along with the number of observations in each cluster. Explain whether any flowers may be outliers.

(b) Perform k-Means CA with k = 3 using `set.seed(2)` and `nstart=20`. Give a table of the cluster sizes and the cluster means. Interpret what the clusters represent based upon the cluster means.

(c) Fit LDA using the k-Means clusters in (b). Give the confusion matrix and error rates for each of the 3 classes. Explain which cluster is not predicted as well.

R-Code

You can read the **Flower Data** into R by downloading *flower.csv* from WyoCourses-Files/DATA to a folder called DataMining and running the following code.

```
                                          # data with all variables #
dat0 = read.csv('C:/DataMining/flower.csv')

dat = dat0[,5:8]                          # dat has V5, V6, V7, V8 #
n = nrow(dat); p = ncol(dat)              # n = 18; p = 4
labs = 1:18                               # labels #

dat.s = scale(dat)                        # scaled data #
dat.d = dist(X.s)                         # distance data #
```