Lab Assignment: Intro to Cluster Analysis

Spencer Swartz

April 25, 2017

Create a Word document from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word document via D2L Dropbox.

## Exercise 1

The depth, area, and perimeter of 31 randomly selected farm ponds in Southeastern Minnesota are recorded in the data file farmpondsize.rda.

### Part 1a

Load the data set and standardize the variables in the file (i.e. find the z-scores for each value). Store the z-scores in a new data frame.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

require(DS705data)

## Loading required package: DS705data

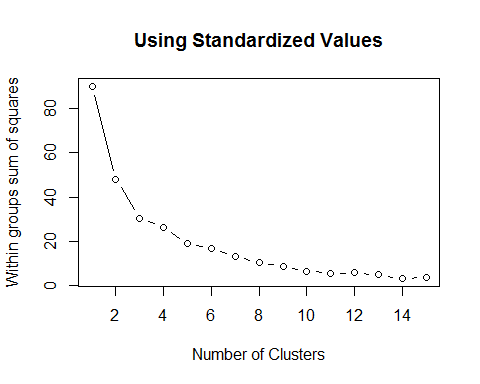
data("farmpondsize")  
farmpondsize.z <- scale(farmpondsize)

### Part 1b

For k-means clustering, plot the within sum of squares for the first 15 clusters against the cluster number and use the plot to determine a good number of clusters to partition the cases into.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

wss <- (nrow(farmpondsize.z)-1)\*sum(apply(farmpondsize.z,2,var))  
for (i in 2:15) wss[i] <- sum(kmeans(farmpondsize.z,  
 centers=i)$withinss)  
plot(1:15, wss, type="b", xlab="Number of Clusters",  
 ylab="Within groups sum of squares",  
 main="Using Standardized Values")



It looks like the knee is at 3 so we will use 3 clusters.

### Part 1c

Perform the k-means clustering on the z-scores for pond depth, area, and perimeter using the number of clusters you determined from the plot. Find the number of cases in each cluster as well as the cluster means for each variable.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

set.seed(0)  
fit <- kmeans(farmpondsize.z,centers = 3,nstart = 25)  
fit

## K-means clustering with 3 clusters of sizes 13, 6, 12  
##   
## Cluster means:  
## depth area perim  
## 1 -0.2291099 -0.82690812 -0.8532843  
## 2 1.5909969 -0.03301262 -0.1112098  
## 3 -0.5472961 0.91232344 0.9799962  
##   
## Clustering vector:  
## [1] 3 1 3 3 3 1 1 1 1 2 1 1 2 3 2 3 3 3 1 3 3 3 3 2 1 2 1 1 1 2 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 7.161449 6.555822 16.870530  
## (between\_SS / total\_SS = 66.0 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

round(aggregate(farmpondsize,by=list(fit$cluster),FUN=mean),2)

## Group.1 depth area perim  
## 1 1 10.10 0.10 123.37  
## 2 2 22.89 0.25 202.00  
## 3 3 7.86 0.44 317.62

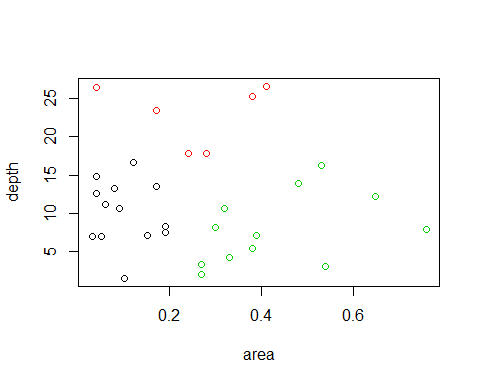
### Part 1d

Create three scatterplots, each with separate plotting colors for each cluster: plot 1: area against depth plot 2: perimeter against depth plot 3: perimeter against area.

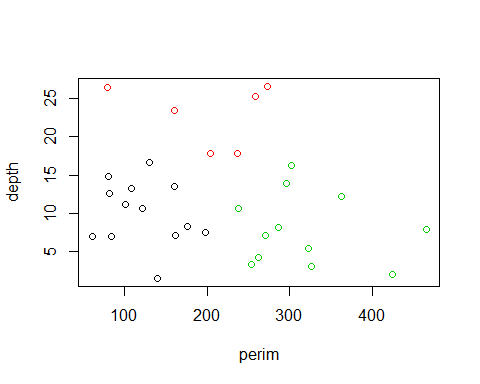
Write a brief summary of your observations about how the ponds have been clustered with regard to the pond size variables.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

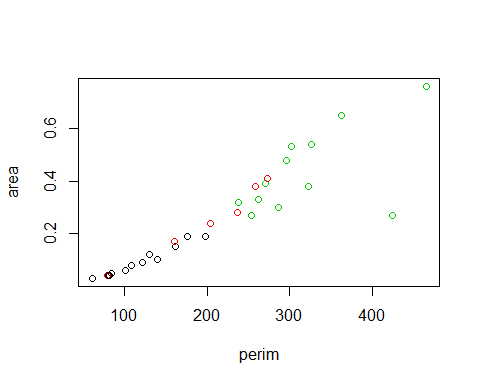
plot(farmpondsize[c("area", "depth")], col = fit$cluster)



plot(farmpondsize[c("perim", "depth")], col = fit$cluster)



plot(farmpondsize[c("perim", "area")], col = fit$cluster)



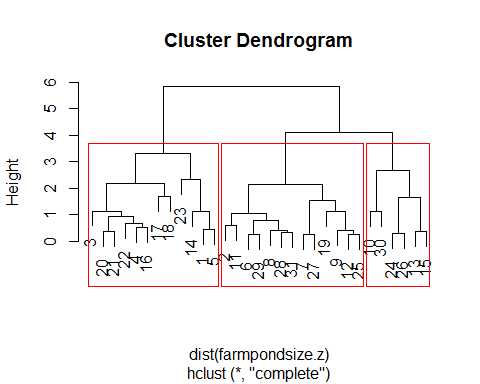
It looks as if there is a pretty strong linear relationship between area and perimeter. The black group has small depth, area and perimeter. The green has a small depth and a large perimeter and area. the red group has high large depth, medium area, and medium to low perimeter.

### Part 1e

Plot the dendrogram for hierarchical clustering using complete linkage and add the rectangles for the number of clusters you chose in part 1b above.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

output <- hclust(dist(farmpondsize.z))  
plot(output)  
rect.hclust(output,h=4)



## Exercise 2

The data set police.rda contains 15 anthropometric and physical fitness measurements for 50 white male applicants to the police department of a major metropolitan city. Note that the first column is simple the ID number of each applicant and is not a variable.

We saw in week 13 how factor analysis could be used to reduce the number of variables for this data set. Now we will see how cluster analysis will group the 50 applicants according to these 15 variables.

### Part 2a

Load the data set and scale the variables in the file by dividing each value by the standard deviation. Store the scaled values in a new data frame.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

data("police")  
police.s <- scale(police, center = FALSE,   
 scale = apply(police, 2, sd, na.rm = TRUE))

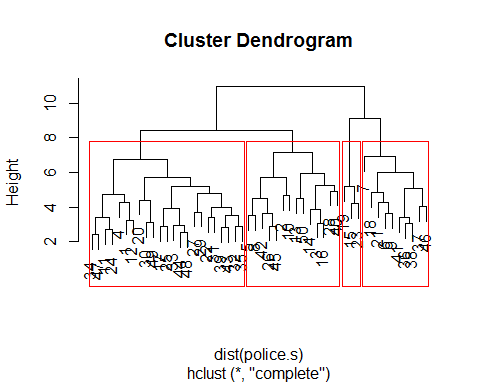
### Part 2b

Plot the dendrogram for hierarchical clustering on the scaled values using complete linkage and add the rectangles for the number of clusters given by cutting the dendrogram at a height of 8.

How many clusters does this create?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

output2 <- hclust(dist(police.s))  
plot(output2)  
rect.hclust(output2,h=8)



4 Clusters are created at this height.

### Part 2c

Append the original data frame (the unscaled one) with the cluster number from cutting the dendrogram at a height of 8. Find the number of applicants in each cluster.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

police$clusternumber <- cutree(output2, h=8)  
summary(as.factor(police$clusternumber))

## 1 2 3 4   
## 23 14 10 3

### Part 2d

Display the sample means rounded to 2 decimal places for the 15 variables for each cluster. Recall that body fat (FAT) was an important variable in the factor analysis. What can you say about the sample means for the variable FAT for each cluster?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

round(aggregate(police,by=list(police$clusternumber),FUN=mean),2)

## Group.1 ID REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE DIAST  
## 1 1 28.57 0.29 177.00 76.80 41.16 28.26 89.68 15.07 69.00 73.30  
## 2 2 21.57 0.34 173.63 67.44 39.54 26.81 85.64 12.86 81.07 76.57  
## 3 3 25.90 0.35 185.53 93.08 42.18 29.21 97.26 20.55 66.40 72.00  
## 4 4 19.00 0.32 179.40 92.13 41.87 29.33 98.90 24.83 89.33 84.00  
## CHNUP BREATH RECVR SPEED ENDUR FAT clusternumber  
## 1 6.83 193.57 113.87 5.54 4.25 11.56 1  
## 2 9.14 177.57 117.93 5.43 4.49 8.41 2  
## 3 2.40 210.10 109.70 5.65 3.71 22.97 3  
## 4 1.67 209.33 136.67 4.67 3.67 25.21 4

It looks like for the FAT variable avg for each group is a little bit different. Group 2 is the lowest (8.41), and Group 4 is the highest (25.21). group 1 and 3 round out the middle with average of 11.56 and 22.97 respectively.