Homework 7
Applied Mutlivariate Analysis

Emorie Beck

October 22, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(psych)
library(gridExtra)
library(knitr)
library(kableExtra)
library(MASS)
library(vegan)
library(smacof)
library(scatterplot3d)
library(ape)
library(ade4)
library(ecodist)
library(cluster)
library(factoextra)
library(ggdendro)
library(lme4)
library(plyr)
library(tidyverse)
```

1.2 data

Ive just obtained a zoo! Well, at least all of the animals from a zoo; the land was lost because of failure to pay property taxes and the owner decided to get out of the zoo business. Ive always wanted to own a zoo, but I dont know beans about it. My first job is to organize the creatures in some sensible way so that the staff can care for them easily and visitors can find what they want to see without too much random walking around.

The file, Set_6.csv, contains one animal per row. The columns are 20 different features that define the animals. All are self-explanatory except the last one (catsize). This is an indicator of whether the animal is smaller than a typical housecat or larger than a typical housecat. All variables are binary with 0 indicating the absence of the feature (e.g., does not produce milk) and 1 indicating the presence of the feature (e.g., has fins).

```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework8"

dat <- sprintf("%s/Set_6(1).csv", wd) %>%
```

```
  read.csv(., stringsAsFactors = F)

head(dat)

##   hair feathers eggs milk airborne aquatic predator toothed backbone
## 1    1        0    0    1        0       0        1       1        1
## 2    1        0    0    1        0       0        0       1        1
## 3    0        0    1    1        0       1        1       1        1
## 4    1        0    0    1        0       0        1       1        1
## 5    1        0    0    1        0       0        1       1        1
## 6    1        0    0    1        0       0        0       1        1
##   breathes venomous fins two_legs four_legs five_legs six_legs eight_legs
## 1        1        0    0        0         1         0        0          0
## 2        1        0    0        0         1         0        0          0
## 3        0        0    1        0         0         0        0          0
## 4        1        0    0        0         1         0        0          0
## 5        1        0    0        0         1         0        0          0
## 6        1        0    0        0         1         0        0          0
##   tail domestic catsize
## 1    0        0       1
## 2    1        0       1
## 3    1        0       0
## 4    0        0       1
## 5    1        0       1
## 6    1        0       1
```

2 Question 1

Analyze these data using each of the following hierarchical clustering procedures: single linkage, complete linkage, average linkage, centroid method, Wards method.

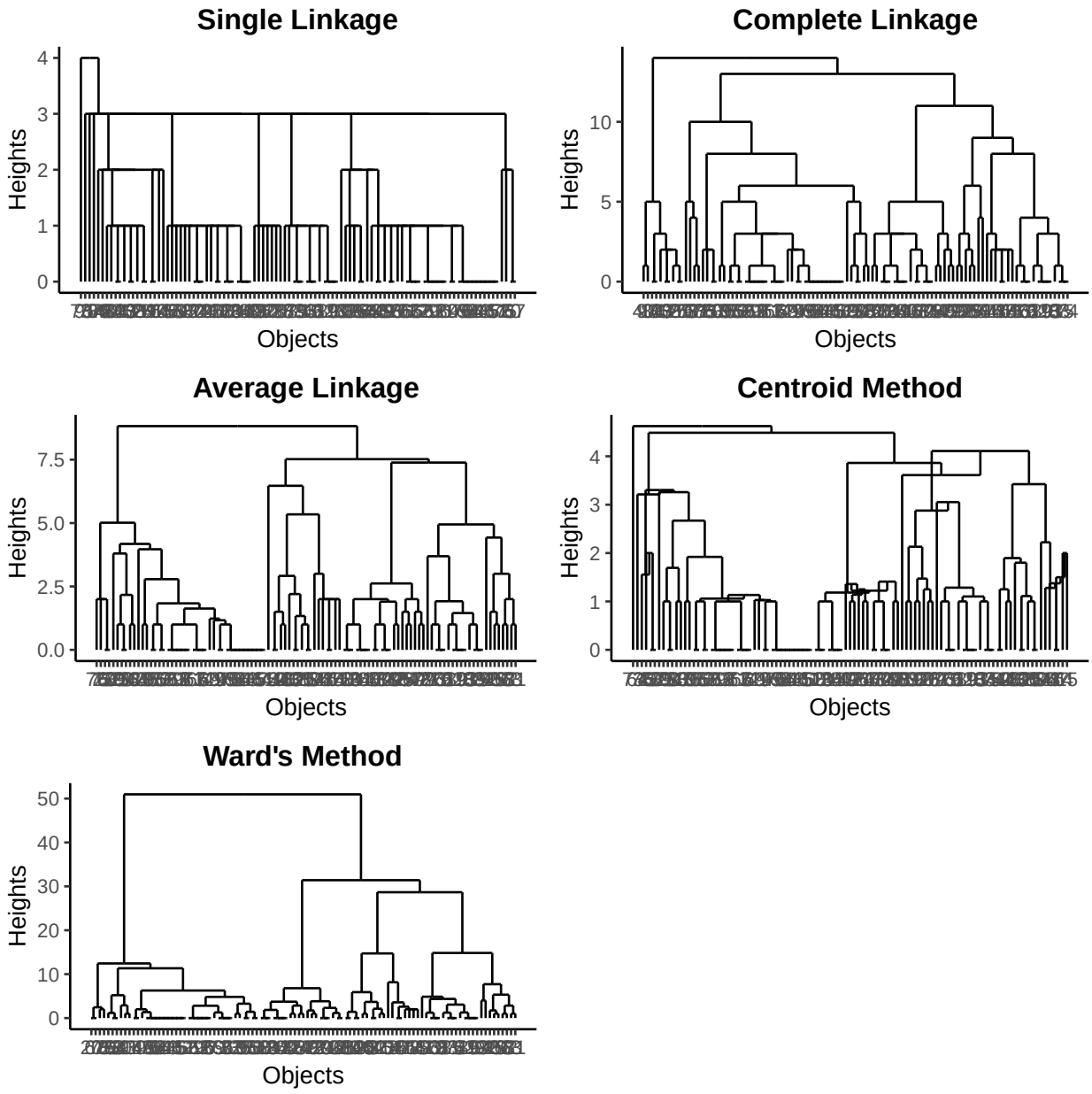
```
# create a fucntion for creating the dendrograms
plot_fun <- function(cl, m){
  ggdendrogram(cl,theme_dendro=FALSE,size=4) +
    labs(x = "Objects", y = "Heights", title = m) +
    theme_classic() +
    theme(plot.title = element_text(face = "bold", hjust = .5))
}

nested.mods <- tribble(
  ~method_arg, ~Method, ~data,
  "single", "Single Linkage", dat,
  "complete",  "Complete Linkage", dat,
  "average", "Average Linkage", dat,
  "centroid", "Centroid Method", dat,
  "ward.D2", "Ward's Method", dat
) %>%
  mutate(dist = map(data, ~(dist(.,method="euclidean"))^2),
         cl = map2(dist, method_arg, ~hclust(.x, method=.y)),
         merge = map(cl, ~data.frame(.$merge)),
         pl = map2(cl, Method, plot_fun))
```

```r
# merges
nested.mods %>% unnest(merge) %>%
  group_by(Method) %>%
  mutate(m_num = 1:n()) %>%
  gather(key = x, value = value, X1, X2) %>%
  select(-method_arg) %>%
  unite(tmp, Method, x, sep = ".") %>%
  spread(tmp, value)
```

```
## # A tibble: 100 x 11
##    m_num `Average Linkag~ `Average Linkag~ `Centroid Metho~
##    <int>            <int>            <int>            <int>
##  1     1               -1               -4               -1
##  2     2               -2               -6               -2
##  3     3              -18                2              -18
##  4     4              -23                3              -23
##  5     5              -29                4              -29
##  6     6              -56                5              -56
##  7     7               -5              -11               -5
##  8     8              -45                7              -45
##  9     9              -46                8              -46
## 10    10              -48                9              -48
## # ... with 90 more rows, and 7 more variables: `Centroid Method.X2` <int>,
## #   `Complete Linkage.X1` <int>, `Complete Linkage.X2` <int>, `Single
## #   Linkage.X1` <int>, `Single Linkage.X2` <int>, `Ward's
## #   Method.X1` <int>, `Ward's Method.X2` <int>
```

```r
do.call("grid.arrange", nested.mods$pl)
```



3 Question 2

Choose the method that you believe provides the easiest interpretation of the data. Based on that method, how many sections or major areas do I need to build in my zoo so that I can house similar animals together? Provide a justification for this conclusion.

In this case, Ward's Method appears to provide the most interpretable clusters, suggesting 3 clusters before you start seeing greater losses in accuracy.

```
new.dat <- dat
new.dat$clustnumber <- cutree((nested.mods %>% filter(method_arg == "ward.D2"))$cl[[1]], k=3)
```

4 Question 3

Based on the dominant features of the clusters you have identified, try to name the major clusters.

```
new.dat %>% gather(feature, value, -clustnumber) %>%
  filter(value == 1) %>%
  group_by(clustnumber, feature) %>%
  summarize(prop = n()/40) %>%
  spread(clustnumber, prop) %>%
  arrange(desc(`1`), desc(`2`), desc(`3`)) %>%
  kable(., "latex", booktab = T, digits = 2) %>%
  kable_styling(full_width = F) %>%
  add_footnote("Note: values represent propotions of across each of the features within a cluster")
```

feature	1	2	3
backbone	1.00	0.57	0.50
toothed	1.00	0.52	NA
breathes	1.00	0.50	0.50
milk	1.00	0.05	NA
hair	0.95	0.12	NA
tail	0.85	0.52	0.50
catsize	0.78	0.18	0.15
four_legs	0.75	0.20	NA
predator	0.52	0.65	0.22
domestic	0.20	0.05	0.08
two_legs	0.18	NA	0.50
aquatic	0.12	0.62	0.15
fins	0.10	0.32	NA
airborne	0.05	0.15	0.40
eggs	NA	0.98	0.50
six_legs	NA	0.25	NA
venomous	NA	0.20	NA
eight_legs	NA	0.05	NA
five_legs	NA	0.02	NA
feathers	NA	NA	0.50

[a] Note: values represent propotions of across each of the features within a cluster

Based on this, it appears the first cluster is mammals, the second is amphibians and water animals, and the third are birds.

5 Question 4

Just for fun. I am really going to get in trouble for one of the animals I will have in my zoo. The previous owner gave me a big discount on the animal collection provided I take this one, which was causing him all kinds of grief. Can you identify it?

```
case <- (dat %>%
  mutate(casenum = 1:n()) %>%
  filter(predator == 1 & toothed == 1 & domestic == 0 & catsize == 1 & venomous == 1))$casenum
```

Based on this, case # 87 is likely to cause all kinds of grief.