Week 2 HW - ISYE 6501

May 26, 2018

Question 4.2

The iris data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. The response values are only given to see how well a specific method performed and should not be used to build the model.

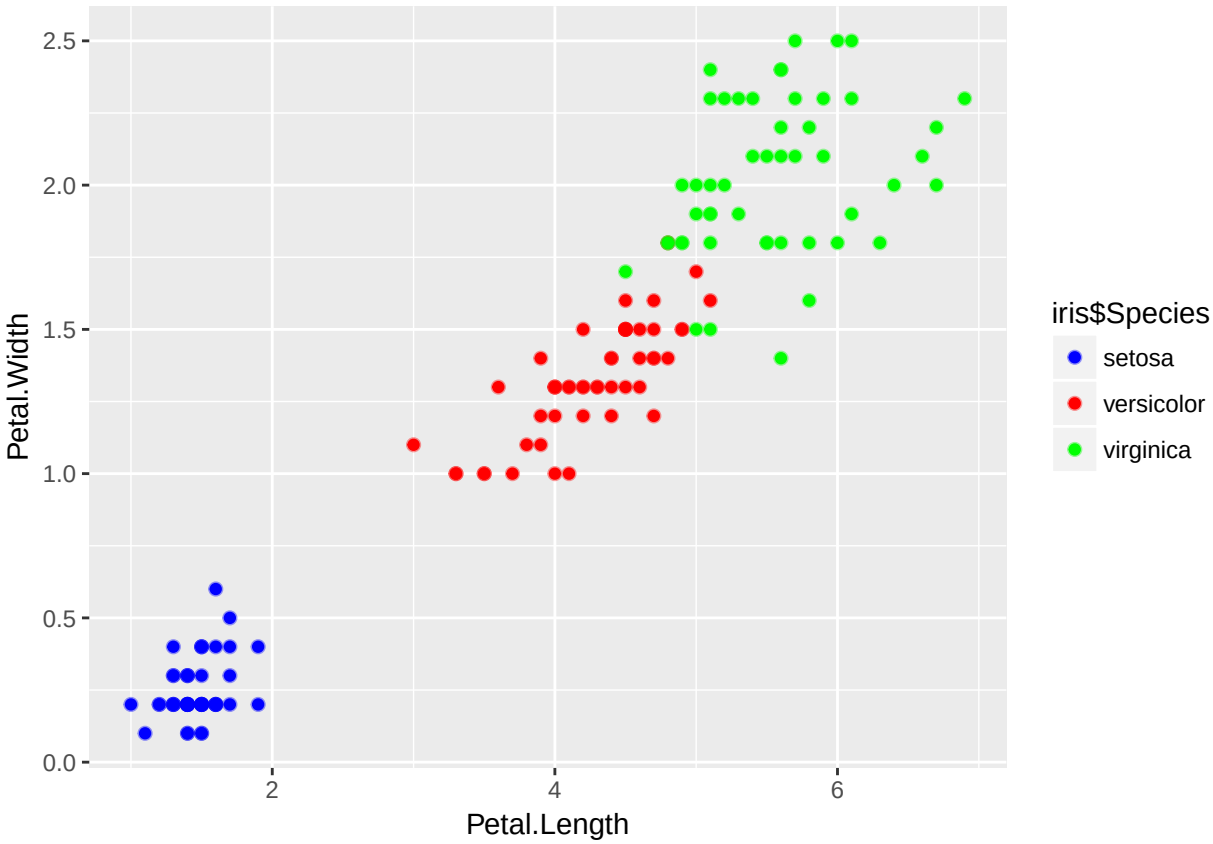
Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

Solution: Possible predictors are sepal.width, sepal.length, petal.width, and petal.length. The response is a categorical variable indicating the type of flower for each data point.

```
data(iris)

library(ggplot2)

ggplot(iris, aes(Petal.Length, Petal.Width, color = iris$Species)) +
  geom_point(alpha = 0.4, size = 2) + geom_point() +
  scale_color_manual(values = c('blue', 'red', 'green'))
```



We are not going to use the Species information to help us cluster the data, but it helps us illustrate what

the problems will be for the unsupervised methods. While the Setosa species is located a large distance away from either the Versicolor or Virginica data, the latter two are much closer together so there is going to be some error in the predictions.

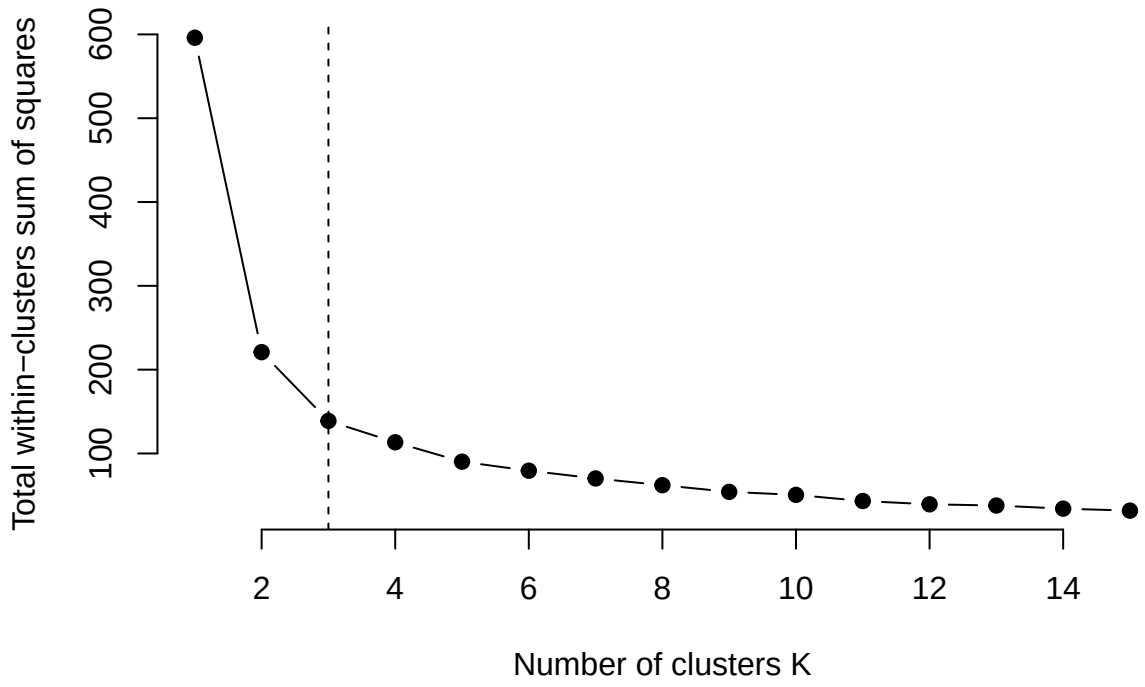
Data Prep: Scale the data:

```
# Remove species column (5) and scale the data
iris.scaled <- scale(iris[, -5])
```

Elbow method for kmeans clustering:

```
set.seed(123)
# Compute and plot wss for k = 2 to k = 15
k.max <- 15 # Maximal number of clusters
data <- iris.scaled
wss <- sapply(1:k.max,
        function(k){kmeans(data, k, nstart=10 )$tot.withinss})

plot(1:k.max, wss,
       type="b", pch = 19, frame = FALSE,
       xlab="Number of clusters K",
       ylab="Total within-clusters sum of squares")
abline(v = 3, lty =2)
```



Question 5.2

Using the crime dataset, test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

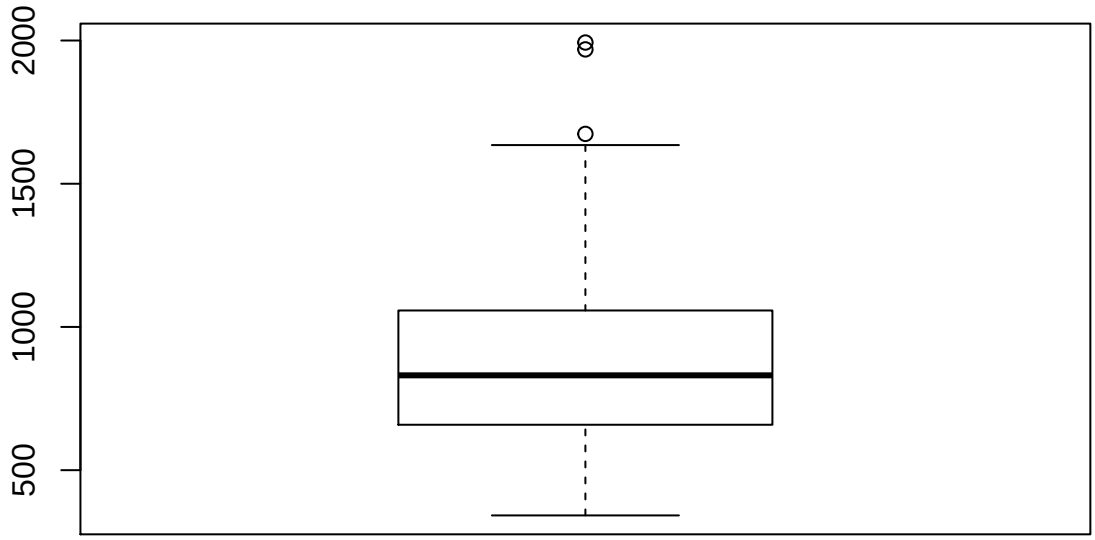
```
crime=read.csv("5.1uscrimeSummer2018.txt",stringsAsFactors = FALSE, header=TRUE, sep='\t')
```

This dataset gives state values of crime occurences for 47 states in the year 1960. We start our by plotting a boxplot of the Crime variable:

```
summary(crime$Crime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   342.0   658.5   831.0   905.1  1057.5  1993.0
```

```
#Basic plot of the Crime column.
boxplot(crime$Crime)
```




```
which.max(crime$Crime)
```

```
## [1] 26
```

```
crime[which.max(crime$Crime),]
```

```
##       M So   Ed Po1  Po2    LF   M.F Pop  NW    U1  U2 Wealth Ineq
## 26 13.1  0 12.1  16 14.3 0.631 107.1   3 7.7 0.102 4.1   6740 15.2
##        Prob    Time Crime
## 26 0.041698 22.1005  1993
```

We wonder if there are any more data points with a value of Crime greater than 1800:

```r
library(outliers)

#testing if there is one outlier in the dataset
grubbs.test(crime$Crime, type=10)
```

```
##
##  Grubbs test for one outlier
##
## data:  crime$Crime
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

This yields a p-value of 0.07887 - while small, it is not below the standard threshold of significance (0.05) and thus we do not reject the null hypothesis (which is that there are no outliers).

Question 6.2

(1) Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. You can use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too.

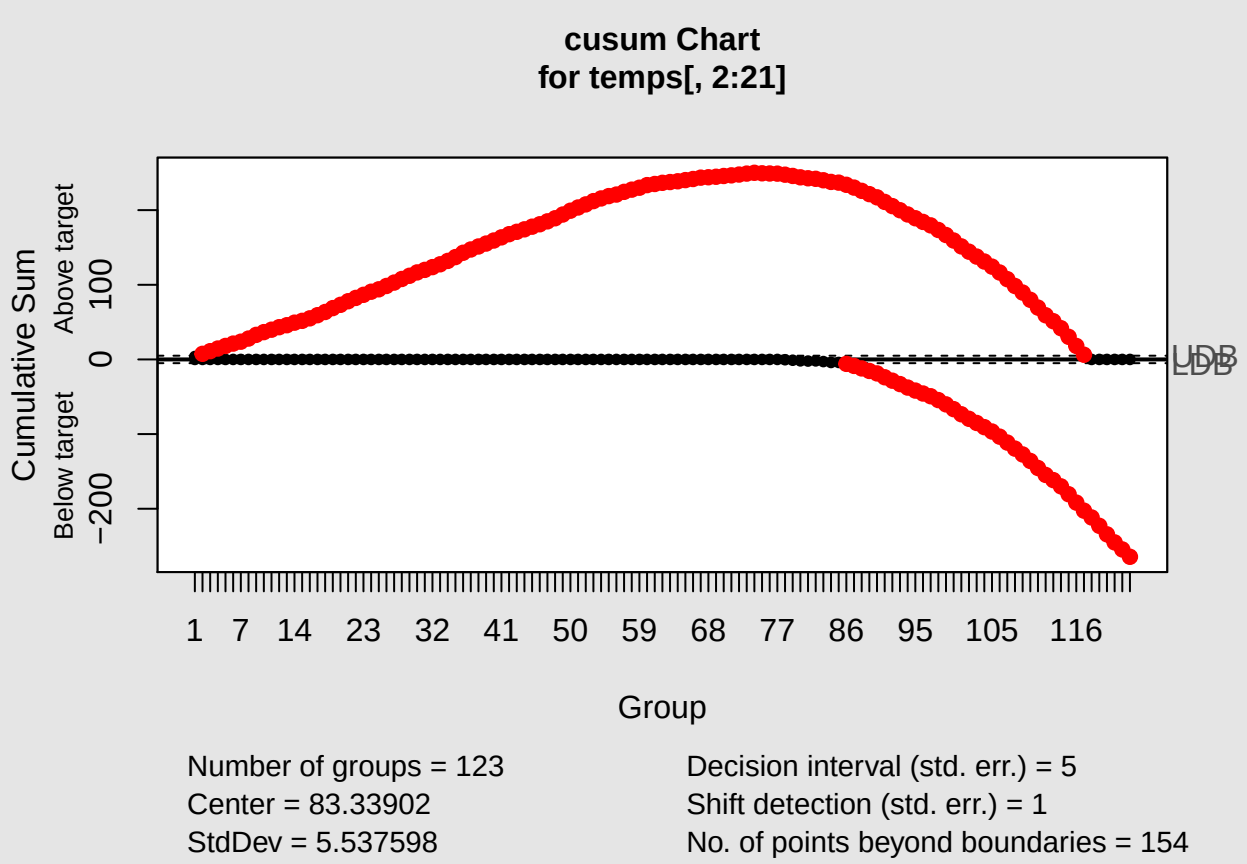
```r
temps=read.csv("6.2tempsSummer2018.txt",stringsAsFactors = FALSE, header=TRUE, sep='\t')

library(qcc)
```

```
## Package 'qcc' version 2.7
```

```
## Type 'citation("qcc")' for citing this R package in publications.
```

```r
cusum(temps[,2:21])
```




```
## List of 14
##  $ call             : language cusum(data = temps[, 2:21])
##  $ type             : chr "cusum"
##  $ data.name        : chr "temps[, 2:21]"
##  $ data             : int [1:123, 1:20] 98 97 97 90 89 93 93 91 93 93 ...
##   ..- attr(*, "dimnames")=List of 2
##  $ statistics       : Named num [1:123] 88.8 88.3 88.4 88.3 88.2 ...
##   ..- attr(*, "names")= chr [1:123] "1" "2" "3" "4" ...
##  $ sizes            : int [1:123] 20 20 20 20 20 20 20 20 20 20 ...
##  $ center           : num 83.3
##  $ std.dev          : num 5.54
##  $ pos              : num [1:123] 3.95 7.5 11.08 14.63 18.1 ...
##  $ neg              : num [1:123] 0 0 0 0 0 0 0 0 0 0 ...
##  $ head.start       : num 0
##  $ decision.interval: num 5
##  $ se.shift         : num 1
##  $ violations       :List of 2
##  - attr(*, "class")= chr "cusum.qcc"
```

(2) Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

References:

(1) Heirarchical Clustering in R https://www.r-bloggers.com/hierarchical-clustering-in-r-2/

(2) Kmeans Clustering in R https://www.r-bloggers.com/k-means-clustering-in-r/

(3) Determining the optimal number of clusters http://www.sthda.com/english/wiki/print.php?id=239