Assignment5.R

Jason

2021-04-25

```
# Assignment 5 Fundamentals of Machine Learing
# Data comes from cereals.csv

library(utils)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(class)
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
library(FNN)
```

```
##
## Attaching package: 'FNN'

## The following objects are masked from 'package:class':
##
##     knn, knn.cv
```

```
library(e1071)
library(reshape2)
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.0.5
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
WD<-setwd("C:/Users/Jason/Documents/MSBA/Fundamentals for Machine Learning/Assignment5")
Cereal<-read.csv("Cereals.csv", header = TRUE)

Cereal2 <- Cereal[complete.cases(Cereal),]

row.names(Cereal2) <- Cereal2[,1]

Cereal2 <- Cereal2[,-1:-3]

NC <- sapply(Cereal2, scale)

CD <- dist(NC, method = 'euclidean')

Hsingle <- agnes(CD, method = 'single')

Hcomp <- agnes(CD, method = 'complete')

Havg <- agnes(CD, method = 'average')

Hward <- agnes(CD, method = 'ward')

Hsingle
```

```
## Call:     agnes(x = CD, method = "single")
## Agglomerative coefficient:  0.6067859
## Order of objects:
##  [1]  1  3  4  2  5 35  6 14 18 71 41 23 28 17 10 34 12 64 46 74 47  8 72 73 30
## [26] 24 29 36  7 48 50 26 27 51 56 13 57 19 55 33 40 21 31 49 20 22 70 32 15 60
## [51] 16 59  9 25 66 58 42 61 62 63 39 45 11 65 43 44 37 67 69 52 38 68 53 54
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.3777  1.7695  1.8668  2.2787  4.0361
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method"
```

```
Hcomp
```

```
## Call:     agnes(x = CD, method = "complete")
## Agglomerative coefficient:  0.8353712
## Order of objects:
##  [1]  1  3  4  2 25 66 58 42 61 62 63 53 54  5 35 46 74 24 30 47 10 34 12  6 17
## [26] 14 18 71 28 23 41 29 64 36  8 72 73  9 31 49 32 13 57 19 33 21 40 55 11 65
## [51] 15 60 16 59 39 20 22 70 37 67 69 52  7 48 45 26 50 43 44 27 51 56 38 68
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.6076  2.3389  2.9321  3.7169 10.9839
##
```

```
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method"
```

```
Havg
```

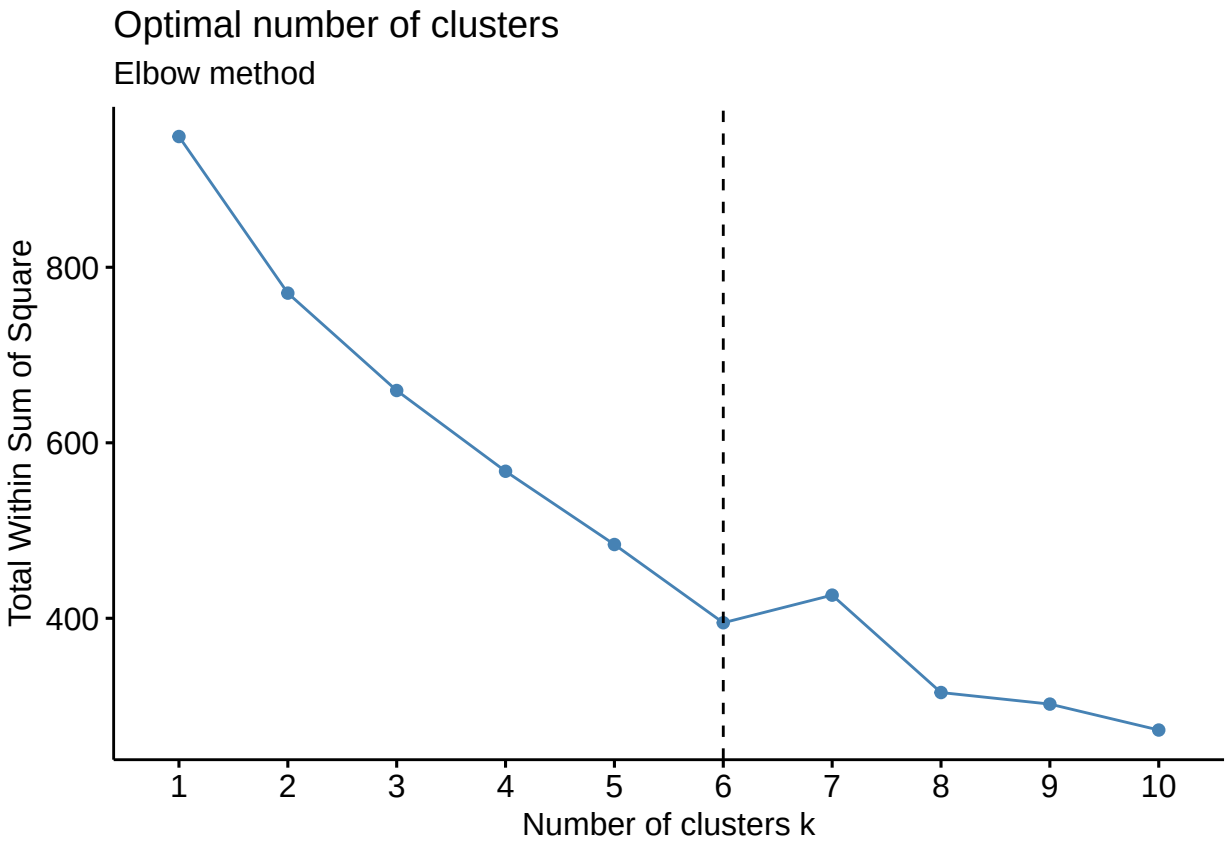
```
## Call:     agnes(x = CD, method = "average")
## Agglomerative coefficient:  0.7766075
## Order of objects:
##  [1]  1  3  4  2  5 35 46 74 24 30 47  6 17 14 18 71 23 41 28 29 64 10 34 12 36
## [26]  8 72 73  9 32 20 22 70 31 49 13 57 19 33 40 55 21 15 60 16 59 39 25 66 58
## [51] 42 61 62 63  7 48 50 45 26 27 51 56 43 44 37 67 69 52 38 68 11 65 53 54
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.4633  2.0666  2.4461  2.9445  7.7243
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method"
```

```
Hward
```

```
## Call:     agnes(x = CD, method = "ward")
## Agglomerative coefficient:  0.9046042
## Order of objects:
##  [1]  1  3  4  2 43 44 13 57 19 33 21 40 55  7 48 45 26 50 27 51 56 38 68  5 35
## [26] 46 74 24 30 47 10 34 12  6 17 29 64 14 18 71 28 23 41 36  8 72 73  9 31 49
## [51] 32 20 22 70 11 65 15 60 16 59 39 37 67 69 52 25 66 58 42 61 62 63 53 54
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.5858  2.3422  3.6092  4.1559 18.5749
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method"
```

```
# WARD clustering method is the best, as it's agglomerative coefficient is
# closest to 1

fviz_nbclust(NC, kmeans, method = "wss") +
  geom_vline(xintercept = 6, linetype = 2)+
  labs(subtitle = "Elbow method")
```




```
# the plot elbow is at k=6, therefore 6 clusters is ideal

Partition <- createDataPartition(NC[,1], p = .5, list = FALSE )
A <- NC[Partition,]
B <- NC[-Partition,]

Award <- agnes(A, method = 'ward')
Bward <- agnes(B, method = 'ward')

Aclus <- cutree(Award, k = 6)
Bclus <- cutree(Bward, k = 6)

Aclus
```

```
##  [1] 1 2 2 3 3 2 3 3 3 4 5 3 3 2 2 2 3 6 3 4 5 5 3 3 5 4 2 5 2 4 4 6 6 6 3 3 3
```

```
Bclus
```

```
##  [1] 1 2 1 2 2 3 2 2 4 2 2 5 5 2 6 3 2 2 2 5 2 2 3 3 6 3 5 6 5 5 6 6 2 4 6 2 2
```

```
# data should not be normalized for evaluating healthy cereal options
# data should be looked at to see which cereal has lowest sugar
# and highest density of vitamins per calorie
```