Project 3 notebook

Meenakshi Nagarajan

Oct 20, 2017

# Dataset – Wholesale customers data

<http://archive.ics.uci.edu/ml/datasets/Wholesale+customers>

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

#Meenakshi Nagarajan  
#nagarajan.12@wright.edu  
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

#Load the data into 'mydata'  
mydata=read.csv(file="/Users/meenakshinagarajan/Desktop/Datamining/Project3/Wholesale customers data.csv")  
head(mydata)

## Channel Region Fresh Milk Grocery Frozen Detergents\_Paper Delicassen  
## 1 2 3 12669 9656 7561 214 2674 1338  
## 2 2 3 7057 9810 9568 1762 3293 1776  
## 3 2 3 6353 8808 7684 2405 3516 7844  
## 4 1 3 13265 1196 4221 6404 507 1788  
## 5 2 3 22615 5410 7198 3915 1777 5185  
## 6 2 3 9413 8259 5126 666 1795 1451

# Background of data

The dataset used in this study is obtained from the UCI Machine learning repository. (<http://archive.ics.uci.edu/ml/datasets/Wholesale+customers>). It consists of clients of a wholesale distributor. It includes annual spending on products in monetary units (m.u.).

Dataset Characteristics: Multivariate

Attribute characteristics: integer

Date Donated: 2014/ 03/31

Number of instances: 440

Number of Attributes: 8

Missing values: None

# Attributes

FRESH: annual spending (m.u.) on fresh products (Continuous)

MILK: annual spending (m.u.) on milk products (Continuous)

GROCERY: annual spending (m.u.)on grocery products (Continuous)

FROZEN: annual spending (m.u.)on frozen products (Continuous)

DETERGENTS\_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)

DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous)

CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)

REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

# Assessing cluster tendency

#removing channel and region from data  
df <- mydata[,-1]  
df <- df[,-1]  
df <- mydata.scaled <- scale(df)  
library("factoextra")

## Loading required package: ggplot2

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

fviz\_pca\_ind(prcomp(df), title = "Wholesale Customers data",   
 habillage = mydata$Channel, palette = "jco",  
 geom = "point", ggtheme = theme\_classic(),  
 legend = "bottom")



It can be seen that this dataset contains 2 clusters.

# Evaluating cluster tendency with Hopkins statistic

library(clustertend)  
set.seed(123)  
hopkins(df, n = nrow(df)-1)

## $H  
## [1] 0.06370234

Here, H is 0.06 which is < 0.5 threshold. Therefore, we can reject the null hypothesis and conclude that dataset has a significantly clusterable data.

# Choosing the best clustering algorithm and number of clusters

library(clValid)

## Loading required package: cluster

# Compute clValid  
clmethods <- c("hierarchical","kmeans","pam")  
intern <- clValid(df, nClust = 2:6,   
 clMethods = clmethods, validation = "stability")

## Warning in clValid(df, nClust = 2:6, clMethods = clmethods, validation =  
## "stability"): rownames for data not specified, using 1:nrow(data)

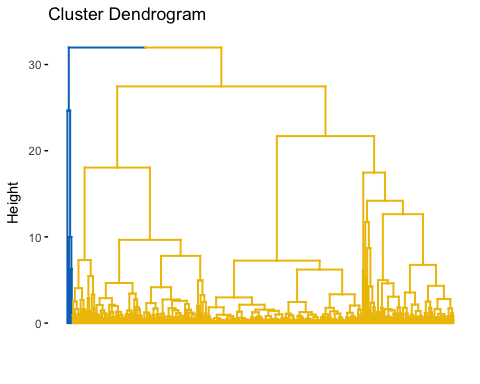
# Summary  
summary(intern)

##   
## Clustering Methods:  
## hierarchical kmeans pam   
##   
## Cluster sizes:  
## 2 3 4 5 6   
##   
## Validation Measures:  
## 2 3 4 5 6  
##   
## hierarchical APN 0.0037 0.0075 0.0083 0.0123 0.0128  
## AD 2.5094 2.4526 2.3596 2.2847 2.2680  
## ADM 0.0431 0.1990 0.1506 0.1130 0.1180  
## FOM 0.9854 0.9609 0.9475 0.8906 0.8781  
## kmeans APN 0.0272 0.0482 0.1040 0.1904 0.1833  
## AD 2.4706 2.2158 2.0491 2.0257 1.8412  
## ADM 0.5173 0.2968 0.4547 0.6367 0.5546  
## FOM 0.9307 0.8682 0.8500 0.8383 0.7909  
## pam APN 0.0620 0.1886 0.3117 0.3577 0.2571  
## AD 2.1928 2.0341 1.9962 1.9089 1.7420  
## ADM 0.1832 0.4669 0.8206 0.8465 0.6112  
## FOM 0.8895 0.8760 0.8730 0.8339 0.8167  
##   
## Optimal Scores:  
##   
## Score Method Clusters  
## APN 0.0037 hierarchical 2   
## AD 1.7420 pam 6   
## ADM 0.0431 hierarchical 2   
## FOM 0.7909 kmeans 6

It can be seen that, for APN and ADM, hierarchical clustering with 2 clusters performs the best

# Performing hierarchical clustering of data

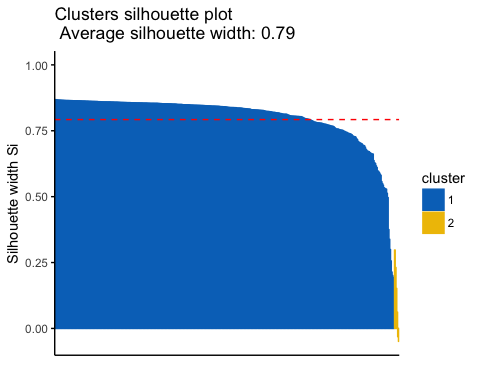
library(factoextra)  
# Hierarchical clustering using eclust() [enhanced clustering]  
hc <- eclust(df, "hclust", k = 2, hc\_metric = "euclidean",   
 hc\_method = "ward.D2", graph = FALSE)  
# Visualize dendrograms  
fviz\_dend(hc, show\_labels = FALSE,  
 palette = "jco", as.ggplot = TRUE)



# Cluster validation

fviz\_silhouette(hc, palette = "jco",   
 ggtheme = theme\_classic())

## cluster size ave.sil.width  
## 1 1 434 0.80  
## 2 2 6 0.11



It can be seen that objects of cluster 1 are well clustered compared to cluster 2.

# Determining the closer clusters

dsil <- hc$silinfo$widths[, 1:3]  
neg\_sil <- which(dsil[, 'sil\_width'] < 0)  
dsil[neg\_sil, , drop = FALSE]

## cluster neighbor sil\_width  
## 184 2 1 -0.03159430  
## 334 2 1 -0.04926217

The closer cluster to two of cluster 2 members are cluster 1.

# Agreement between channel and hierarchical clusters

library("fpc")  
# Compute cluster stats  
channel <- as.numeric(mydata$Channel)  
d\_clust\_stats <- cluster.stats(d = dist(df),   
 channel, hc$cluster)  
# Corrected Rand index  
d\_clust\_stats$corrected.rand

## [1] 0.02256532

Agreement between Channel type and cluster solution is 0.022 which is very low.

# extracting sample data

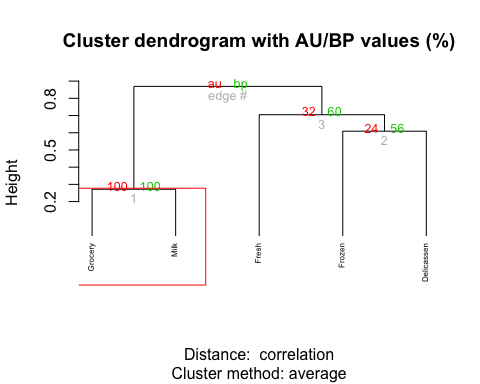
set.seed(123)  
ss <- sample(1:8, 5)  
df <- mydata[, ss]

# computing p value and extracting significant clusters

library(pvclust)  
set.seed(123)  
pv <- pvclust(df, method.dist="cor",   
 method.hclust="average", nboot = 10)

## Bootstrap (r = 0.5)... Done.  
## Bootstrap (r = 0.6)... Done.  
## Bootstrap (r = 0.7)... Done.  
## Bootstrap (r = 0.8)... Done.  
## Bootstrap (r = 0.9)... Done.  
## Bootstrap (r = 1.0)... Done.  
## Bootstrap (r = 1.1)... Done.  
## Bootstrap (r = 1.2)... Done.  
## Bootstrap (r = 1.3)... Done.  
## Bootstrap (r = 1.4)... Done.

# Default plot  
plot(pv, hang = -1, cex = 0.5)  
pvrect(pv)



#extract objects from significant clusters  
clusters <- pvpick(pv)  
clusters

## $clusters  
## $clusters[[1]]  
## [1] "Grocery" "Milk"   
##   
##   
## $edges  
## [1] 1

Milk and grocery are the two significant cluster objects in determining the channel type.