Wyniki

Started on	czwartek, 7 grudnia 2023, 10:22
State	Finished
Completed on	czwartek, 7 grudnia 2023, 11:41
Time taken	1 hour 19 mins
Grade	**49.00** out of 50.00 (**98%**)

Question 1

Complete Mark 30.00 out of 30.00

1. Download **the facebook.csv** dataset from the folder [data folder]
2. **Clustering kmeans**
    1. Remove from the dataset the first 3 variables.
    2. Find out the number of clusters in this dataset, in kmeans clustering. Write why you choose a given number of clusters.
    3. Clustering the data using the kmeans method.
    4. The observations from the largest cluster put in the data frame, describe this largest cluster.
    5. Calculate the value of the average silhouette for the whole dataset in this clustering.
3. **Hierarchical clustering**
    1. Produce a dendogram of the dataset(attach a graphic file by email) using Euclidean distance and complete.
    2. Using the tree above, split the dataset into 4 clusters.
    3. Show observations from the smallest cluster.

-

```
data <- read.csv("facebook.csv")

# 1. Remove from the dataset the first 3 variables

df <- data[-1:-3]

# 2. Find out the number of clusters in this dataset, in kmeans clustering. Write why
# you choose a given number of clusters.

df.scaled <- scale(df)
library(NbClust)
NbClust(df.scaled, distance="euclidean", min.nc=2, max.nc=10,
method="kmeans", index="all")

# The best number of clusters is 2 because NbClust is a function that is based on
# the majority rule and it returns the value 2
```

```
# 3. Clustering the data using the kmeans method.

f.km <- kmeans(df.scaled, centers = 2, nstart = 25)

f.km$cluster
f.km$centers
f.km$size
f.km$withinss

# 4. The observations from the largest cluster put in the data frame, describe
# this largest cluster.

# With f.km$size we have seen that the cluster 2 is the largest. Let's obtain
# some information

facebook2<-cbind(df,f.km$cluster)
aggregate(facebook2, by=list(cluster = f.km$cluster), mean)

# In this case, the cluster 2 corresponds to those users that are not popular in
# facebook, since the number of reactions, comments, shares, likes, loves, etc. is
# significantly lower than these average numbers from cluster 1.

# 5. Calculate the value of the average silhouette for the whole dataset in this
# clustering.

kms = silhouette(f.km$cluster,dist(df.scaled))
summary(kms)
avg.silhouette <- (0.3247719 + 0.8063032)/2
avg.silhouette

# 1. Produce a dendogram of the dataset(attach a graphic file in your answer)
# using Euclidean distance and complete.


m.dist <- dist(df)
tree.facebook<-hclust(m.dist, method="complete")
library(factoextra)
fviz_dend(tree.facebook, k=2, cex = 0.5 , main = "Facebook dataset tree - complete")
```

2. Using the tree above, split the dataset into 4 clusters.

clust.facebook <- cutree(tree.facebook,4)

3. Show observations from the smallest cluster.

Question 2

Complete Mark 19.00 out of 20.00

Download the dataset contact-lense1.csv, from the folder rules

For this dataset

1. Draw a histogram of the frequency of items in the dataset.
2. Display and give the dimensions of the transaction matrix.
3. Find association rules with a support of 0.01 and a confidence value of 0.8.
    1. give the number of rules found.
    2. give(write) the number of rules of length 5.
    3. give an interpretation of the rule with the highest support value.
4. Give the information that results, from the fact that the contact.lenses=hard.

library(arules)

1. Draw a histogram of the frequency of items in the dataset.

d.tr<-read.transactions(file = "contact-lenses1.csv", format = "basket",
sep = ",", header = T,
rm.duplicates = FALSE,
quote = "", skip = 0,
encoding = "unknown")

itemFrequencyPlot(d.tr, support = 0.1, main = "Histogram of Frequency of Items")

2. Display and give the dimensions of the transaction matrix.

inspect(d.tr)
dimension <- dim(d.tr)
dimension

```
# 3. Find association rules with a support of 0.01 and a confidence value of 0.8.
# give the number of rules found.

rules <- apriori(d.tr, parameter = list(supp = 0.01, conf = 0.8))
length(rules)

# give(write) the number of rules of length 5.

rules2 <- apriori(d.tr, parameter = list(supp = 0.01, conf = 0.8, minlen = 5, maxlen = 5))
length(rules2)

# give an interpretation of the rule with the highest support value.

rules.sort <- sort (rules , by="support", decreasing=TRUE)
inspect(rules.sort[1])

# The result is the following rule:
# lhs rhs support confidence coverage lift count
# [1] {reduced} => {none} 0.5 1 0.5 1.6 12

# The meaning of this rule is that in 50% of the total transactions, we have
# reduced tears production and none contact-lenses at the same time. It also
# shows that with confidence 1 (in 100% of cases), if the person has reduced
# tears production, then he or she will not have contact lenses


# 4. Give the information that results, from the fact that the contact.lenses=hard.

# In this case, we are searching for rules in which contact.lenses=hard is a lhs

rules3 <- apriori(d.tr,
parameter = list(support=0.1, confidence=0.3),
appearance = list(default="rhs",lhs = c("hard")))
inspect(rules[1:5])

# Some information that results from hard contact lenses is reduced tears prod,
# myope, young, etc
```