INTERMEDIATE ANALYTICS

ALY6015, FALL 2019

MODULE 4 ASSIGNMENT

DATA MINING

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

CRN: 71933

SUBMITTED TO: PROF. LI, TENGLONG

DATE: 03/12/2019

Introduction

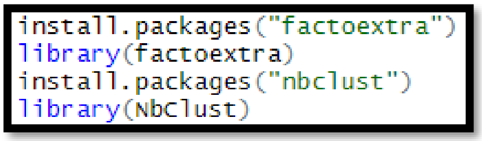
We have worked on the “Titanic” dataset to perform K-means clustering and decision trees using R.

Analysis

Part 1

Loading of packages

We have loaded and installed the factoextra and NbClust packages, using the following functions:

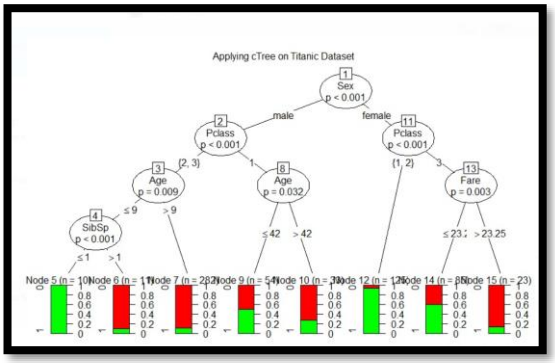


```
install.packages("factoextra")
library(factoextra)
install.packages("nbclust")
library(NbClust)
```

Part 2

Decision Tree

- Decision tree is a tool which uses decision and their consequences to show the chance of an event occurring
- We have used the “Titanic” dataset and used the ctree() and predict() functions to build the decision tree



Observations: We can see that the 9th node shows a better decision in predicting the survival rates. Moreover, we get to know that, the female passengers in the first and second class survived. Whereas, the male passengers in the second, third class and aged could not survive.

Fig.1: Decision Tree

Part 3

K- means clustering

- The algorithm divides the dataset into clusters based on the Euclidean distance between the clusters by comparing the mean values.
- As seen in the Fig.2, it can be noted that the Elbow Method is a very efficient method in plotting the optimal clusters, used for interpretation in cluster analysis.
- We will be considering 12 as the optimal number of clusters, as the graph shows a drastic change at that point.
- As mentioned in Fig.3,We have used the ggplot() function to plot the clusters.

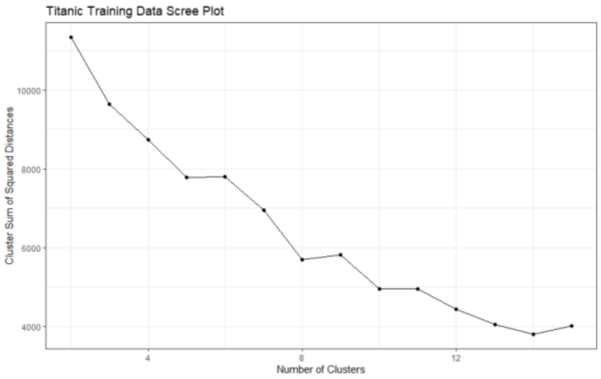
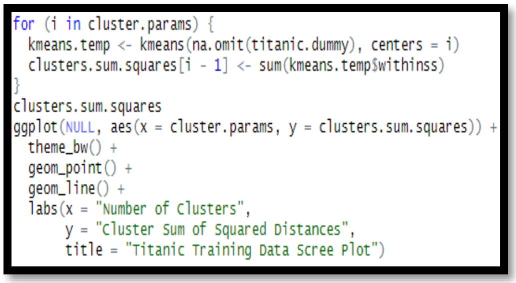


Fig.2: Elbow Graph



```
for (i in cluster.params) {
  kmeans.temp <- kmeans(na.omit(titanic.dummy), centers = i)
  clusters.sum.squares[i - 1] <- sum(kmeans.temp$withinss)
}
clusters.sum.squares
ggplot(NULL, aes(x = cluster.params, y = clusters.sum.squares)) +
  theme_bw() +
  geom_point() +
  geom_line() +
  labs(x = "Number of Clusters",
       y = "Cluster Sum of Squared Distances",
       title = "Titanic Training Data Scree Plot")
```

Fig.3: K-Means Clustering Code

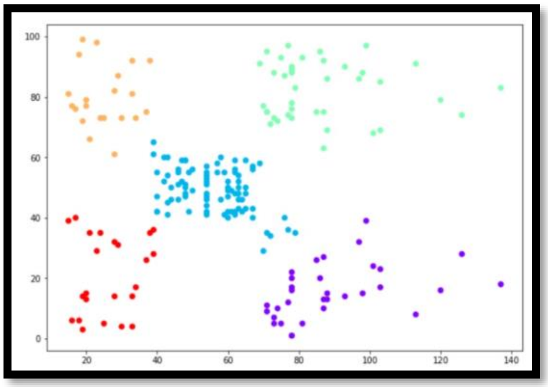


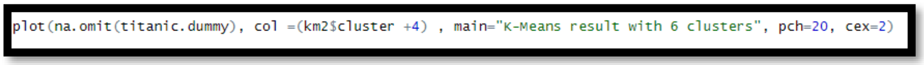
Fig.4: K-means clustering

Observations: The optimal number of clusters that can be formed are 6 clusters. Using the Euclidean distance, and comparing the least difference between the clusters, the clusters have been formed.

Part 4

Density Based Cluster

- Density based clusters show the densities of the sparse and dense areas on the graph.
- We have used the below function to plot the densities of the dataset:



```
plot(na.omit(titanic.dummy), col =(km2$cluster +4) , main="K-Means result with 6 clusters", pch=20, cex=2)
```

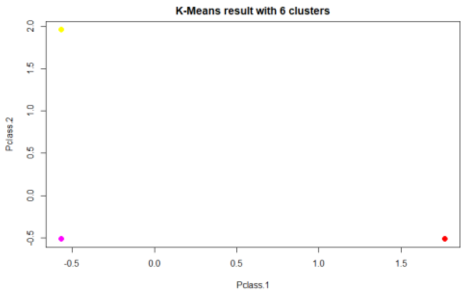


Fig.5: Density Cluster

References

- Rpubs.com. (2019). *RPubs - Linear Model Selection and Regularization Using Boston Housing Data*. [online] Available at: https://rpubs.com/Swidle/368819 [Accessed 19 Nov. 2019].
- https://cran.r-project.org/web/packages/biglasso/vignettes/bigla