Godavari college of Engineering , Jalgaon

Subject Name: Machine Learning

Practical No: 04 **Date:**

Title: Clustering Algorithm and Evaluation in R

Aim: Study and implementation of Clustering Algorithm and Evaluation in R

Theory:

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k groups $G = \{G_1, G_2, \ldots, G_k\}$ so as to minimize the within-cluster sum of squares (WCSS) defined as follows −

$$\operatorname{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2$$

The later formula shows the objective function that is minimized in order to find the optimal prototypes in k-means clustering. The intuition of the formula is that we would like to find groups that are different with each other and each member of each group should be similar with the other members of each cluster.

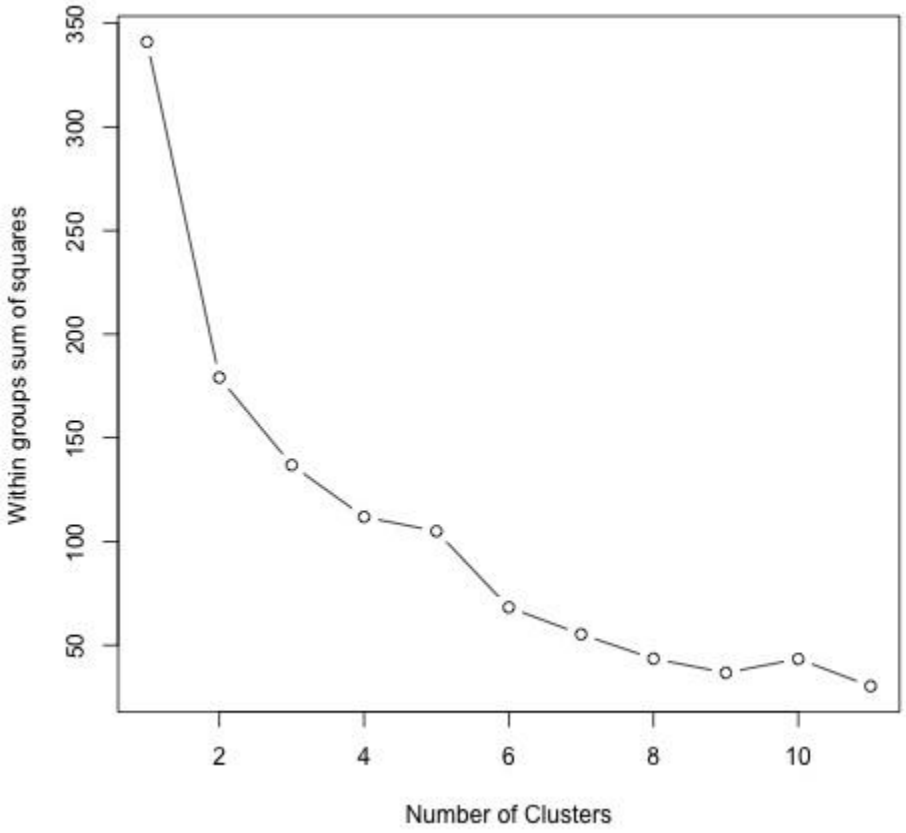
The following example demonstrates how to run the k-means clustering algorithm in R.

Input Data:

```
library(ggplot2)
# Prepare Data
data=mtcars

# We need to scale the data to have zero mean and unit variance
data<- scale(data)

# Determine number of clusters
wss<-(nrow(data)-1)*sum(apply(data,2,var))
for(i in2:dim(data)[2]){
wss[i]<- sum(kmeans(data, centers = i)$withinss)
}

# Plot the clusters
plot(1:dim(data)[2],wss, type ="b",xlab="Number of Clusters",
ylab="Within groups sum of squares")
```

In order to find a good value for K, we can plot the within groups sum of squares for different values of K. This metric normally decreases as more groups are added, we would like to find a point where the decrease in the within groups sum of squares starts decreasing slowly. In the plot, this value is best represented by K = 6.

Output:



Now that the value of K has been defined, it is needed to run the algorithm with that value.

Input data:

```
# K-Means Cluster Analysis
fit<- kmeans(data, 5) # 5 cluster solution

# get cluster means
aggregate(data,by = list(fit$cluster),FUN = mean)

# append cluster assignment
data<- data.frame(data, fit$cluster)
```

Output data:

```
Group.1        mpg         cyl       disp         hp        drat
wt
1       1  0.1082193 -0.58493208 -0.4486701 -0.6496905 -0.04967936
-0.02346989
2       2 -0.0565170  0.08165718 -0.1977204  0.4980222  0.63297355
-0.36767056
3       3  1.3739630 -1.22485777 -1.1370289 -0.9643131  1.03241235
-1.21514154
4       4 -1.2395370  1.01488215  1.4981449  1.1859553 -0.70427805
1.50088285
5       5 -0.5483556  1.01488215  0.6850701  0.3400164 -1.02222599
0.48169844
        qsec         vs         am       gear       carb
fit.cluster
1  1.1854841  1.1160357 -0.8141431 -0.1573201 -0.4145882
3.142857
2 -1.1483763 -0.8680278  1.1899014  1.3271364  1.1479487
4.333333
3  0.4763722  1.1160357  1.1899014  0.6171790 -0.8568156
2.000000
4 -0.5342923 -0.8680278 -0.8141431 -0.9318192  0.7352031
1.000000
5 -0.2958964 -0.8680278 -0.8141431 -0.9318192 -0.2376972
4.000000
  fit.cluster.1 fit.cluster.2 fit.cluster.3
1             4             2             5
2             5             1             3
3             1             3             4
4             3             4             2
5             2             5             1
```

Conclusion:

In this practical we Study and implementation of K-means Clustering Algorithm and Evaluation in R.