Alok Bhawankar
PDog

Page No.
Date
YOUVA

# BI Assignment 3

* Aim:- To apply clustering analysis to examine the patterns in given dataset. Using R programming.

* Theory:
Brief theory on k-means algorithm

Clustering is the most common exploratory data analysis technique used to get an intuition about the structure of data. It covers under unsupervised learning algorithm.

K-means algorithm is an iterative algorithm that tries to partition the dataset into k predefined non-overlapping subgroups i.e clusters where each data point belongs to only one group. It tries to make the intra-cluster points as similar as possible while also keeping the cluster as different as possible. It assigns data points to a cluster such that sum of the squared distance between the data points and the cluster centroid is minimum.

Page No.:
Date:

- Steps
1) Specify number of clusters K.
2) Initialise centroid by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3) Keep iterating until there is no change to the centroids i.e. assignment of data points to cluster isn't changing

-the approach followed by k-means is Expectation Minimization.

★ Input : Dataset contains 620 high school students grade of three subjects areas English, maths & Science.

★ output : Group 620 high schod students based on their grades.

★ conclusion : Hence, learned the k-means clustering using R and determine the value of K in k-means algorithm.

★ FAQ

1. compare and contrast any two clustering algorithms.

=>

| K-means clustering | Hierarchical clustering |
|---|---|
| 1. It has prefined numbers of clusters and the method assigns records based on distance | 1. It can be stopped at any point if found appropriate by interpreting the dendogram. |
| 2. Start is random and hence results vary if run multiple times. | 2. Results are reproducible |
| 3. computation intensive | 3. Less computation intensive. |
| 4. Simple division of data | 4. Tree-based structure division. |

Q2. How to determine the optimal values of k in k-means algorithm?

=) Elbow method is used to determine the optimal value of k in k-means
It runs k-means clustering on the dataset for a range of values for k and then for each value of k compute an average score for all clusters. The distortion score is computed, the sum of square distances from each point to its assigned centre.
Optimal 'k' is the point at the elbow i.e. the point after which the distortion/inertia starts decreasing in a linear fashion.