What is clustering? Explain in detail about k-means clustering with an Example?

clustering:- "cluster analysis or clustering is the task of grouping a set of objects or data points in such a way that objects in the same group are more similar to Each other than to those in other groups

(Or)

clustering is the process of grouping objects into different groups which are meaningful, useful or both"

K-Means:-

A prototype based one level partitional clustering technique that attempts to find a user- specified number of cluster(k)

It is the oldest and most widely used clustering algorithm.

k-means defines a prototype as a 'centroid'

Applied to objects in a Continuous n-dimensional space

The Basic k-means Algorithm:-

choose k initial Centroids, k-user specified and k indicates number of clusters desired

Each point is then assined to the closest centroid.

centroid of each cluster is then updated based on the assigned to the cluster

Scanned with CamScanner

Algorithm:-

Algorithm Basic k-means algorithm

1. select k points as initial. Centroids
2. repeat
3. Form k clusters by assigning each point to its closest control.
4. Recompute the centroid of each cluster
5. Untill centroids do not change.

Assigning points to the closest centroid

To assign a point into a cluster, a measure must be calculated for finding the closest of point to a centroid.

Various Measures are:-

Euclidean L2

Cosine similar. is more appropriate for documents

Manhattan (L1)

Jaccard measure

symbol	Description
x	An object
Ci	The ith cluster
ci	The centroid of cluster Ci
c	The centroid of all points
mi	The number of objects in the ith cluster
m	The number of objects in the dataset.
k	The number of clusters.

Scanned with CamScanner

Centroids and obiective Functions;

Minimize the squared distance of each point to its closest centroid

Data in Euclidean space

Obiective function. measures the quality of a clustering calculate the Error of each datapoint.

SSE is formally defined as follows:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2$$

Where dist is the standard Euclidean (L_2) distance between two obiects in Euclidean space.

The centroid that minimizes.

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

To illustrate, the centroid of a cluster containing the three two-dimensional points (4,1), (2,3) and (6,2)

Document data:

k means is also used for document data

Document data is represented as a document-term matrix

$$\text{Total Cohesion} = \sum_{i=1}^{k} \sum_{x \in C_i} Cosione(x, c_i)$$

Scanned with CamScanner

Scanned with CamScanner

Choosing initial Centroids! Randomly selected initial centroids may be poor. One technique that is Commonly used to address the Problem of choosing initial Centroids is to perform multiple runs, each with a different set of randomly. chosen initial Centroids. and then select the set of clusters with the minimum SSE

Scanned with CamScanner

Scanned with CamScanner