Non-hierarchical clustering.

K-means clustering

when you have large amount of dataset heirachical clustering is not useful (Aglomerative and division)

when we have large amount of data we will use partition clustering (Specifying the cluster how many we want) one of the way of partition clustering is K-means clustering.

partition = K = 3

K = no. of clustering in dataset
means = method is to identify the cluster

Note :- always confirm K values from domain expert.

PAGE NO.
DATE / /

process of K-means clustering

step 1 :- set K values

for example I have a dataset

A[= [1, 2, 4, 7, 7, 5]]

A	B
1	1
2	1
4	5
7	7
5	7

-> So I have dataset which has two variable A and B which have 5 rows and select K value as 2

K = 2

step 2 : partition of dataset

we can do partition into three ways

1. user-specified initial partition
2. user-specified initial centroids
3. random partition

* 1 2 is depend of user input

Note :- mostly we will use random partition.

Steps:

How random partion happens.
Firstly it will check the k values and give centroids that here no. k = no. centroids but how it give centroid position?

Solution

Let say we have co-ordinates of data points like (1,1) (2,1) (4,5) (5,7)(7,7) now random partion will happen Let say cluster

A = (1,1) (2,1) (4,5)
B = (5,7) (7,7)

and we will take out the centroid position of cluster through this formula.

$$X = \frac{X_1 + X_2 + X_3 + X_n}{\text{no. coordinates}}$$

$$Y = \frac{Y_1 + Y_2 + Y_3 + Y_n}{\text{no. co-ordinates.}}$$

(X, Y) will be centroid position.

PAGE No.

DATE / /

Let compute. for A cluster

$$x = \frac{(1+2+4)}{3} = 3.33$$

$$y = \frac{(1+1+5)}{3} = 2.33$$

(3.33, 2.33) will be my centroid position.

For B cluster.

$$x = \frac{(7+5)}{2} = \frac{12}{2} = 6$$

$$y = \frac{(7+7)}{2} = \frac{14}{2} = 7$$

(6, 7) will be my centroid position

step 4

Now compute Euclidean distance of each record from each centroid and reassigns to closet cluster.

A

$\sqrt{(1-2.33)^2 + (1-2.33)^2} = 1.89$

Similarly we got

A	B
1.89	7.81
1.37	7.21
3.14	2.93
6.60	1
5.37	

So whoever in A cluster is higher values than B cluster will move toward B cluster and B cluster value will move to A cluster value so in this example 3.14 is less than 2.93 so we will move

So we have move the datapoint into A cluster and find the New centroid values for each cluster. and this cluster goes on until we dont reach perfect values which dont overlap each other

How to select K values.

Step 1:- get the mean values of each cluster.

Step 2:- take the average distance in cluster by giving different K values.

Now plot elbow chart

Step 3:- whoever get the least values of Average distance for example 6 cluster

Step 4:- Now take out the mead point 6 cluster is 3. the 3 values will be our best cluster.

elbow chart.

because it not rising too much. No decline or flatten too much.