10 Oct 2023

K-Means Clustering

clustering :- it is unsupervised machine learning techniques in which we combine similar type of data into the group.

* K-mean clustering is very easy to visualize in 2D ad 3D but in the higher dimensions (large number of columns) then it is harder to visualize it. we only perform it
* for the higher dimension it is very effective
* the steps in K-mean clustering:

(1.) Decide n-clusters

↓

(2.) Initialize Centroids

↓

(3.) Assign cluster

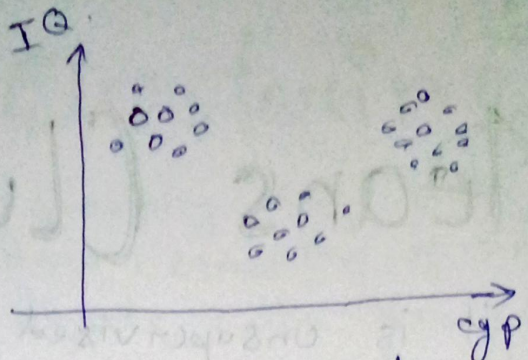
↑ ↓

(4.) Move Centroids

↓

(5.) Finish

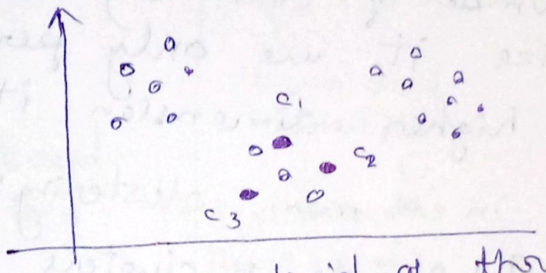
इस algorithm में programmer को इस algorithm को बताना पड़ता है कि हमें कितने clusters करने है।



Step-1 let us assum in the above graph there is have to make a three cluster.

$k = 3$

Step-2 initialize centroid
algorith take any three points in the graph and assume it as the centroides



c_1, c_2, c_3 is centroid of three cluster.

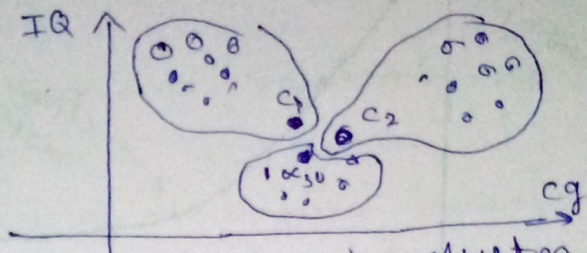
Step-3 Assign clusters.
use the euclidean distan. for each point on the graph calculat the distance. from each centroid. which one is minimum them that point is member of that cluster when centroid lies.
exampl

P_1 distance from $c_1 = 20$
P_1 distance from $c_2 = 10$
P_1 distance from $c_3 = 22$

Hence. P_1 is the part of C_2 cluster.

the above approach is apply for each. point

Step-4 Move centroid.



according to above graph. cluster is formed. Now for moving centroid for each cluster. calculate the intersection point of mean of cgpa and mean of IQ for each cluster

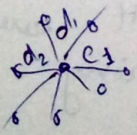
Step-5 Finish.

if the location of each centroid in previous and present step is same then stop it else continue from step-3

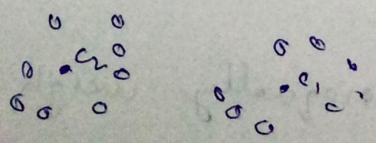
How to decide the number of cluster:

for finding the number of cluster we use elbow method. in which we plot a graph that is called elbow curve and has two axis number of clusters and inertia (within cluster sum of square distance (WCSS)).

C_1 = centroid.

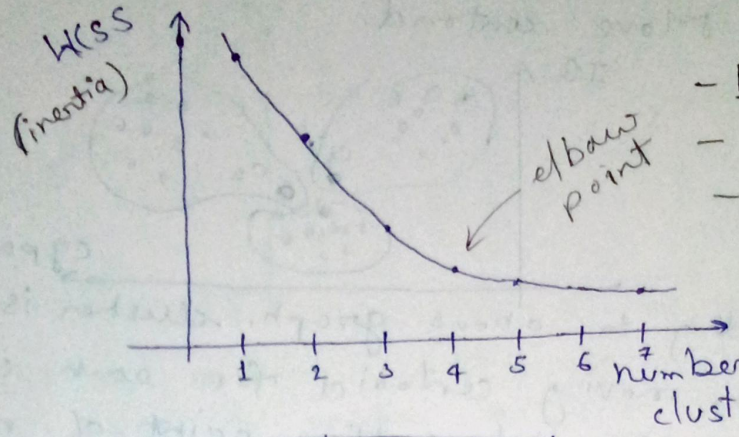


$WCSS = d_1^2 + d_2^2 + d_3^2 + \dots d_n^2$



C_1, C_2 = centroid.

$WCSS = WCSS_1 + WCSS_2 + \dots WCSS_n$



methods to find 'K'
- Elbow method
- Silhouette Score
- Gap statistic

no. of cluster	WCSS
1	$WCSS_1$
2	$WCSS_2$
3	$WCSS_3$

according to above table $WCSS_1 > WCSS_2 > WCSS_3$

for finding number of cluster we want to select ~~the~~ detect the elbow point in elbow curve.

elbow point = a point in which curve is parallel to x-axis.

Some Drawback of k-mean clustering:
- it is difficult to determine the optimal K for a given dataset
- Different initial partitions can result in different final clusters
- It doesn't work well with clusters of different size and density.
- it is ~~sent~~ sensitive to scaling and requires careful pre-processing.
- it does not produce the same result every time
- it is only for spherical ~~data~~ clustering
- it requires manual work
- it clusters everything.
- euclidean distance can unequally weigh the factors.

- it gives the local optima of the squared error function.
- choosing the centroids randomly can't give fruitful results.
- As the number of dimensions increases, a distance based similarity measure converges to a constant value between any given examples.

Question.

Solve a numerical problem on k mean clustering ~~The problem~~ problem has given 15 points. we are also given the information that we need to make 3 clusters. it means k = 3. we will solve this numerical on k-means clustering ~~using~~

first randomly choose 3 centroids from the given points. let us consider. A2(2,6), A7(5,10), A15(6,11) as the centroids of the initial clusters.

Centroid-1 = (2,6) is associated with c_1

Centroid-2 = (5,10) is associated with c_2

Centroid-3 = (6,11) is associated with c_3

Now we will find the euclidean distance between each point and the centroids. Based on the minimum distance of each point from the centroids. we will assign the points to a cluster.

point	1st centroid (2,6)	2nd centroid (5,10)	3rd centroid (6,11)	
(2,10)	4	3	4.13	C2
(2,6)	0	5	6.4	C1
(11,11)	10.29	6.08	5	C3
(6,9)	5	1.41	2	C2
(6,4)	4.47	6.08	7	C1
(1,2)	4.12	8.94	10.2	C1
(5,10)	5	0	1.41	C2
(4,9)	3.6	1.41	2.8	C2
(10,12)	10	5.38	4.1	C3
(7,5)	5.09	5.38	6.08	C1
(9,11)	8.60	4.12	3	C3
(4,6)	2	4.12	5.38	C1
(8,10)	4.12	2	3.16	C2
(3,8)	2.23	2.82	4.24	C1
(6,11)	6.4	1.41	0	C3

calculate new centroid

$$C_1(x) = \left(\frac{2+6+1+7+4+3}{6}, \frac{6+4+2+5+6+8}{6}\right)$$

$$= (3.83, 5.16)$$

$$C_2(x) = \left(\frac{2+6+5+4+3}{5}, \frac{10+9+10+9+10}{5}\right)$$

$$= (4, 9.6)$$

$$C_3(x) = \left(\frac{11+10+9+6}{4}, \frac{11+12+11+11}{4}\right)$$

$$= (9, 11.25)$$

Now that we have calculated new centroid for each cluster, we will calculate the distance of each data point from the new centroids. Then we will assign the points to clusters based on their distance from the centroids.

point	1st centroid (3.83, 5.16)	2nd centroid (4, 9.6)	3rd centroid (9, 11.25)	
(2, 10)	5.169	2.040	7.111	c_2
(2, 6)	2.013	4.118	8.750	c_1
(11, 11)	9.291	7.139	2.016	c_3
(6, 9)	4.403	2.088	3.750	c_2
(6, 4)	2.461	5.946	7.846	c_1
(1, 2)	4.249	8.171	12.230	c_1
(5, 10)	4.972	1.077	4.191	c_2
(4, 9)	3.837	0.600	5.483	c_2
(10, 12)	9.204	6.462	1.250	c_3
(7, 5)	3.171	5.492	6.562	c_1
(9, 11)	7.792	5.192	0.250	c_3
(4, 6)	0.850	3.600	7.25	c_1
(3, 10)	4.904	1.077	6.129	c_2
(3, 8)	2.950	1.887	6.824	c_2
(6, 11)	6.223	2.441	3.010	c_2

Now calculate new centroid.

$C_1(X) = (4, 9.6)$

$C_2(X) = (4.143, 9.571)$

$C_3(X) = (10, 11.333)$

the above iteration is continue till our present centroids and previous centroids are similar.

point	1st centroid (4.46)	2nd centroid (4.143, 9.571)	3rd centroid (10, 11.33)	Assigned cluster
(2, 10)	5.7	2.1	8.1	C_2
(2, 6)	2.4	4.1	9.6	C_1
(11, 11)	9.4	7.0	1.0	C_3
(6, 9)	4.8	1.9	4.6	C_2
(6, 4)	2.0	5.8	8.3	C_1
(1, 2)	3.9	8.1	12.8	C_1
(5, 10)	5.4	0.9	5.1	C_2
(4, 9)	4.4	0.5	6.4	C_2
(10, 12)	9.5	6.3	0.6	C_3
(7, 5)	3.0	5.3	7.6	C_1
(4, 6)	1.4	3.5	8.0	C_1
(3, 10)	5.4	1.2	7.1	C_2
(3, 8)	3.5	1.9	7.7	C_2
(6, 11)	6.7	2.3	4.6	C_2
(9, 11)	8.1	5.0	1.0	C_3

Calculate new centroid:

$C_1(X) = (4, 4.6)$

$C_2(X) = (4.143, 9.571)$

$C_3(X) = (10, 11.33)$

Hence here we see that the previous and present centroid is same hence we stop there our iteration.

- point lies in cluster-1:
  (2,6), (6,4), (1,2), (7,5), (4,6)

points lies in cluster-2:
  (2,10), (6,9), (5,10), (7,9), (3,8), (6,11), (3,10)

points lies in cluster-3:
  (11,11), (10,12), (9,11)

Applications of k-mean clustering:
- Document classification
- Customer segmentation
- Cyber profiling
- Image segmentation
- Fraud detection in banking and insurance.

Advantages of k-means clustering Algo:
- easy to implement
- Scalability
- Convergence
- Generalization
- Choice of centroids

Disadvantages of k-means clustering Algo:
- Deciding the number of clusters
- choice of initial centroids
- effects of outliers
- curse of dimensionality