Introduction to Data Analytics

ITE 5201

Lecture10-K Mean Clustering

Instructor: Parisa Pouladzadeh
Email: parisa.pouladzadeh@humber.ca

K-means clustering

A process of organizing objects into groups.

A cluster is a collection of objects where these objects are similar and dissimilar to the other cluster.

K-means clustering

K-mean clustering is unsupervised clustering algorithm where you know how many clusters are appropriate.

An unsupervised algorithm which is using for quickly predicting groups from an unlabeled dataset.

The main goal of this algorithm to find groups in data and the number of groups is represented by K.

Predictions are based on the number of centroids present(K) and nearest mean values, given an Euclidean distance measurement between observations.

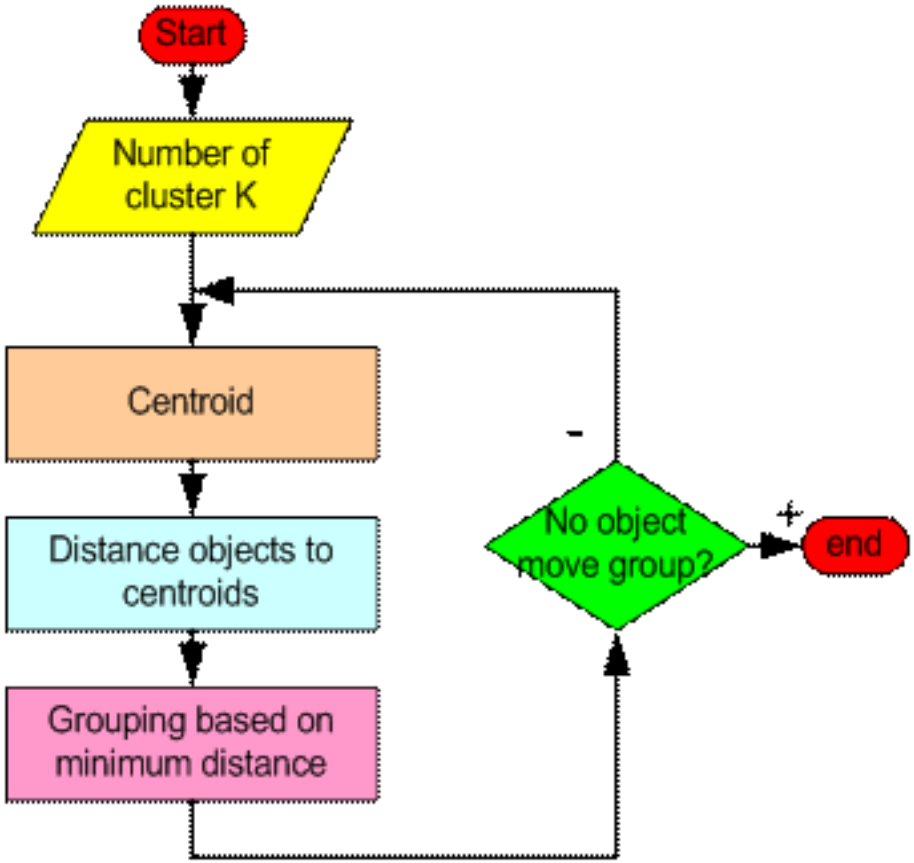
K-means

- K-Means algorithm starts with initial estimates of K centroids, which are randomly selected from the dataset.
- The algorithm iterates between two steps assigning data points and updating Centroids.
- The data point is assigned to its nearest centroid based on the squared Euclidean distance.
- Let us assume a Cluster with c as centroid and a data point x is assigned to this cluster, based on the distance between c,x.

K-mean algorithm

- The algorithm:
- K points are placed into the object data space representing the initial group of centroids.
- Each object or data point is assigned into the closest k.
- After all objects are assigned, the positions of the k centroids are recalculated.
- Steps 2 and 3 are repeated until the positions of the centroids no longer move.

How the K-Mean Clustering algorithm works?



K-mean algorithm

- **Step 1:** Begin with a decision on the value of k = number of clusters .

- **Step 2**: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:

- 1.Take the first k training sample as single-element clusters
- 2. Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment, recompute the centroid of the gaining cluster.

K-mean algorithm

Step 3: Take each sample in sequence and compute its distance from the centroid of each of the clusters.

If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Step 4 . Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

K-mean clustering example

- 1. Consider 4 data points A,B,C,D as below:

	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

- 2. Choose two centroids AB and CD, calculated as
- AB = Average of A, B
- CD = Average of C,D

	X1	X2
AB	4	2
CD	2	1

3. Calculate squared euclidean distance between all data points to the centroids AB, CD. For example distance between A(2,3) and AB (4,2) can be given by s = $(2–4)^2 + (3–2)^2$.

	A	B	C	D
AB	5	5	9	5
CD	4	16	2	2

4. If we observe in the fig, the highlighted distance between (A, CD) is 4 and is less compared to (AB, A) which is 5. Since point A is close to the CD we can move A to CD cluster.

5. There are two clusters formed so far, let recompute the centroids

- i.e, B, ACD similar to step 2.
- ACD = Average of A, C, D
- B = B

	X1	X2
B	6	1
ACD	2	1.67

- 6. As we know K-Means is iterative procedure now we have to calculate the distance of all points (A, B, C, D) to new centroids (B, ACD) similar to step 3.

	A	B	C	D
B	20	0	26	10
ACD	1.78	16.44	1.11	3.78

7. In the above picture, we can see respective cluster values are minimum that A is too far from cluster B and near to cluster ACD. All data points are assigned to clusters (B, ACD) based on their minimum distance. The iterative procedure ends here.

8. To conclude, we have started with two centroids and end up with two clusters, K=2.