K-Means Problem

By – Prof. Ansar Sheikh

K-Means Algorithm for Clustering

- kMeans algorithm is an unsupervised learning algorithm
- Given a data set of items, with certain features, and values for these features, the algorithm will categorize the items into k groups or clusters of similarity.
- To calculate the similarity, we can use the Euclidean distance,
 Manhattan distance, Hamming distance, Cosine distance as measurement.

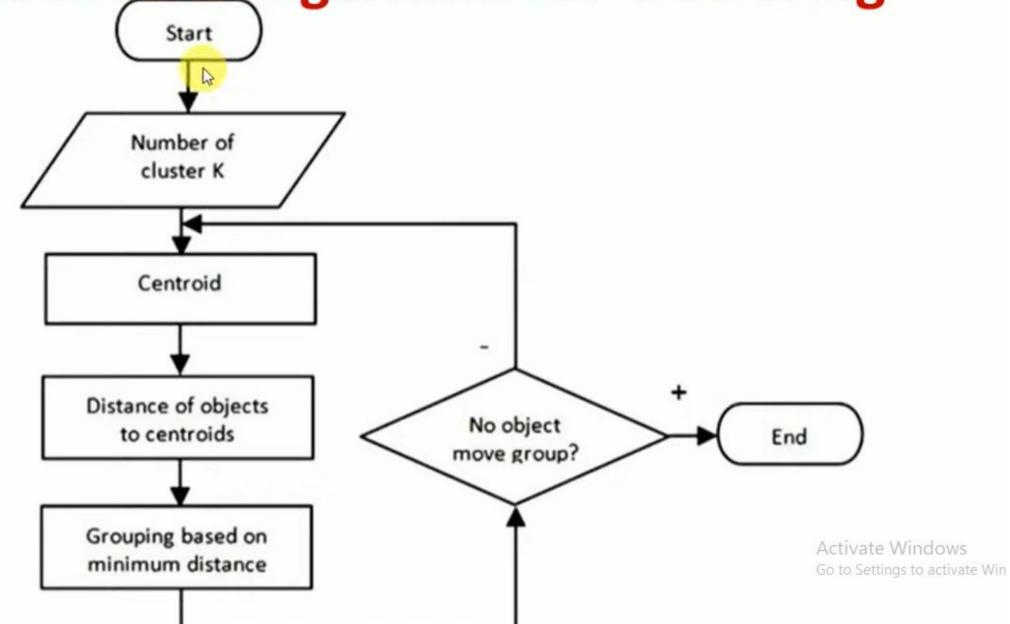
K-Means Algorithm for Clustering

Here is the pseudocode for implementing a K-means algorithm.

Input: Algorithm K-Means (K number of clusters, D list of data points)

- Choose K number of random data points as initial centroids (cluster centers).
- 2. Repeat till cluster centers stabilize:
 - a. Allocate each point in D to the nearest of Kth centroids.
 - b. Compute centroid for the cluster using all points in the cluster.

K-Means Algorithm for Clustering



Advantages and Disadvantages of K-Means Algorithm

Advantages of K-Means Algorithm

- 1. K-means algorithm is simple, easy to understand, and easy to implement.
- It is also efficient, in which the time taken to cluster K-means rises linearly with the number of data points.
- 3. No other clustering algorithm performs better than K-means.

Disadvantages of K-Means Algorithm

- The user needs to specify an initial value of K.
- The process of finding the clusters may not converge.
- 3. It is not suitable for discovering clusters that are not hyper ellipsoids or hyper spheres).

K-Means Clustering

```
A1(2, 10), A2(2, 5),
A3(8, 4), B1(5, 8),
 B2(7, 5), B3(6, 4),
 C1(1, 2), C2(4, 9)
```

Solved Example

- Suppose that the data mining task is to cluster points into three clusters,
- · where the points are
- A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).
- The distance function is Euclidean distance.
- Suppose initially we assign A1, B1, and C1 as the center of each cluster,

Da	ata Poir	nts		Dista	nce to		Cluster	New
	ta Foii	11.5					Cluster	Cluster
A1	2	10						
A2	2	5						
А3	8	4						
B1	5 _k	8						
B2	7	5						
В3	6	4						
C1	1	2						
C2	4	9						

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

Da	ata Poir	atc.			Distar	nce to			Cluster	New
De	ita Poli	11.5	2	10	5	8	1	2	Cluster	Cluster
A1	2	10								
A2	2	5								
А3	8	4								
B1	5	8								
B2	7	5								
В3	6	4								
C1	1	2								
C2	4	9								

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

	ta Dair	·+c			Distar	nce to			Cluston	New
Da	ita Poir	its	2	10	5	8	1	2	Cluster	Cluster
A1	2	10				2.0				
A2	2	5								
А3	8	4								
B1	5	8								
B2	7	5								
В3	6	4								
C1	1	2								
C2	4	9								

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

Da	ita Poir	atc.			Distar	nce to			Cluster	New
De	ita Poli	11.5	2	10	5	8	1	2	Ciustei	Cluster
• A1	2	10	0.0	00				-		
A2	2	5	5.0	00						
А3	8	4	8.4	49						
B1	5	8	3.	61						
B2	7	5	7.0	07						
В3	6	4	7.3	7.21						
C1	1	2	8.0	06						
C2	4	9	2.:	24						

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

	ata Poir	at c		1 0	Distar	nce to			Cluster	New
Da	ita Poli	its	2	10	5	8	1	2	Cluster	Cluster
→ A1	2	10	0.	00	3.	61	8.0	06		
A2	2	5	5.	00	4.:	24	3.3	16		
А3	8	4	8.	49	5.00		7.2	28		
B1	5	8	3.	61	0.0	00	7.2	21		
B2	7	5	7.	07	3.	61	6.71			
В3	6	4	7.	21	4.:	12	5.39			
C1	1	2	8.06		7.21		0.00			
C2	4	9	2.	24	1.4	41	7.62			

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

Da	to Doi:	2+0			Dista	nce to			Clustor	New
Da	ata Poir	its	2	10	5	8	1	2	Cluster	Cluster
A1	2	10	0.	00	3.	61	8.	06	1	
A2	2	5	5.	00	4.	24	3.	16	3	
A3	8	4	8.	8.49		5.00		28	2	
B1	5	8	3.	61	0.00		7.	21	2	
B2	7	5	7.	07	3.	61	6.	71	2	
В3	6	4	7.	21	4.	4.12		39	2	
C1	1	2	8.06		7.	21	0.	00	3	
C2	4	9	2.	24	1.	41	7.	62	2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

C1: (1, 2)

New Centroids:

A1: (2, 10) ~

B1: (6, 6) —

	to Doin	.+.			Distar	nce to			Cluston	New
Da	ita Poir	its	2	10	5	8	1	2	Cluster	Cluster
A1	2	10	0.0	00	3.0	61	8.	06	1	
A2	2	5	5.0	00	4.2	24	3.	16	3	
A3	8	4	8.4	49	5.0	00	7.	28	2	
B1	5	8	3.	61	0.0	00	7.	21	2	
B2	7	5	7.0	07	3.0	51	6.	71	2	
В3	6	4	7.:	21	4.:	12	5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.:	24	1.4	41	7.62		2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Current Centroids:

A1: (2, 10)

B1: (6, 6)

					··	-				
	ata Poir	ntc			Distar	nce to			Cluster	New
: D	ata Fuii	ILS	2	10	6	6	1.5	1.5	Cluster	Cluster
A1	2	10							1	
A2	2	5							3	
A3	8	4							2	
B1	5	8							2	
B2	7	5							2	
В3	6	4							2	
C1	1	2							3	
C2	4	9							2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids

A1: (2, 10)

B1: (6, 6)

C1: (1.5, 3.5)

						Dista	nce to			GL .	New
5:	Da	ita Poir	nts	2	10	6	6	1.5	1.5	Cluster	Cluster
	A1	2	10	0.	00	5.	66	6.	52	1	
	A2	2	5	5.	00	4.	12	1.5	58	3	
	А3	8	4	8.	49	2.	83	6.5	52	2	
	B1	5	8	3.	61	2.24		5.	70	2	
	B2	7	5	7.	07	1.41		5.	70	2	
	В3	6	4	7.	21	2.00		4.53		2	
	C1	1	2	8.06		6.40		1.58		3	
	C2	4	9	2.	24	3.	61	6.04		2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Go to Settings to activate Windows.

Current Centroids:

A1: (2, 10)

B1: (6, 6)

C1: (1.5, 3.5)

	ta Dair	.+.			Distar	nce to			Cluston	New
: Da	ita Poir	its	2	10	6	6	1.5	1.5	Cluster	Cluster
A1	2	10	0.0	00	5.0	66	6.	52	1	1
A2	2	5	5.	00	4.:	12	1.	58	3	3
А3	8	4	8.	49	2.83		6.	52	2	2
B1	5	8	3.	51	2.3	24	5.	70	2	2
B2	7	5	7.0	07	1.4	1.41		5.70		2
В3	6	4	7.:	21	2.0	2.00		4.53		2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.:	24	3.0	3.61 6.04		04	2	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Go to Settings to activate Windows.

Current Centroids:

A1: (2, 10)

B1: (6, 6)

C1: (1.5, 3.5)

New Centroids:

A1: (3, 9.5) -

B1: (6.5, 5.25)

	Da	ita Poir	atc			Distar	nce to			Cluster	New
-	Da	ta Poli	its	2	10	6	6	1.5	1.5	Cluster	Cluster
F	41	2	10	0.	00	5.	66	6.	52	1	1
F	42	2	5	5.	00	4.	12	1.	1.58		3
F	43	8	4	8.	49	2.	83	6.	52	2	2
E	B1	5	8	3.	61	2.	24	5.	70	2	2
E	B2	7	5	7.	07	1.	41	5.	70	2	2
E	В3	6	4	7.	21	2.	2.00		53	2	2
(C1	1	2	8.06		6.	6.40		1.58		3
(C2	4	9	2.	24	3.	61	6.	04	2	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Current Centroids:

A1: (3, 9.5)

B1: (6.5, 5.25)

Da	ita Poir	nts		Dista	nce to		Cluster	New Cluster
A1	2	10					1	Cluster
A2	2	5					3	
А3	8	4					2	
B1	5	8					2	
B2	7	5					2	
В3	6	4					2	
C1	1	2					3	
C2	4	9					1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Window

Current Centroids:

A1: (3, 9.5)

B1: (6.5, 5.25)

C1: (1.5, 3.5)

Data Baints				Cluston	New					
: Da	Data Points			9.5	6.5 5.25		1.5	3.5	Cluster	Cluster
A1	2	10	1.	1.12		6.54		6.52		
A2	2	5	4.	4.61		51	1.58		3	
А3	8	4	7.	7.43		1.95		6.52		
B1	5	8	2.	2.50		13	5.70		2	
B2	7	5	6.	6.02		56	5.	5.70		
В3	6	4	6.	6.26		1.35		4.53		
C1	1	2	7.	7.76		6.39		1.58		
C2	4	9	1.	1.12		4.51		6.04		

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

rger to rigerryser that are at longer a

Current Centroids:

A1: (3, 9.5)

B1: (6.5, 5.25)

_											
	Da	Nata Daints				Cluston	New				
:	Data Points			3	9.5	9.5 6.5 5.25 1.5 3		3.5	Cluster	Cluster	
	A1	2	10	1.12		6.54		6.52		1	1
	A2	2	5	4.61		4.51		1.58		3	3
	А3	8	4 ·	7.43		1.95		6.52		2	2
	B1	5	8	2.50		3.13		5.70		2	(1)
	B2	7	5	6.0	02	0.	56	6 5.70		2	2
	ВЗ	6	4	6.	6.26		1.35		4.53		2
	C1	1	2	7.76		6.39		1.58		3	3
	C2	4	9	1.	12	4.51		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids:

A1: (3, 9.5)

B1: (6.5, 5.25)

C1: (1.5, 3.5)

New Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

Data Points				Cluster	New					
: Da	ita Poli	11.5	3	3 9.5		5.25	1.5 3.5		Cluster	Cluster
A1	2	10	1.	1.12		6.54		6.52		1
A2	2	5	4.61		4.	51 1.58		58	3	3
А3	8	4	7.43		1.95		6.52		2	2
B1	5	8	2.	2.50		3.13		5.70		1
B2	7	5	6.	6.02		0.56		5.70		2
В3	6	4	6.	6.26		1.35		4.53		2
C1	1	2	7.76		6.39		1.58		3	3
C2	4	9	1.	1.12		4.51		6.04		1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

Data Daints				Chuston	New					
Da	Data Points			9	7	4.33	1.5	·3.5	Cluster	Cluster
A1	2	10							1	
A2	2	5							3	
А3	8	4							2	
B1	5	8							1	
B2	7	5							2	
В3	6	4							2	
C1	1	2							3	
C2	4	9							1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

Data Points		*			Cluston	New				
Da	Data Points			9	7 4.33		1.5	3.5	Cluster	Cluster
A1	2	10	1.94		7.56		6.52		1	
A2	2	5	4.33		5.04		1.58		3	
А3	8	4	6.62		1.05		6.52		2	
B1	5	8	1.67		4.18		5.70		1	
B2	7	5	5.21		0.67		5.70		2	
В3	6	4	5.52		1.05		4.53		2	
C1	1	2	7.4	7.49		6.44		1.58		
C2	4	9	0.33		5.55		6.04		1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
Activate Windows

Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

Data Baints				Cluston	New					
: Da	Data Points			9	7 4.33		1.5	3.5	Cluster	Cluster
A1	2	10	1.9	1.94		7.56		6.52		1
A2	2	5	4.3	4.33		5.04		1.58		3
А3	8	4	6.0	6.62		1.05		6.52		2
B1	5	8	1.0	1.67		4.18		5.70		1
B2	7	5	5.2	5.21		0.67		5.70		2
В3	6	4	5.5	5.52		1.05		4.53		2
C1	1	2	7.4	7.49		6.44		1.58		3
C2	4	9	0.3	0.33		5.55		6.04		.1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Thank