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)

1. 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.

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.
2. 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.

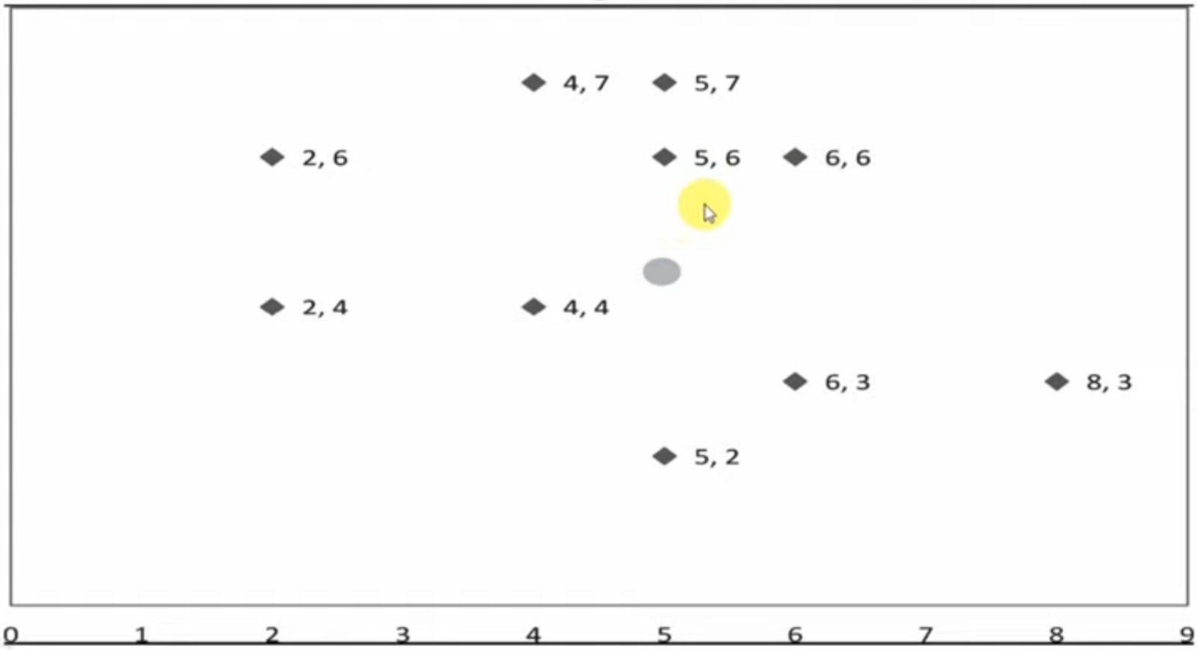
Disadvantages of K-Means Algorithm

1. The user needs to specify an initial value of K.
2. 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).

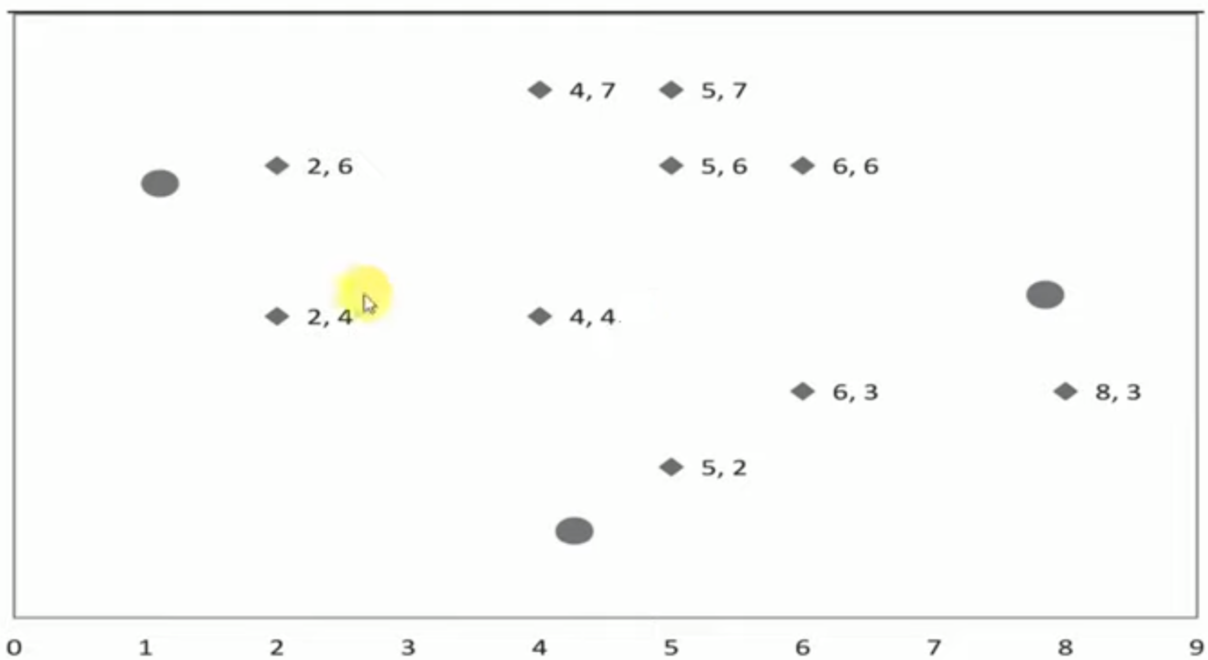
Clustering Exercise

X	Y
2	4
2	6
5	6
4	7
8	3
6	6
5	2
5	7
6	3
4	4

Clustering Exercise



K-Means Algorithm for Clustering



Clustering Exercise

Iteration - 1

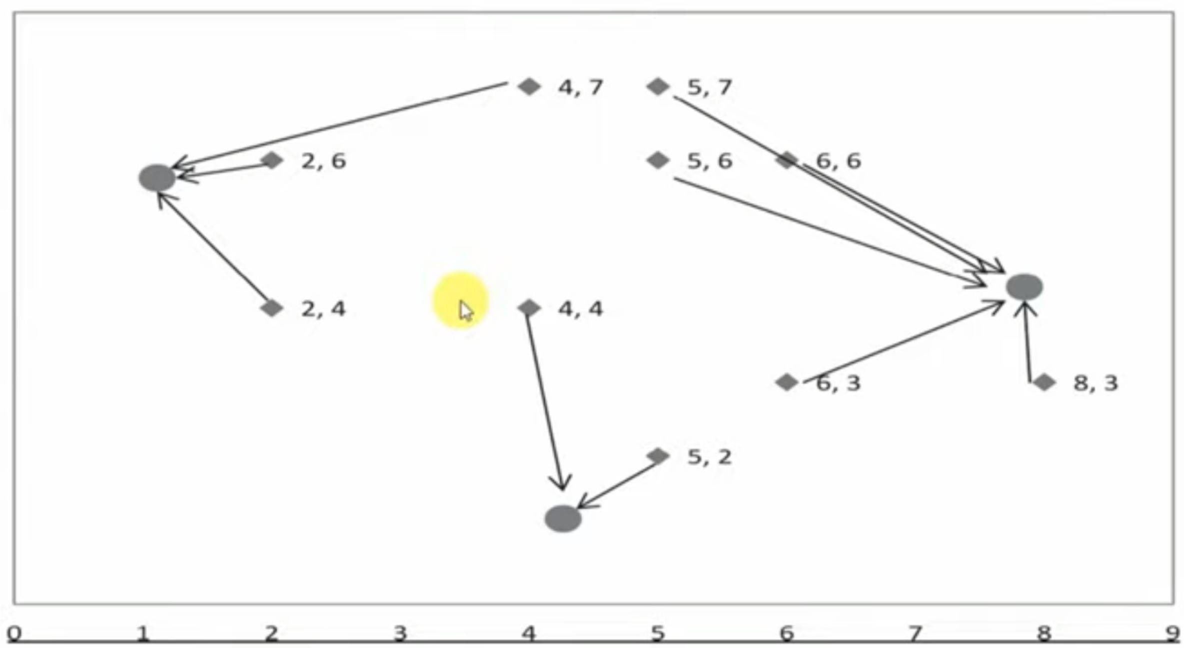
C1 - Seed Point1 – (1, 5)
C2 - Seed Point2 – (4, 1)
C3 - Seed Point3 – (8, 4)

$$D = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$$

C1 – Centroid – (2.66, 5.66)
C2 – Centroid – (4.5, 3)
C3 – Centroid – (6, 5)

		Distance to			Cluster
X	Y	(1, 5)	(4, 1)	(8, 4)	Number
2	4	1.41	3.61	6.00	C1
2	6	1.41	5.39	6.32	C1
5	6	4.12	5.10	3.61	C3
4	7	3.61	6.00	5.00	C1
8	3	7.28	4.47	1.00	C3
6	6	5.10	5.39	2.83	C3
5	2	5.00	1.41	3.61	C2
5	7	4.47	6.08	4.24	C3
6	3	5.39	2.83	2.24	C3
4	4	3.16	3.00	4.00	C2

K-Means Algorithm for Clustering



Clustering Exercise

Iteration - 2

C1 – Centroid – (2.66, 5.66)
C2 – Centroid – (4.5, 3)
C3 – Centroid – (6, 5)

C1 – Centroid – (2.66, 5.66)
C2 – Centroid – (5, 3)
C3 – Centroid – (6, 5.5)

		Distance to			Cluster
X	Y	(2.66, 5.66)	(4.5, 3)	(6, 5)	Number
2	4	1.79	2.69	4.12	C1
2	6	0.74	3.91	4.12	C1
5	6	2.36	3.04	1.41	C3
4	7	1.90	4.03	2.83	C1
8	3	5.97	3.5	2.83	C3
6	6	3.36	3.35	1	C3
5	2	4.34	1.12	3.16	C2
5	7	2.70	4.03	2.24	C3
6	3	4.27	1.5	2	C2
4	4	2.13	1.12	2.24	C2

Clustering Exercise

Iteration - 3

C1 – Centroid – (2.66, 5.66)
C2 – Centroid – (5, 3)
C3 – Centroid – (6, 5.5)

C1 – Centroid – (2.66, 5.66)
C2 – Centroid – (5.75, 3)
C3 – Centroid – (5.33, 6.33)

		Distance to			Cluster
X	**Y**	**(2.66, 5.66)**	**(5, 3)**	**(6, 5.5)**	**Number**
2	4	1.79	3.16	4.27	C1
2	6	0.74	4.24	4.03	C1
5	6	2.36	3.00	1.12	C3
4	7	1.90	4.12	2.50	C1
8	3	5.97	3.00	3.20	C2
6	6	3.36	3.16	0.50	C3
5	2	4.34	1.00	3.64	C2
5	7	2.70	4.00	1.80	C3
6	3	4.27	1.00	2.50	C2
4	4	2.13	1.41	2.50	C2

Clustering Exercise

Iteration - 4

C1 – Centroid – (2.66, 5.66)
C2 – Centroid – (5.75, 3)
C3 – Centroid – (5.33, 6.33)

C1 – Centroid – (2, 5)
C2 – Centroid – (5.75, 3)
C3 – Centroid – (5, 6.5)

		Distance to			Cluster
X	**Y**	**(2.66, 5.66)**	**(5.75, 3)**	**(5.33, 6.33)**	**Number**
2	4	1.79	3.88	4.06	C1
2	6	0.74	4.80	3.35	C1
5	6	2.36	3.09	0.47	C3
4	7	1.90	4.37	1.49	C3
8	3	5.97	2.25	4.27	C2
6	6	3.36	3.01	0.75	C3
5	2	4.34	1.25	4.34	C2
5	7	2.70	4.07	0.75	C3
6	3	4.27	0.25	3.40	C2
4	4	2.13	2.02	2.68	C2

Clustering Exercise

Iteration - 5

C1 – Centroid – (2, 5)
C2 – Centroid – (5.75, 3)
C3 – Centroid – (5, 6.5)

No movement of data Points
Hence these are the final positions

		Distance to			Cluster
X	Y	(2, 5)	(5.75, 3)	(5, 6.5)	Number
2	4	1.00	3.88	3.91	C1
2	6	1.00	4.80	3.04	C1
5	6	3.16	3.09	0.50	C3
4	7	2.83	4.37	1.12	C3
8	3	6.32	2.25	4.61	C2
6	6	4.12	3.01	1.12	C3
5	2	4.24	1.25	4.50	C2
5	7	3.61	4.07	0.50	C3
6	3	4.47	0.25	3.64	C2
4	4	2.24	2.02	2.69	C2

K-Means Clustering – 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, respectively.

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

New Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			2, 10	5, 8	1, 2		
A1	2	10	0.00	3.61	8.06	1	
A2	2	5	5.00	4.24	3.16	3	
A3	8	4	8.49	5.00	7.28	2	
B1	5	8	3.61	0.00	7.21	2	
B2	7	5	7.07	3.61	6.71	2	
B3	6	4	7.21	4.12	5.39	2	
C1	1	2	8.06	7.21	0.00	3	
C2	4	9	2.24	1.41	7.62	2	

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

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

New Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			2, 10	6, 6	1.5, 1.5		
A1	2	10	0.00	5.66	6.52	1	1
A2	2	5	5.00	4.12	1.58	3	3
A3	8	4	8.49	2.83	6.52	2	2
B1	5	8	3.61	2.24	5.70	2	2
B2	7	5	7.07	1.41	5.70	2	2
B3	6	4	7.21	2.00	4.53	2	2
C1	1	2	8.06	6.40	1.58	3	3
C2	4	9	2.24	3.61	6.04	2	1

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

K-Means Clustering – Solved Example

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)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			3, 9.5	6.5, 5.25	1.5, 3.5		
A1	2	10	1.12	6.54	6.52	1	1
A2	2	5	4.61	4.51	1.58	3	3
A3	8	4	7.43	1.95	6.52	2	2
B1	5	8	2.50	3.13	5.70	2	1
B2	7	5	6.02	0.56	5.70	2	2
B3	6	4	6.26	1.35	4.53	2	2
C1	1	2	7.76	6.39	1.58	3	3
C2	4	9	1.12	4.51	6.04	1	1

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

K-Means Clustering – Solved Example

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			3.67, 9	7, 4.33	1.5, 3.5		
A1	2	10	1.94	7.56	6.52	1	1
A2	2	5	4.33	5.04	1.58	3	3
A3	8	4	6.62	1.05	6.52	2	2
B1	5	8	1.67	4.18	5.70	1	1
B2	7	5	5.21	0.67	5.70	2	2
B3	6	4	5.52	1.05	4.53	2	2
C1	1	2	7.49	6.44	1.58	3	3
C2	4	9	0.33	5.55	6.04	1	1

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

K-Means Clustering Algorithm – Solved Example

- Use K Means clustering to cluster the following data into two groups.
- Data Points: { 2, 4, 10, 12, 3, 20, 30, 11, 25 }
- The distance function used is Euclidean distance.
- Initial cluster centroid are M1 = 4 and M2 = 11.

K-Means Clustering Algorithm – Solved Example

Initial Centroids:
M1: 4
M2: 11

Therefore
C1= {2, 4, 3}
C2= {10, 12, 20, 30, 11, 25}

New Centroids: ✓
M1: 3 ✓
M2: 18 ✓

Data Points	Distance to M1	Distance to M2	Cluster	New Cluster
2	2	9	C1	
4	0	7	C1	
10	6	1	C2	
12	8	1	C2	
3	1	8	C1	
20	16	9	C2	
30	26	19	C2	
11	7	0	C2	
25	21	14	C2	

$$d(x_2, x_1) = \sqrt{(x_2 - x_1)^2}$$

K-Means Clustering Algorithm – Solved Example

Current Centroids:
M1: 3
M2: 18

Data Points	Distance to M1	Distance to M2	Cluster	New Cluster
2	1	16	C1	C1
4	1	14	C1	C1
10	7	8	C2	C1
12	9	6	C2	C2
3	0	15	C1	C1
20	17	2	C2	C2
30	27	12	C2	C2
11	8	7	C2	C2
25	22	7	C2	C2

Therefore
C1= {2, 4, 10, 3}
C2= {12, 20, 30, 11, 25}

New Centroids:
M1: 4.75
M2: 19.6

$$d(x_2, x_1) = \sqrt{(x_2 - x_1)^2}$$

K-Means Clustering Algorithm – Solved Example

Current Centroids:
M1: 4.75
M2: 19.6

Data Points	Distance to M1	Distance to M2	Cluster	New Cluster
2	2.75	17.6	C1.	C1
4	0.75	15.6	C1	C1
10	5.25	9.6	C1	C1
12	7.25	7.6	C2	C1
3	1.75	16.6	C1	C1
20	15.25	0.4	C2	C2
30	25.25	10.4	C2	C2
11	6.25	8.6	C2	C1
25	20.25	5.4	C2	C2

Therefore
C1= {2, 4, 10, 11, 12, 3}
C2= {20, 30, 25}

New Centroids:
M1: 7
M2: 25

$$d(x_2, x_1) = \sqrt{(x_2 - x_1)^2}$$

Current Centroids:
M1: 7
M2: 25

Final Cluster are:
C1= {2, 4, 10, 11, 12, 3}
C2= {20, 30, 25}

Data Points	Distance to		Cluster	New Cluster
	M1	M2		
2	5	23	C1	C1
4	3	21	C1	C1
10	3	15	C1	C1
12	5	13	C1	C1
3	4	22	C1	C1
20	13	5	C2	C2
30	23	5	C2	C2
11	4	14	C1	C1
25	18	0	C2	C2

$$d(x_2, x_1) = \sqrt{(x_2 - x_1)^2}$$

K Means Clustering using L1 Distance Euclidean Distance Machine Learning by Dr. M...

K-Means Clustering L1 Distance

Data Point	C1: (4, 0.33, 3)	C2: (0.5, 1.5. 2.5)	Cluster
P1: (1, 2, 3)	4.67	1.5	C2
P2: (0, 1, 2)	5.67	1.5	C2
P3: (3, 0, 5)	3.33	6.5	C1
P4: (4, 1, 3)	0.67	4.5	C1
P5: (5, 0, 1)	3.33	7.5	C1

Solved Example

K-Means Clustering using L1 Distance

- Consider the 5 data points shown below:

P1: (1, 2, 3)

P2: (0, 1, 2)

P3: (3, 0, 5)

P4: (4, 1, 3)

P5: (5, 0, 1)

- Apply the **Kmeans** clustering algorithm, to group those data points into 2 clusters, using the L1 distance measure.
- Consider the initial centroids are C1: (1, 0, 0) and C2: (0, 1, 1).

K-Means Clustering using L1 Distance

- L1 distance is just manhattan distance: sum of differences in each dimension – **ITERATION 1**

Formula: $|x_2-x_1|+|y_2-y_1|+|z_2-z_1|$

Data Point	C1: (1, 0, 0)	C2: (0, 1, 1)	Cluster
P1: (1, 2, 3)	0+2+3 5 ✓	4	C2
P2: (0, 1, 2)	1+1+2 4 ✓	1	C2
P3: (3, 0, 5)	7	8	C1
P4: (4, 1, 3)	7	6	C2
P5: (5, 0, 1)	5	6	C1

K-Means Clustering using L1 Distance

- L1 distance is just manhattan distance: sum of differences in each dimension - **ITERATION 2**

Data Point	C1: (4, 0, 3)	C2: (1.6, 1.3, 2.6)	Cluster
P1: (1, 2, 3)	3+ 5	1.7	C2
P2: (0, 1, 2)	6	2.5	C2
P3: (3, 0, 5)	3	5.3	C1
P4: (4, 1, 3)	1	3.1	C1
P5: (5, 0, 1)	3	6.3	C1

K-Means Clustering using L1 Distance

- L1 distance is just manhattan distance: sum of differences in each dimension - **ITERATION 3**

Data Point	C1: (4, 0.33, 3)	C2: (0.5, 1.5. 2.5)	Cluster
P1: (1, 2, 3)	3+1.7 +0 4.67	1.5	C2
P2: (0, 1, 2)	5.67	1.5	C2
P3: (3, 0, 5)	3.33	6.5	C1
P4: (4, 1, 3)	0.67	4.5	C1
P5: (5, 0, 1)	3.33	7.5	C1