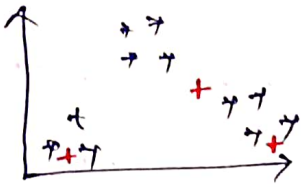
K-Means Clustering - Manual Mathematical Calculation

- K Means is an unsupervised ML algo used for clustering.
- Goal is to categorize the data into K-no of clusters.
- It begins by selecting K-points randomly as the initial centroids then assign each data points to the nearest centroid.
- After every point is assigned to a centroid, the centroids are recalculated by taking the mean of all the datapoints that were assigned to each centroid then the datapoints are reassigned to the nearest centroid based on the updated once.
- This process repeats until the cluster no longer change.

(or)

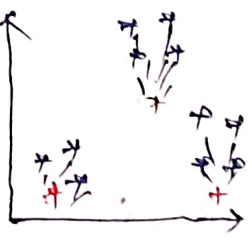
Step 1 : Choose the number of clusters (K) you wish to put the data into, let us choose 3 (K).

Step 2 : Randomly initialize K points, these are called centroids.



Step 3 : Identify the points closest to each centroid.

Step 4 : Calculate the mean of the points in each cluster and move each centroid to that mean point.



Step 5 : Repeat Step 3 & 4 until the centroid value stays the same.

K Means Clustering - Solved Example

- Datapoints:
  A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4)
  C1(1,2), C2(4,9)
- Distance function: Euclidean distance
- Suppose initially we assign A1, B1 and C1 as the centers of each cluster respectively.

Iteration 1:

Initial Centroids	Datapoints	x_1	y_1	Distance to (x_2 y_2) = (2 10)	(5 8)	(1 2)	Cluster	New Cluster
A1: (2,10)	A1	2	10	0	3.61	8.06	1	
B1: (5,8)	A2	2	5	5	4.24	3.16	3	
C1: (1,2)	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	

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

Simple calculation

Euclidean distance $d_1(P_1, P_2) = \sqrt{(2-2)^2 + (10-10)^2} = 0$

$d_2(P_1, P_2) = \sqrt{(2-2)^2 + (10-5)^2} = \sqrt{(5)^2} = 5$

$d_3(P_1, P_2) = \sqrt{(2-8)^2 + (10-4)^2} = \sqrt{(6)^2 + (6)^2} = 8.49$

$d_4(P_1, P_2) = \sqrt{(2-5)^2 + (10-8)^2} = \sqrt{(3)^2 + (2)^2} = \sqrt{13} = 3.61$

$d_5(P_1, P_2) = \sqrt{(2-7)^2 + (10-5)^2} = \sqrt{(5)^2 + (5)^2} = 7.07$

$d_6 = 7.21$, $d_7 = 8.06$, $d_8 = 2.24$

New Centroid Calculation for cluster

1st cluster: A1 (2,10) → (2, 10)

2nd cluster → $\frac{8+5+7+6+4}{5}$, $\frac{4+8+5+4+9}{5}$ → (6, 6)

3rd cluster → $(2+1)/2 = 1.5$, $(5+2)/2 = 3.5$ → (1.5, 3.5)

Current Centroids

A1(2, 10) B1(6, 6) C1(1.5, 3.5)

Distance to

Datapoints		2 10	6 6	1.5 3.5	Cluster	New cluster
A1	2 10	0.0	5.66	6.52	1	1
A2	2 5	5.0	4.12	1.58	3	3
A3	8 4	8.49	2.83	6.52	2	2
B1	5 8	3.61	2.84	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

New Centroids

(3, 9.5) (6.5, 5.25) (1.5, 3.5)

Distance to

	Data points	3 9.5	6.5 5.25	1.5 3.5	Cluster	New clustw
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

New Centroids.

A1 (3.67, 9) B1 (7, 4.33) c1 (1.5, 3.5)

	Datapoints		Distance to 3.67, 9	7	4.33	1.5	3.5	Cluster	New cluster
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

Cluster = = New cluster

Stop the iteration.

A1, B1, c2 → belongs to → cluster 1

A3, B2, B3 → belongs to → cluster 2

A2, c1 → belongs to → cluster 3