Chapter 11: Kmeans Clustering

- **Clustering:**
- Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.

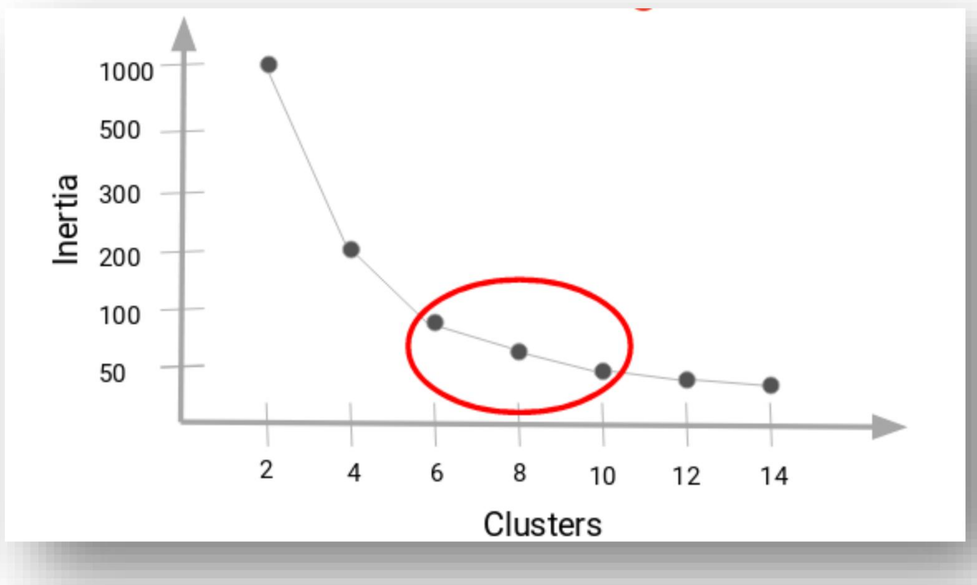
- **K-Means Clustering:**
- The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset.

KMeans Clustering:

- KMeans is a centroid based algorithm or distance based algorithm, i.e. each cluster is associated with a centroid and we calculate the distances to assign a point to a cluster.
- How to apply:
  - S1: choose the number of clusters "k"
  - S2: select k random points from the data as centroids.
  - S3: assign all the points to the closest cluster centroid.
  - S4: recomputed the centroids of newly formed clusters.
  - Repeat S3 and S4 until the centroids of newly formed clusters do not change or maximum number of iterations are reached.

How to choose K:

- The cluster value where this decrease in inertia value(**inertia actually calculates the sum of distances of all the points within a cluster from the centroid of that cluster.**) becomes constant can be chosen as the right cluster value for our data.



P1

Instance	X	Y
P1	1	1
P2	1.5	2
P3	3	4
P4	5	7
P5	3.5	5
P6	4.5	5
P7	3.5	4.5

	C1 = (1,1)	C2 = (1.5, 2)	Cluster
	Dist(P, C1)	Dist(P,C2)	
P1(1,1)	sqrt($(1-1)^2 + (1-1)^2$) = 0	sqrt($(1-1.5)^2 + (1-2)^2$) = 1.12	
P2(1.5,2)	sqrt($(1.5-1)^2 + (2-1)^2$) = 1.12	sqrt($(1.5-1.5)^2 + (2-2)^2$) = 0	
P3(3,4)	sqrt($(3-1)^2 + (4-1)^2$) = 3.61	sqrt($(3-1.5)^2 + (4-2)^2$) = 2.5	
P4(5,7)	sqrt($(5-1)^2 + (7-1)^2$) = 7.21	sqrt($(5-1.5)^2 + (7-2)^2$) = 6.10	
P5(3.5,5)	sqrt($(3.5-1)^2 + (5-1)^2$) = 4.72	sqrt($(3.5-1.5)^2 + (5-2)^2$) = 3.61	
P6(4.5,5)	sqrt($(4.5-1)^2 + (5-1)^2$) = 5.32	sqrt($(4.5-1.5)^2 + (5-2)^2$) = 4.24	
P7(3.5,4.5)	sqrt($(3.5-1)^2 + (4.5-1)^2$) = 4.30	sqrt($(3.5-1.5)^2 + (4.5-2)^2$) = 3.20	

C1 =

C2 =

Instance	X	Y
P1	1	1
P2	1.5	2
P3	3	4
P4	5	7
P5	3.5	5
P6	4.5	5
P7	3.5	4.5

	C1 = (1,1)	C2 = (3.5, 4.58)	Cluster
	Dist(P, C1)	Dist(P,C2)	
P1(1,1)	sqrt($(1-1)^2 + (1-1)^2$) = 0	sqrt($(1-3.5)^2 + (1-4.58)^2$) = 4.37	
P2(1.5,2)	sqrt($(1.5-1)^2 + (2-1)^2$) = 1.12	sqrt($(1.5-3.5)^2 + (2-4.58)^2$) = 3.26	
P3(3,4)	sqrt($(3-1)^2 + (4-1)^2$) = 3.61	sqrt($(3-3.5)^2 + (4-4.58)^2$) = 0.77	
P4(5,7)	sqrt($(5-1)^2 + (7-1)^2$) = 7.21	sqrt($(5-3.5)^2 + (7-4.58)^2$) = 2.85	
P5(3.5,5)	sqrt($(3.5-1)^2 + (5-1)^2$) = 4.72	sqrt($(3.5-3.5)^2 + (5-4.58)^2$) = 0.42	
P6(4.5,5)	sqrt($(4.5-1)^2 + (5-1)^2$) = 5.32	sqrt($(4.5-3.5)^2 + (5-4.58)^2$) = 1.08	
P7(3.5,4.5)	sqrt($(3.5-1)^2 + (4.5-1)^2$)=4.30	sqrt($(3.5-3.5)^2 + (4.5-4.58)^2$)=0.08	

C1 =

C2 =

Instance	X	Y
P1	1	1
P2	1.5	2
P3	3	4
P4	5	7
P5	3.5	5
P6	4.5	5
P7	3.5	4.5

	C1 = (1.25,1.5)	C2 = (3.9, 5.1)	Cluster
	Dist(P, C1)	Dist(P,C2)	
P1(1,1)	sqrt($(1-1.25)^2 + (1-1.5)^2$) = 0.56	sqrt($(1-3.9)^2 + (1- 5.1)^2$) = 5.02	
P2(1.5,2)	sqrt($(1.5-1.25)^2 + (2-1.5)^2$) = 0.56	sqrt($(1.5-3.9)^2 + (2 - 5.1)^2$) = 3.92	
P3(3,4)	sqrt($(3-1.25)^2 + (4-1.5)^2$) = 3.05	sqrt($(3-3.9)^2 + (4-5.1)^2$) = 1.42	
P4(5,7)	sqrt($(5-1.25)^2 + (7-1.5)^2$) = 6.66	sqrt($(5-3.9)^2 + (7-5.1)^2$) = 2.2	
P5(3.5,5)	sqrt($(3.5-1.25)^2 + (5-1.5)^2$) = 4.16	sqrt($(3.5-3.9)^2 + (5-5.1)^2$) = 0.41	
P6(4.5,5)	sqrt($(4.5-1.25)^2 + (5-1.5)^2$) = 4.78	sqrt($(4.5-3.9)^2 + (5-5.1)^2$) = 0.61	
P7(3.5,4.5)	sqrt($(3.5-1.25)^2 + (4.5-1.5)^2$) = 3.75	sqrt($(3.5-3.9)^2 + (4.5-5.1)^2$) = 0.72	

C1 =

C2 =

P2

Instance	A	B
P1	2	2
P2	1	14
P3	10	7
P4	1	11
P5	3	4
P6	11	8
P7	4	3
P8	12	9

	C1 = (2,2)	C2 = (1,14)	C3=(4,3)	Cluster
	Dist(P, C1)	Dist(P,C2)	Dist(P,C3)	
P1(2,2)	sqrt($(2-2)^2 + (2-2)^2$) = 0	sqrt($(2-1)^2 + (2-14)^2$) = 12.04	sqrt($(2-4)^2 + (2-3)^2$) = 2.24	
P2(1,14)	sqrt($(1-2)^2 + (14-2)^2$) =12.04	sqrt($(1-1)^2 + (14-14)^2$) = 0	sqrt($(1-4)^2 + (14-3)^2$) = 11.4	
P3(10,7)	sqrt($(10-2)^2 + (7-2)^2$) = 9.43	sqrt($(10-1)^2 + (7-14)^2$) = 11.4	sqrt($(10-4)^2 + (7-3)^2$) = 7.21	
P4(1,11)	sqrt($(1-2)^2 + (11-2)^2$) = 9.06	sqrt($(1-1)^2 + (11-14)^2$) = 3	sqrt($(1-4)^2 + (11-13)^2$) = 8.54	
P5(3,4)	sqrt($(3-2)^2 + (4-2)^2$) = 2.24	sqrt($(3-1)^2 + (4-14)^2$) = 10.2	sqrt($(3-4)^2 + (4-3)^2$) = 1.41	
P6(11,8)	sqrt($(11-2)^2 + (8-2)^2$) = 10.82	sqrt($(11-2)^2 + (8-2)^2$) = 11.66	sqrt($(11-4)^2 + (8-3)^2$) = 8.60	
P7(4,3)	sqrt($(4-2)^2 + (3-2)^2$) = 2.24	sqrt($(4-1)^2 + (3-14)^2$) = 11.4	sqrt($(4-4)^2 + (3-3)^2$) = 0	
P8(12,9)	sqrt($(12-2)^2 + (9-2)^2$) = 12.21	sqrt($(12-1)^2 + (9-14)^2$) = 12.08	sqrt($(12-4)^2 + (9-3)^2$) = 10	

C1 =

C2 =

C3 =

Instance	A	B
P1	2	2
P2	1	14
P3	10	7
P4	1	11
P5	3	4
P6	11	8
P7	4	3
P8	12	9

	C1 = (2,2)	C2 = (1, 12.5)	C3=(8, 6.2)	Cluster
	Dist(P, C1)	Dist(P,C2)	Dist(P,C3)	
P1(2,2)	sqrt($(2-2)^2 + (2-2)^2$) = 0	sqrt($(2-1)^2 + (2-12.5)^2$) = 10.55	sqrt($(2-8)^2 + (2-6.2)^2$) = 7.32	
P2(1,14)	sqrt($(1-2)^2 + (14-2)^2$ = 12.04	sqrt($(1-1)^2 + (14-12.5)^2$) = 1.5	sqrt($(1-8)^2 + (14-6.2)^2$) = 10.48	
P3(10,7)	sqrt($(10-2)^2 + (7-2)^2$) = 9.43	sqrt($(10-1)^2 + (7-12.5)^2$) = 10.55	sqrt($(10-8)^2 + (7-6.2)^2$) = 2.15	
P4(1,11)	sqrt($1-2)^2 + (11-2)^2$) = 9.06	sqrt($(1-1)^2 + (11-12.5)^2$) = 1.5	sqrt($(1-8)^2 + (11-6.2)^2$) = 8.49	
P5(3,4)	sqrt($(3-2)^2 + (4-2)^2$) = 2.24	sqrt($(3-1)^2 + (4-12.5)^2$) = 8.73	sqrt($(3-8)^2 + (4-6.2)^2$) = 5.46	
P6(11,8)	sqrt($(11-2)^2 + (8-2)^2$) = 10.82	sqrt($(11-1)^2 + (8-12.5)^2$) = 10.97	sqrt($(11-8)^2 + (8-6.2)^2$) = 3.50	
P7(4,3)	sqrt($(4-2)^2 + (3-2)^2$) = 2.24	sqrt($(4-1)^2 + (3-12.5)^2$) = 9.96	sqrt($(4-8)^2 + (3-6.2)^2$) = 5.12	
P8(12,9)	sqrt($(12-2)^2 + (9-2)^2$) = 12.21	sqrt($(12-1)^2 + (9-12.5)^2$) = 11.54	sqrt($(12-8)^2 + (9-6.2)^2$) = 4.88	

C1 =

C2 =

C3 =

Instance	A	B
P1	2	2
P2	1	14
P3	10	7
P4	1	11
P5	3	4
P6	11	8
P7	4	3
P8	12	9

	C1 = (3,3)	C2 = (1, 12.5)	C3=(11, 8)	Cluster
	Dist(P, C1)	Dist(P,C2)	Dist(P,C3)	
P1(2,2)	sqrt($(2-3)^2 + (2-3)^2$) = 1.41	sqrt($(2-1)^2 + (2-12.5)^2$) = 10.55	sqrt($(2-11)^2 + (2-8)^2$) = 10.82	
P2(1,14)	sqrt($(1-3)^2 + (14-3)^2$) = 11.18	sqrt($(1-1)^2 + (14-12.5)^2$) = 1.5	sqrt($(1-11)^2 + (14-8)^2$) = 11.66	
P3(10,7)	sqrt($(10-3)^2 + (7-3)^2$) = 8.06	sqrt($(10-1)^2 + (7-12.5)^2$) = 10.55	sqrt($(10-11)^2 + (7-8)^2$) = 1.41	
P4(1,11)	sqrt($(1-3)^2 + (11-3)^2$) = 8.25	sqrt($(1-1)^2 + (11-12.5)^2$) = 1.5	sqrt($(1-11)^2 + (11-8)^2$) = 10.44	
P5(3,4)	sqrt($(3-3)^2 + (4-3)^2$) = 1.00	sqrt($(3-1)^2 + (4-12.5)^2$) = 8.73	sqrt($(3-11)^2 + (4-8)^2$) = 8.94	
P6(11,8)	sqrt($(11-3)^2 + (8-3)^2$) 9.43	sqrt($(11-1)^2 + (8-12.5)^2$) = 10.97	sqrt($(11-11)^2 + (8-8)^2$) = 0	
P7(4,3)	sqrt($(4-3)^2 + (3-3)^2$) =1.0	sqrt($(4-1)^2 + (3-12.5)^2$) 9.96	sqrt($(4-11)^2 + (3-8)^2$) 8.6	
P8(12,9)	sqrt($(12-3)^2 + (9-3)^2$) = 10.82	sqrt($(12-1)^2 + (9-12.5)^2$) = 11.54	sqrt($(12-11)^2 + (9-8)^2$) 1.41	

C1 =

C2 =

C3 =

P3

NAME	RUNS	WICKETS
SACHIN	99	2
DRAVID	95	1
SRINATH	8	20
ANKIT	20	15
VIRAT	97	5

NAME	RUNS	WICKETS
SACHIN	1.000000	0.052632
DRAVID	0.956044	0.000000
SRINATH	0.000000	1.000000
ANKIT	0.131868	0.736842
VIRAT	0.978022	0.210526

	C1 = (1.00, 0.05)	**C2 = (0.00, 1.00)**	**Cluster**
	Dist(P, C1)	**Dist(P,C2)**	
Sachin (1.00, 0.05)	sqrt($(1.00-1.00)^2$ + $(0.05-0.05)^2$)	sqrt($(1.00-0.00)^2$ + $(0.05-1.00)^2$)	
Dravid (0.96, 0.00)	sqrt($(0.96-1.00)^2$ + $(0.00-0.05)^2$)	sqrt($(0.96-0.00)^2$ + $(0.00-1.00)^2$)	
Srinath (0.00, 1.00)	sqrt($(0.00-1.00)^2$ + $(1.00-0.05)^2$)	sqrt($(0.00-0.00)^2$ + $(1.00-1.00)^2$)	
Ankit (0.13, 0.74)	sqrt($(0.13-1.00)^2$ + $(0.74-0.05)^2$)	sqrt($(0.13-0.00)^2$ + $(0.74-1.00)^2$)	
Virat (0.98, 0.21)	sqrt($(0.98-1.00)^2$ + $(0.21-0.05)^2$)	sqrt($(0.98-0.00)^2$ + $(0.21-1.00)^2$)	

C1 =

C2 =

NAME	RUNS	WICKETS
SACHIN	99	2
DRAVID	95	1
SRINATH	8	20
ANKIT	20	15
VIRAT	97	5

NAME	RUNS	WICKETS
SACHIN	1.000000	0.052632
DRAVID	0.956044	0.000000
SRINATH	0.000000	1.000000
ANKIT	0.131868	0.736842
VIRAT	0.978022	0.210526

	C1 = (0.98, 0.09)	**C2 = (0.07, 0.87)**	**Cluster**
	Dist(P, C1)	**Dist(P,C2)**	
Sachin (1.00, 0.05)	sqrt($(1.00-0.98)^2$ + $(0.05-0.09)^2$)	sqrt($(1.00-0.07)^2$ + $(0.05-0.87)^2$)	
Dravid (0.96, 0.00)	sqrt($(0.96-0.98)^2$ + $(0.00-0.09)^2$)	sqrt($(0.96-0.07)^2$ + $(0.00-0.87)^2$)	
Srinath (0.00, 1.00)	sqrt($(0.00-0.98)^2$ + $(1.00-0.09)^2$)	sqrt($(0.00-0.07)^2$ + $(1.00-0.87)^2$)	
Ankit (0.13, 0.74)	sqrt($(0.13-0.98)^2$ + $(0.74-0.09)^2$)	sqrt($(0.13-0.07)^2$ + $(0.74-0.87)^2$)	
Virat (0.98, 0.21)	sqrt($(0.98-0.98)^2$ + $(0.21-0.09)^2$)	sqrt($(0.98-0.07)^2$ + $(0.21-0.87)^2$)	

C1 =

C2 =

Applications of K-Means:

- **Customer Segmentation:**
  One of the most common applications of clustering is customer segmentation in Banking, Telecom, E-Commerce, Sports, Advertising, Sales, etc.

- **Document Clustering:**
  Clustering helps us group these documents such that similar documents are in the same clusters.

- **Image Segmentation**
  We can apply clustering to create clusters having similar pixels in the same group.

- **Recommendation Engines**
  Clustering can also be used in recommendation engines. Let's say you want to recommend songs to your friends. You can look at the songs liked by that person and then use clustering to find similar songs and finally recommend the most similar songs.