28/02/23

UNIT-IV

* Types of Unsupervised learning :-

Clustering	Association
1) method of grouping the objects into clusters based on similarities.	1) It involve discovering the patterns in data finding co occurance.
2) It discover inherent grouping in data	2) used in finding the relationship b/w variables in the large data base.
3) K-means clustering, hirarchical clustering, density based clustering, EM clustering [Expatization, maximization clustering]	3) Market basket Analysis, Eg:- machine + Cover + Tempuglass milk + bread + egg milk + bread + cheese milk + curd + Bread.

* Challenges in Unsupervised learning :-

→ Unsupervised learning is computationally complex.
→ Cannot get precise information regarding data sorting
→ Human intervention to validate the output variables.
→ High risk of inaccurate results & lack of transparency.

* K-Means Clustering :-

1) It is an unsupervised learning algorithm.
2) given a data set of items with certain features and values for this features the algorithm will caligorise the items into K group of clusters of similarities.
3) to calculate the similarity we can use the Euclidean, Manhattan & Hamming distances for measurement.

4) K-Means : 'K' indicates no. of clusters & 'd' indicates list of data points.

5) Choose K no. of random data points as initial centroids.

6) Repeat till clusters centers stabilize.

7) Allocate each point in D to the nearest of K-th centroid.

8) Compute centroid for the cluster using all points in the clusters.

* Advantages :-

→ It is simple, easy to understand.

→ No other clustering algos performs better than K-means.

→ It is also efficient in which time taken to cluster K-means rises linear with data point.

* Disadvantages :- no. of.

→ the user needs to specify an initial value of K.

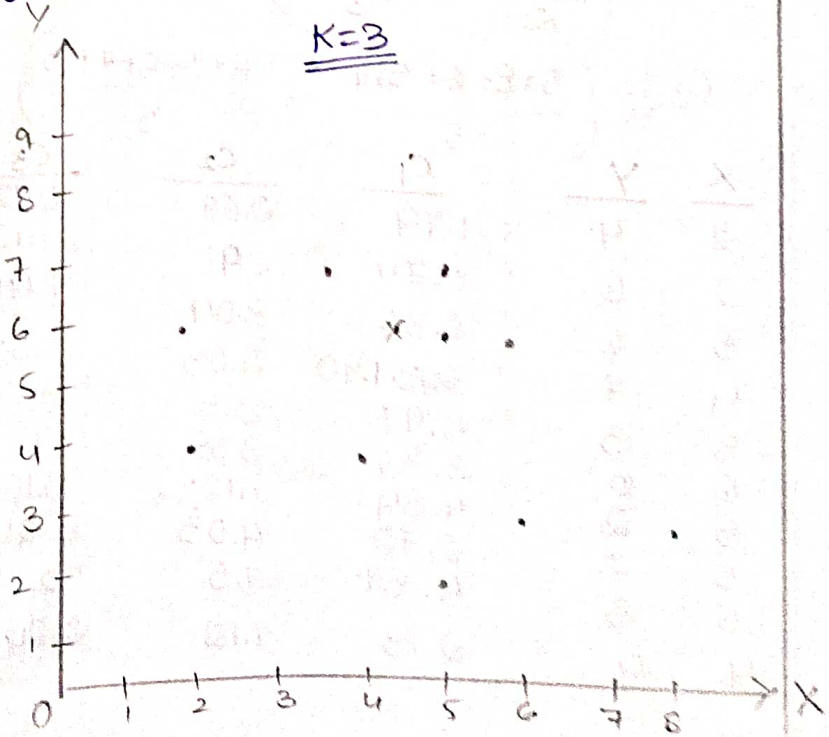
→ the process of finding the clusters may not converge.

→ It is not suitable for discovering clusters that are not hyper ellipsoids (or) hyper spheres.

Example :-

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

K=3



X	Y	(1,5)	(4,1)	(8,4)	cluster number
2	4	1.41	3.61	6	C_1
2	6	1.41	5.39	6.0	C_1
5	6	4.12	5.10	3.61	C_3
4	7	3.61	6	5	C_1
8	3	7.28	4.47	1	C_3
6	6	5.10	5.39	2.83	C_3
5	2	5	1.41	3.61	C_2
5	7	4.47	6.08	4.24	C_3
6	3	5.39	2.83	2.24	C_3
4	4	3.16	3	4.	C_2

$C_1 = (1,5)$ $C_2 = (4,1)$ $C_3 = (8,4)$.

~~$C_1 = \left[\frac{13}{3}, \frac{10}{3}\right]$~~

$C_1 = \left(\frac{2+2+4}{3}, \frac{4+6+7}{3}\right) = \left(\frac{8}{3}, \frac{17}{3}\right) = (2.66, 5.66)$

$C_2 = \left(\frac{9+4}{2}, \frac{2+4}{2}\right) = (4.5, 3)$

$C_3 = \left(\frac{5+8+6+5+6}{5}, \frac{6+3+6+7+3}{5}\right) = (6,5)$

X	Y	C_1	C_2	C_3	cluster no.
2	4	1.79	2.69	4.12	C_1
2	6	0.74	3.91	4.12	C_1
5	6	2.36	3.04	1.41	C_3
4	7	~~2.72~~ 1.90	4.03	2.83	C_1
8	3	5.97	3.5	2.83	C_3
6	6	3.36	3.35	1	C_3
5	2	4.34	1.12	3.16	C_2
5	7	2.70	4.03	2.24	C_3
6	3	4.37	1.5	2	C_2
4	4	2.13	1.12	2.24	C_2

$C_1 = \left(\frac{0+2+4}{3}, \frac{4+6+7}{3}\right) = (2.66, 5.66)$

$C_2 = \left(\frac{5+6+4}{3}, \frac{6+3+4}{3}\right) = (5.3)$

$C_3 = \left(\frac{5+6+6+5}{4}, \frac{6+3+6+7}{4}\right) = (5.5, 5.5)$

Iteration-3

Iteration-3

X	Y	C_1 (2.66, 5.66)	C_2 (5.3)	C_3 (5.5, 5.5)	Cluster no
2	4	1.79	3.46	4.32	C_1
2	6	0.74	4.84	4.03	C_1
5	6	3.36 0.36	3	1.12	C_3
4	7	1.90	4.42	0.90	C_1
8	3	5.97	3	3.60	C_2
6	6	3.36	3.46	0.50	C_3
5	2	4.34	1	3.64	C_2
5	7	0.70	4	1.50	C_3
6	3	4.27	1	0.50	C_2
4	4	0.13	1.41	0.50	C_2

$C_1 = \left(\frac{0+2+4}{3}, \frac{4+6+7}{3}\right) = \left(\frac{6}{3}, \frac{17}{3}\right) = (2.66, 5.66)$

$C_2 = \left(\frac{5+5+6+4}{4}, \frac{3+2+3+4}{4}\right) = \left(\frac{23}{4}, \frac{12}{4}\right) = (5.75, 3)$

$C_3 = \left(\frac{5+6+5}{3}, \frac{6+6+7}{3}\right) = (5.33, 6.33)$

Iteration-4

X	Y	C_1 (2.66, 5.66)	C_2 (5.75, 3)	C_3 (5.33, 6.33)	Cluster no
2	4	1.79	3.66	4.06	C_1
2	6	0.74	4.80	3.35	C_1
5	6	0.36	3.04	0.47	C_3
4	7	1.90	4.37	1.49	C_3
8	3	5.97	0.25	4.37	C_2
6	6	3.36	3.01	0.75	C_3
5	2	4.34	1.95	4.34	C_2
5	7	2.70	4.07	0.75	C_3
6	3	4.27	0.25	3.40	C_2
4	4	2.13	2.02	0.68	C_2

$C_1 = (2.5)$

$C_2 = \left(\frac{23}{4}, 3\right) = (5.75, 3)$

$C_3 = (5, 6.5)$

$C_1 = \left(\frac{2+2+4}{3}, \frac{4+6+7}{3}\right) = (2.66, 5.66)$

$C_2 = \left(\frac{5+6+4}{3}, \frac{2+3+4}{3}\right) = (5, 3)$

$C_3 = \left(\frac{5+8+6+5}{4}, \frac{6+3+6+7}{4}\right) = (6, 5.5)$

13/23

Situation - 3.

X	Y	C_1 (2.66, 5.66)	C_2 (5, 3)	C_3 (6, 5.5)	Cluster no
2	4	1.79	3.16	4.27	C_1
2	6	0.74	4.24	4.03	C_1
5	6	2.36	3	1.12	C_3
4	7	1.90	4.12	2.50	C_1
8	3	5.97	3	3.20	C_2
6	6	3.36	3.16	0.50	C_3
5	2	4.34	1	3.64	C_2
5	7	2.70	4	1.80	C_3
6	3	4.27	1	0.50	C_2
4	4	2.13	1.41	0.50	C_2

$C_1 = \left(\frac{2+2+4}{3}, \frac{4+6+7}{3}\right) = \left(\frac{8}{3}, \frac{17}{3}\right) = (2.66, 5.66)$

$C_2 = \left(\frac{8+5+6+4}{4}, \frac{3+2+3+4}{4}\right) = \left(\frac{23}{4}, \frac{12}{4}\right) = (5.75, 3)$

$C_3 = \left(\frac{5+6+5}{3}, \frac{6+6+7}{3}\right) = (5.33, 6.33)$

Situation - 4

X	Y	C_1 (2.66, 5.66)	C_2 (5.75, 3)	C_3 (5.33, 6.33)	Cluster no
2	4	1.79	3.88	4.06	C_1
2	6	0.74	4.80	3.35	C_1
5	6	2.36	3.09	0.47	C_3
4	7	1.90	4.37	1.49	C_3
8	3	5.97	2.25	4.27	C_2
6	6	3.36	3.01	0.75	C_3
5	2	4.34	1.25	4.34	C_2
5	7	2.70	4.07	0.75	C_3
6	3	4.27	0.95	3.40	C_2
4	4	2.13	2.02	2.68	C_2

$C_1 = (2, 5)$

$C_2 = \left(\frac{23}{4}, 3\right) = (5.75, 3)$

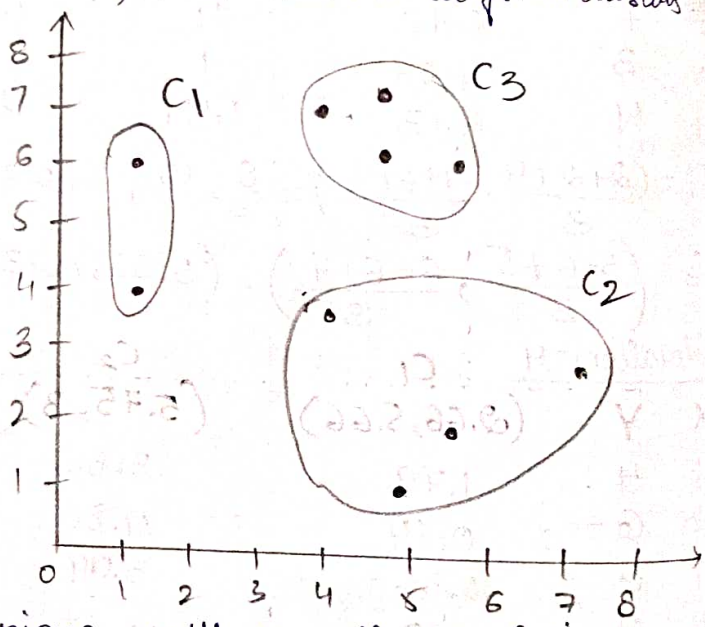
$C_3 = (5, 6.5)$

iteration 5

X	Y	C_1 (2,5)	C_2 (5.75,3)	C_3 (5,6.5)	Cluster number
2	4	1	3.88	3.91	C_1
2	6	1	4.80	3.04	C_1
5	6	3.16	3.09	0.50	C_3
4	7	2.83	3.37	1.12	C_3
8	3	6.32	2.25	4.61	C_2
6	6	4.17	3.01	1.12	C_3
5	2	4.24	1.25	4.50	C_2
5	7	3.61	4.07	0.50	C_3
6	3	4.47	0.25	3.64	C_2
4	4	3.24	2.02	2.69	C_2

there is no movements of data points, hence these are the final clusters

$C_1 = \left(\frac{2+2}{2}, \frac{4+6}{2}\right) = (2,5)$



3/3/23 * Agglomerative clustering :-

for the given dataset find the clusters using a single link technique, we'll use the euclidian distance and draw the Dendrogram.

1) Single link
2) Complete link
3) Average link.

(1) Single link :-

Eg:

Sample no	X	Y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	0.30

$$d(P_1,P_2)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

Step1 :- Compute the distance matrix

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.23	0.22	0.37	0.34	0.24
P_2	0.23	0	0.14	0.19	0.14	0.24
P_3	0.22	0.14	0	0.15	0.28	0.10
P_4	0.37	0.19	0.15	0	0.28	0.22
P_5	0.34	0.14	0.28	0.28	0	0.37
P_6	0.24	0.24	0.10	0.22	0.37	0

Step2 :-

	P_1	P_2	P_3,P_6	P_4	P_5
P_1	0				
P_2	0.23	0			
P_3,P_6	0.22	0.14	0		
P_4	0.37	0.19	0.13	0	
P_5	0.34	0.14	0.28	0.28	0

merge (P_3,P_6)

merge (P_3,P_6,P_4)

	P_1	P_2	$P_3P_6P_4$	P_5
P_1	0			
P_2	0.23	0		
$P_3P_6P_4$	0.22	0.14	0	
P_5	0.34	0.14	0.28	0

merge (P_2P_5)

	P_1	P_2P_5	$P_3P_6P_4$
P_1	0		
P_2P_5	0.23	0	
$P_3P_6P_4$	0.22	0.14	0

merge ($P_2P_5P_3P_6P_4$)

	P_1	$P_2P_5P_3P_6P_4$
P_1	0	
$P_2P_5P_3P_6P_4$	0.23	0

4/3/23

2/* **Complete linkage:-**

given a 1D dataset {1,5,8,10,2} use the agglomerative clustering algorithm with the complete link by using the Euclidian distance to establish the hierarcial grouping relationship.

→ By using the calling threshold value of 5, how many clusters are there?

→ We need to calculate the distance matrix using 1 dimensional dataset.

1 point $ED = \sqrt{(x_2 - x_1)^2}$

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

(i) distance matrix

	1	5	8	10	2
1	0	4	7	9	1
5	4	0	3	5	3
8	7	3	0	2	6
10	9	5	2	0	8
2	1	3	6	8	0

	1	2	3	4	5
1	0	4	7	9	(1)
2	4	0	3	5	3
3	7	3	0	2	6
4	9	5	2	0	8
5	(1)	3	6	8	0

merge (1,5)

compare.

$d(2,\{1,5\}) = \max\{d(2,1), d(2,5)\} = \max\{4,3\} = 4$

$d(3,\{1,5\}) = \max\{d(3,1), d(3,5)\} = \max\{7,6\} = 7$

$d(4,\{1,5\}) = \max\{d(4,1), d(4,5)\} = \max\{9,8\} = 9$

merge (3,4)

$d(\{1,5\},\{3,4\}) = \max\{d(\{1,5\},3), d(\{(1,5),4\}) = \max\{7,9\} = 9$

$d\{2,\{3,4\}\} = \max\{d(2,3), d(2,4)\} = \max\{3,5\} = 5$

(ii)

	1,5	2	3	4
1,5	0	4	7	9
2	4	0	3	5
3	7	3	0	(2)
4	9	5	(2)	0

(iii)

	1,5	2	3,4
1,5	0	4	9
2	4	0	5
3,4	9	5	0

merge ({1,5}, 2)

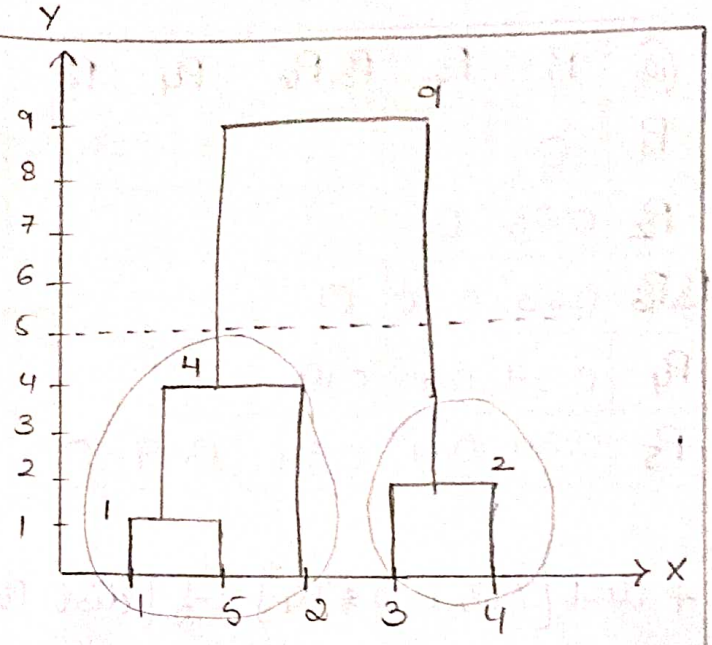
$d((3,4),(\{1,5\},2)) = \max(d(\{3,4\},\{1,5\})\ d(\{3,4\},2))$

$= \max(9,5) = 9$

(iv)

	1,5,2	3,4
1,5,2	0	9
3,4	9	0

there are 2 clusters in the given data.



3) Average Link :-

Find the average link using.

	X	Y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	0.30

①	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.23	0				
P_3	0.22	0.15	0			
P_4	0.37	0.20	0.15	0		
P_5	0.34	0.14	0.28	0.29	0	
P_6	0.23	0.25	0.11	0.22	0.39	0

Now merge (P_3, P_6)

- distance $[(P_3, P_6), P_1] = \frac{1}{2}[dist(P_3, P_1) + (P_6, P_1)]$
  $= \frac{1}{2}[0.22 + 0.23] = 0.23$.
- distance $[(P_3, P_6), P_2] = \frac{1}{2}[dist(P_3, P_2) + (P_6, P_2)] = \frac{1}{2}[0.15 + 0.25] = 0.20$
- distance $[(P_3, P_6), P_4] = \frac{1}{2}[dist(P_3, P_4) + (P_6, P_4)] = \frac{1}{2}[0.15 + 0.22] = 0.14$.
- distance $[(P_3, P_6), P_5] = \frac{1}{2}[dist(P_3, P_5) + (P_6, P_5)] = \frac{1}{2}[0.28 + 0.39] = 0.34$.

②	P_1	P_2	P_3, P_6	P_4	P_5
P_1	0				
P_2	0.23	0			
P_3, P_6	0.23	0.20	0		
P_4	0.37	0.20	0.19	0	
P_5	0.34	(0.14)	0.34	0.29	0

merge (P_2, P_5)

* $dist[(P_2, P_5), P_1] = \frac{1}{2}[dist(P_2, P_1) + dist(P_5, P_1)]$

$= \frac{1}{2}[0.23 + 0.34]$

$= 0.29$.

* $dist[(P_2, P_5), (P_3, P_6)] = \frac{1}{2}[dist(P_2, (P_3, P_6)) + (P_5, (P_3, P_6))]$

$= \frac{1}{2}[0.20 + 0.34] = 0.27$

* $dist[(P_2, P_5) * P_4] = \frac{1}{2}[dist(P_2, P_4) + dist(P_5, P_4)] = \frac{1}{2}[0.20 + 0.29]$

$= 0.25$

③	P_1	$P_2 P_5$	$P_3 P_6$	P_4
P_1	0			
$P_2 P_5$	0.29	0		
$P_3 P_6$	0.23	0.27	0	
P_4	0.37	0.25	(0.19)	0

merge $[(P_3, P_6), P_4]$

* $dist[(P_3, P_6, P_4), P_1] = \frac{1}{2}[dist((P_3, P_6), P_1) + dist(P_4, P_1)]$

$= \frac{1}{2}[0.23 + 0.37] = 0.30$

* $dist[(P_3, P_6, P_4), (P_2, P_5)] = \frac{1}{2}[dist((P_3, P_6), (P_2, P_5)) + dist(P_4, (P_2, P_5))]$

$= \frac{1}{2}[0.27 + 0.25] = 0.26$

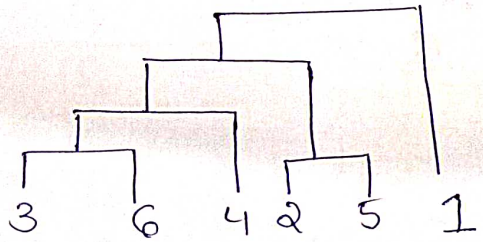
④	P_1	$P_2 P_5$	$P_3 P_6 P_4$
P_1	0		
$P_2 P_5$	0.29	0	
$P_3 P_6 P_4$	0.30	(0.26)	0

merge $(P_2P_5, P_3P_6P_4)$

* dist $(P_2P_5P_4P_3P_6, P_1) = \frac{1}{2}[0.29 + 0.30]$

$= 0.30.$

⑤	P_1	$P_2P_3P_4P_5P_6$.
P_1	0	
$P_2P_3P_4P_5P_6$	0.30.	0



14/03/23

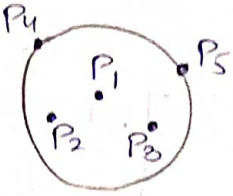
* DBSCAN:-

(Density based spatial clustering of applications with Noise)

→ the main aim of DBSCAN is to create clusters with a minimum size & density.

→ Density is defined as a minimum no. of points with a certain distance of each other.

→ the concept of threshold values & minimum points and epsilon.



P_1 => Center point

ε = 2 units (EPSILON)

minimum points is 4

→ DBSCAN divides the data points into 3 types.

1) Core point

2) Noise point

3) Border point

(1) Core point :-

P_1 can form a cluster by using P_2, P_3, P_4, P_5 then P_1 is called as Core point.

If Minimum points = 6, P_1 cannot form a cluster, then P_1 is not a core point.

then P_1 may be Noise point or Border point.

(2) Noise point :-

P_1 is center point

ε = 2 units

minimum points = 4

As it has only 2 points in its cluster, then it is called as Noise point.

Example:-

Point	X	Y
P_1	3	2
P_2	4	6
P_3	5	5
P_4	6	4
P_5	7	3
P_6	6	2
P_7	7	2
P_8	8	4
P_9	3	3
P_{10}	2	6
P_{11}	3	5
P_{12}	2	4

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
P_1	0											
P_2	1.41	0										
P_3	2.83	1.41	0									
P_4	4.24	2.83	1.41	0								
P_5	5.66	4.24	2.83	1.41	0							
P_6	5.83	4.47	3.16	2.00	1.41	0						
P_7	6.40	5.00	3.61	2.24	1.00	1.00	0					
P_8	5.83	4.47	3.16	2.00	1.41	2.83	2.24	0				
P_9	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P_{10}	1.41	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P_{11}	2.00	1.41	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41	0	
P_{12}	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41	2.00	1.41	0

Epsilon = 1.9

min points = 4

P_1: P_2, P_{10}

P_2: P_1, P_3, P_{11} ✓

P_3: P_2, P_4

P_4: P_3, P_5

P_5: P_4, P_6, P_7, P_8 ✓

P_6: P_5, P_7

P_7: P_5, P_6

P_8: P_5

P_9: P_{12}

P_{10}: P_1, P_{11}

P_{11}: P_2, P_{10}, P_{12} ✓

P_{12}: P_9, P_{11}

Core points ≥ 4 (min pts)

i.e, P_2, P_5, P_{11}

Points	Status	
P_1	Noise	Border
P_2	Core	
P_3	Noise	Border
P_4	Noise	Border
P_5	Core	
P_6	Noise	Border
P_7	Noise	Border
P_8	Noise	Border
P_9	Noise	
P_{10}	Noise	Border
P_{11}	Core	
P_{12}	Noise.	Border

$P_1 = P_3, P_{10}$.

P_1 has a core point so, it is border point

$P_3 = P_3, P_4$.

P_3 is a border point bcoz P_3 is a core point or it has core point.

Eg:

Points	X	Y
P_1	2	10
P_2	2	5
P_3	8	4
P_4	5	8
P_5	~~7~~	5
P_6	6	4
P_7	1	2
P_8	4	9

$\varepsilon = 2$ units

min points = 3

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
P_1	0							
P_2	5	0						
P_3	8.46	6.08	0					
P_4	3.60	4.24	5	0				
P_5	7.07	5	1.41	3.61	0			
P_6	7.21	4.12	2	4.12	1.41	0		
P_7	8.06	3.16	7.28	7.21	~~6.70~~ 6.70	5.39	0	
P_8	2.2	4.47	~~6.40~~	~~4.24~~ 1.41	6.40 5	~~5.81~~ 5.38	7.62	0

P_1:

P_2:

P_3: P_5, P_6 ✓

P_4: P_6

P_5: P_3, P_6 ✓

P_6: P_3, P_5 ✓

P_7:

P_8: P_4

Points	Status
P_1	Noise
P_2	Noise
P_3	Core
P_4	Noise
P_5	Core
P_6	Core
P_7	Noise
P_8	Noise

* **Evaluation of Clustering Algorithms:-**
Clustering can be evaluated based on:
① Clustering tendancy.
② No. of clusters
③ Clustering quality.

https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc

→ **Clustering tendancy:-** if the data doesn't contain clustering tendancy then clusters identified by any clustering algorithms may be irrelavant.
⇒ Non uniform distribution of points in datasets becomes important in clustering.
⇒ Some of the clustering algos like K-means require 'K' as clustering parameter

→ **no. of clusters:-**
⇒ K should not be too high nor too low.
⇒ There are 2 major approaches to find optimal no. of clusters.
① Domain knowledge
② Data driven approach.

* **Domain knowledge:-** On what data we are performing clustering for ex: In dataset, if we have person knowledge of species (virginica, versi color) then K=3.
* **Data driven approach:-** if the domain knowledge is not available, Mathematical methods help in finding out right no. of clusters

→ **Clustering Quality**:- Once clustering is done, how well the clustering is performed can be qualified by a no. of metrics. There are majorly 2 types of measures to access the clustering performance.

① Extrinsic measures
② Intrinsic measures.

⇒ **Extrinsic measures**:- requires ground labels.

Ex: adjusted random index, mutual information based scores, homogenity, completeness, B measures.

⇒ **Intrinsic measures**:- doesn't require ground labels.

Ex: Coefficient, Davis - Bouldin Index

18/3/23 **Internal measures**:-

① Cohesion
② Separation (Coupling).

→ **Cohesion**:- How closely related objects present in a cluster.

→ **Separation**:- Measures how distinct or well separated cluster is from other cluster.

* **Sample**:- Find Cohesion and Coupling for the data points.

Cluster	F_1	F_2	Centroid
C_1	1	0	$\left(\frac{1+1}{2}, \frac{0+1}{2}\right) = (1, 1/2) = (1, 0.5)$
C_1	1	1	
C_2	1	2	$\left(\frac{1+2+2+1}{4}, \frac{2+3+2+2}{4}\right) = \left(\frac{3}{2}, \frac{9}{4}\right)$
C_2	2	3	$= (1.5, 2.25)$
C_2	2	2	
C_2	1	2	
C_3	3	1	$\left(\frac{3+3+2}{3}, \frac{1+3+1}{3}\right) = \left(\frac{8}{3}, \frac{5}{3}\right)$
C_3	3	3	$= (2.67, 1.67)$
C_3	2	1	

Cohesion (C_1):

Euclidian distance.

(1,0) (1,0.5) → 0.5
(1,1) (1,0.5) → 0.5 } = 1

Cohesion (C_2):

(1,2) (1.5,2.25) → 0.55
(2,3) (1.5,2.25) → 0.90
(2,2) (1.5,2.25) → 0.55 } = 2.155
(1,2) (1.5,2.25) → 0.55

Cohesion (C_3):

(3,1) (2.67,1.67) → 0.75
(3,3) (2.67,1.67) → 1.37 } = 3.07
(2,1) (2.67,1.67) → 0.95

Separation (C_1, C_2) : d(Centroid (C_1), Centroid (C_2)) :

(1,0.5), (1.5,2.25) → $\sqrt{(0.5)^2 + (1.75)^2} = 1.82$

Separation (C_2, C_3) :

(1.5,2.25), (2.67,1.67) → $\sqrt{(1.17)^2 + (0.58)^2} = 1.30$

Separation (C_1, C_3) :

(1,0.5), (2.67,1.67) → $\sqrt{(1.67)^2 + (1.17)^2} = 2.03$

Silhouette Coefficient :- It combines ideas of both Cohesion and Separation. But for individual points as well as clusters and clusterings. For an individual point i. Calculate a = average distance of i to the points in its cluster.

Calculate b = minimum average distance of i to points in another cluster.

Then, Silhouette coefficient is given by :

$$S = 1 - \frac{a}{b}$$

Ex:

Point	cluster label.
P_1	1
P_2	1
P_3	2
P_4	2

Compute silhouette coefficient for each point, each of the 2 clusters & overall cluster.

Dissimilarity matrix :-

Points	P_1	P_2	P_3	P_4
P_1	0	0.10	0.65	0.55
P_2	0.10	0	0.70	0.60
P_3	0.65	0.70	0	0.30
P_4	0.55	0.60	0.30	0

Pointwise :

Point P_1 :

$$a = \frac{0.1}{1} = 0.1$$

$$b = \frac{0.65+0.55}{2} = 0.6$$

$$S = 1 - \frac{a}{b} = 1 - \frac{0.1}{0.6}$$

$$= 5/6$$

$$= 0.83$$

Point P_2 :

$$a = \frac{0.1}{1} = 0.1$$

$$b = 0.65$$

$$S = 0.84$$

Point P_3 :

$$a = \frac{0.3}{1} = 0.3$$

$$b = \frac{0.65+0.7}{2}$$

$$= \frac{1.35}{2}$$

$$= 0.67$$

$$S = 1 - \frac{a}{b}$$

$$= 1 - \frac{0.3}{0.67}$$

$$= 0.553$$

Point P_4 :

$$a = 0.3$$

$$b = 0.575$$

$$S = 1 - \frac{a}{b}$$

$$= 0.48$$

Cluster-wise :-

for C1 = Avg of P1 & P2 $= \frac{0.833+0.846}{2} = 0.839$

for C2 = Avg of P3 & P4 $= \frac{0.55+0.48}{2} = 0.517$

overall Silhouette coefficient $= \frac{0.84+0.517}{2} = 0.68$.

* External Measures of cluster Validity :- Purity.

Cluster	entertainment	Financial	foreign	metro	Station	Sport	Purity
1	3	5	40	506	96	27	0.7475
2	4	7	280	29	39	2	0.7756
3	1	1	1	7	4	671	0.9796
4	10	162	3	119	73	2	0.4390
5	33	22	5	70	13	23	0.7134
6	5	358	12	212	48	13	0.5525

Purity: one cluster is dominated by one class.

If a cluster has only 1 class then the purity is 1.

$$\text{Purity} = \frac{\text{max val}}{\text{Sum of all val.}}$$

Purity for cluster1 $= \frac{506}{3+5+40+506+96+27}$

$= \frac{506}{677} = 0.747.$

Purity for cluster 2 :- $\frac{280}{2+4+7+280+89+39+2}$

$= 0.7756$.

Purity for cluster 3: $\frac{671}{1+1+1+7+4+671}$.

$= 0.9796$.

Purity for cluster 4: $\frac{162}{10+162+3+119+73+2}$

$= 0.4390$.

Purity for cluster 5 :- $\frac{70}{33+22+5+70+13+23}$

$= 0.7134$

Purity for cluster 6 :-

$$\frac{358}{5+358+12+212+48+13}$$

$$= 0.5525.$$

* Dimensionality Reduction :-

It refers to the techniques that are used to reduce the no. of input variables in large given data set. while reducing the dimensionality we try to preserve most of the info in the given large data set. that means we remove only unwanted/unimportant features.

Techniques:

1) factor analysis :- It is a variable reduction technique. reduction of set of variables into a small no. of latent factors.
→ unobserved factors (latent factors) account for correlation among observed variables.

Ex: 1 academic ability of a student can be done by
quantative ability
verbal ability

→ quantative ability covers the mathematics, physics, computer programming.

→ verbal ability covers the english, verbal reasoning.

Ex: 2 restaurants are categorised on the basis of 6 variables.

1) Waiting time
2) Cleanliness
3) Staff behaviour
4) taste of food
5) food freshness
6) food temp (hot or cool)

first 3 indicates service, next 3 indicates food quality.

latent variables are variables that are not directly observed but are infered from other variables.

* **Principal component Analysis (PCA):-**

→ A method used to project data in higher dimensional space into a lower dimensional space by maximising the variance of each new dimension.

→ PCA is mostly used as a tool in exploratory data analysis & for making the predictive models.

Ex:- Given following data use PCA to reduce dimension from 2 to 1

Feature	Ex1	Ex2	Ex3	Ex4.
x	4	8	13	7
y	11	4	5	14.

Step1: No. of features = 2 , n=2

No. of samples = 4. , N=4

Step2: computation of mean of variables.

$$\bar{x} = \frac{4+8+13+7}{4} = 8.$$

$$\bar{y} = \frac{11+4+5+14}{4.} = 8.5$$

Step3: computation of co-variance matrix

→ order pairs for xy are

(x,x) (x,y) (y,x) (y,y)

→ covariance of all order pairs

$$Cov(x,x) = \frac{1}{N-1}\sum_{K=1}^{N}(x_i - \bar{x})^2 = \frac{1}{4-1}\left[(4-8)^2+(8-8)^2+(13-8)^2+(7-8)^2\right]$$

$$\boxed{Cov(x,x) = 14}$$

$$Cov(x,y) = \frac{1}{4-1}\left[(4-8)(11-8.5)+(8-8)(4-8.5)+(13-8)(5-8.5)+(7-8)(14-8.5)\right]$$

$$\boxed{Cov(x,y) = -11}$$

$$Cov(y,y) = Cov(x,x)\ \frac{1}{4-1}\left[(11-8.5)^2+(4-8.5)^2+(5-8.5)^2+(14-8.5)^2\right]$$

$$\boxed{Cov(y,y) = 23}$$

Covariance matrix

$$S = \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix} = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}_{2\times 2}$$

Step 4: Eigen value, Eigen vector & Normalized Eigen vector.

(1) Eigen value: $\det(S-\lambda I) = 0$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}_{2\times 2} \Rightarrow \lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$S-\lambda I = \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix}$$

$$\det(S-\lambda I) = \det \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \Rightarrow (14-\lambda)(23-\lambda) - (-11)(-11) = 0.$$

$$\Rightarrow \lambda^2 - 37\lambda + 201 = 0.$$

$$\lambda = 30.3849, 6.6151 \quad \Rightarrow \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

$$\lambda_1 = 30.3849 \qquad \lambda_2 = 6.6151$$

Eigen vector of λ_1 :-

$$(S-\lambda_1 I)\, U_1 = 0 \qquad U_1 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0.$$

$$(14-\lambda_1)u_1 - 11u_2 = 0 \quad \Rightarrow (14-\lambda_1)u_1 = 11u_2$$

$$-11u_1 + (23-\lambda_1)u_2 = 0 \qquad \frac{u_1}{11} = \frac{u_2}{14-\lambda_1} = t$$

when $t = 1$

$u_1 = 11$

$u_2 = 14-\lambda_1$

$\Rightarrow u_2 = 14 - 30.3849$

→ Eigen vector u_1 of λ_1 =

$$= \begin{bmatrix} 11 \\ 14-30.3849 \end{bmatrix}$$

$$= \begin{bmatrix} 11 \\ -16.3849 \end{bmatrix}$$

⇒ Normalize eigen vector u_1

$$e_1 = \begin{bmatrix} \frac{11}{\sqrt{(11)^2+(-16.3849)^2}} \\ \frac{-16.3849}{\sqrt{(11)^2+(-16.3849)^2}} \end{bmatrix} \qquad e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix} \qquad e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 5:- Derive new data set

	Ex1	Ex2	Ex3	Ex4
First PC (PC1)	P_{11}	P_{12}	P_{13}	P_{14}

$$P_{11} = e_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$$

$$P_{11} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

$$P_{11} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

$$P_{11} = -4.305$$

$$P_{12} = \begin{bmatrix} 0.5584 & -0.8303 \end{bmatrix} \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix}$$

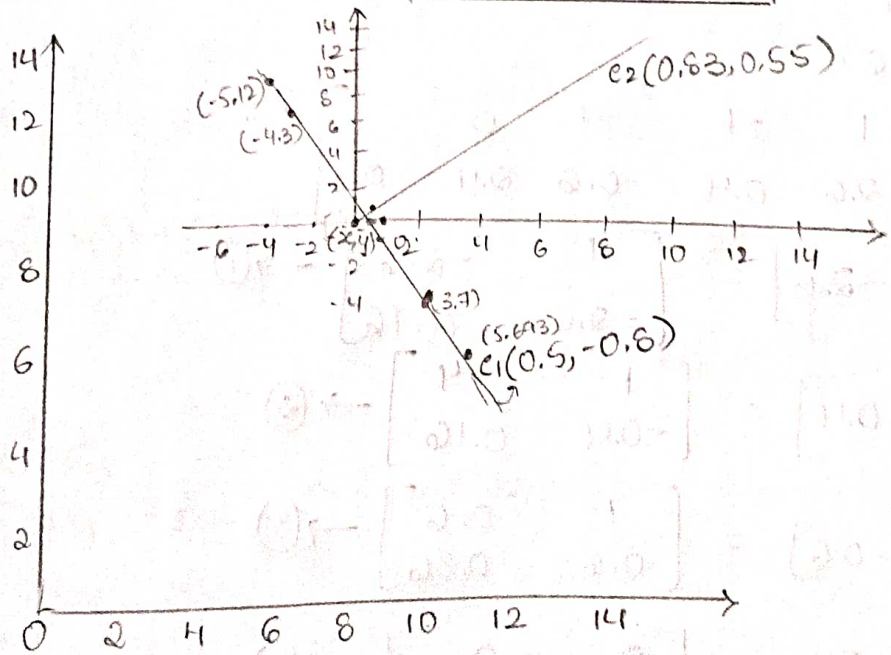
$$P_{12} = 3.736$$

$$P_{13} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} 13-8 \\ 5-8.5 \end{bmatrix} = 5.693.$$

$$P_{14} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} 7-8 \\ 14-8.5 \end{bmatrix} = -5.124.$$

feature

First PC (PC1)	Ex1	Ex2	Ex3	Ex4
	-4.305	3.736	-5.693	-5.124.



* Linear discriminant Analysis (LDA):-

→ LDA is a dimensionality reduction technique used as a pre processing step for pattern classification and ML applications.

→ LDA is similar to PCA (principal component analysis) but LDA finds the axis that maximizes the separation btw multiple classes.

Ex:-

$C_1 = \{x_1 = (x_1, x_2) = \{(4,1)(2,4)(2,3)(3,6)(4,4)\}$

$C_2 = \{x_2 = (x_1, x_2) = \{(9,10)(6,8)(9,5)(8,7)(10,8)\}$

Step 1: Compute within class scatter matrix (S_w)

$$S_w = S_1 + S_2$$

S_1 is the Co-variance matrix for the class C1

S_2 is the Co-variance matrix for the class C2

$$S_1 = \sum_{x \in \omega_1} (x - \mu_1)(x - \mu_1)^T$$

μ_1 is the mean of the class C1

μ_2 is the mean of the class C2

$\mu_1 = (3, 3.6)$

$\mu_2 = (8.4, 7.6)$

$$(x - \mu_1) = \begin{bmatrix} 1 & -1 & -1 & 0 & 1 \\ -2.6 & 0.4 & -0.6 & 2.4 & 0.4 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ -2.6 \end{bmatrix} \begin{bmatrix} 1 & -2.6 \end{bmatrix} = \begin{bmatrix} 1 & -2.6 \\ -2.6 & 6.76 \end{bmatrix} \rightarrow (1)$$

$$\begin{bmatrix} -1 \\ 0.4 \end{bmatrix} \begin{bmatrix} -1 & 0.4 \end{bmatrix} = \begin{bmatrix} 1 & -0.4 \\ -0.4 & 0.16 \end{bmatrix} \rightarrow (2)$$

$$\begin{bmatrix} -1 \\ -0.6 \end{bmatrix} \begin{bmatrix} -1 & -0.6 \end{bmatrix} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 0.36 \end{bmatrix} \rightarrow (3)$$

$$\begin{bmatrix} 0 \\ 2.4 \end{bmatrix} \begin{bmatrix} 0 & 2.4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 5.76 \end{bmatrix} \rightarrow (4)$$

$$\begin{bmatrix}1\\0.4\end{bmatrix}\begin{bmatrix}1 & 0.4\end{bmatrix} = \begin{bmatrix}1 & 0.4\\0.4 & 0.16\end{bmatrix} \rightarrow (5)$$

S_1 = add $\dfrac{(1)+(2)+(3)+(4)+(5)}{5}$ $= S_1 = \begin{bmatrix}0.8 & -0.4\\-0.4 & 0.6\end{bmatrix}$

$$S_2 = \sum_{x=\omega_2} (x-\mu_2)(x-\mu_2)^T$$

$\mu_2 = (8.4, 7.6)$

$$(x-\mu_2) = \begin{bmatrix}0.6 & -2.4 & 0.6 & -0.4 & 1.6\\2.4 & 0.4 & -2.6 & -0.6 & 0.4\end{bmatrix}$$

$$\begin{bmatrix}0.6\\2.4\end{bmatrix}\begin{bmatrix}0.6 & 2.4\end{bmatrix} = \begin{bmatrix}0.36 & 1.44\\1.44 & 5.76\end{bmatrix} \rightarrow (1)$$

$$\begin{bmatrix}-2.4\\0.4\end{bmatrix}\begin{bmatrix}-2.4 & 0.4\end{bmatrix} = \begin{bmatrix}5.76 & -0.96\\-0.96 & 0.16\end{bmatrix} \rightarrow (2)$$

$$\begin{bmatrix}0.6\\-2.6\end{bmatrix}\begin{bmatrix}0.6 & -2.6\end{bmatrix} = \begin{bmatrix}0.36 & -1.56\\-1.56 & 6.76\end{bmatrix} \rightarrow (3)$$

$$\begin{bmatrix}-0.4\\-0.6\end{bmatrix}\begin{bmatrix}-0.4 & -0.6\end{bmatrix} = \begin{bmatrix}0.16 & 0.24\\0.24 & 0.36\end{bmatrix} \rightarrow (4)$$

$$\begin{bmatrix}1.6\\0.4\end{bmatrix}\begin{bmatrix}1.6 & 0.4\end{bmatrix} = \begin{bmatrix}2.56 & 0.64\\0.64 & 0.16\end{bmatrix} \rightarrow (5)$$

$$S_2 = \frac{(1)+(2)+(3)+(4)+(5)}{5} = S_2 = \begin{bmatrix}\frac{9.2}{5} & \frac{-0.20}{5}\\ \frac{-0.20}{5} & \frac{13.2}{5}\end{bmatrix} = \begin{bmatrix}1.84 & -0.04\\-0.04 & 2.64\end{bmatrix}$$

$$S_W = S_1 + S_2 = \begin{bmatrix}0.8 & -0.4\\-0.4 & 2.6\end{bmatrix} + \begin{bmatrix}1.84 & -0.04\\-0.04 & 2.64\end{bmatrix} = \begin{bmatrix}2.64 & -0.44\\-0.44 & 5.28\end{bmatrix}$$

$$S_W = \begin{bmatrix}2.64 & -0.44\\-0.44 & 5.28\end{bmatrix}$$

Step 2: Compute b/w class scatter matrix (S_B)

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$= \begin{bmatrix} -5.4 \\ 4 \end{bmatrix} \begin{bmatrix} -5.4 & 4 \end{bmatrix} = \begin{bmatrix} 29.16 & 21.6 \\ 21.6 & 16.00 \end{bmatrix}$$

Step 3: find the best LDA projection vector

Similar to PCA we find this using eigen vectors having largest eigen value.

$$S_W^{-1} S_B V = \lambda V$$

↳ projection vector

$$\left| S_W^{-1} \quad S_B - \lambda I \right| = 0$$

$$\begin{vmatrix} 11.89 - \lambda & 8.81 \\ 5.081 & 3.76 - \lambda \end{vmatrix} = 0$$

28/5

We get $\lambda = 15.65$

$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = 15.65 \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$

(or)

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = S_W^{-1} (\mu_1 - \mu_2)$$

$$= \begin{bmatrix} 0.1921 & -0.032 \\ -0.032 & 0.08 \end{bmatrix} \begin{bmatrix} -5.4 \\ -4 \end{bmatrix} = \begin{bmatrix} -0.91 & -0.39 \end{bmatrix}^T$$

$$S_W^{-1} = \frac{1}{13.74} \begin{bmatrix} 5.28 & 0.44 \\ 0.44 & 2.64 \end{bmatrix}$$

$$S_W^{-1} = \begin{bmatrix} 0.384 & 0.032 \\ 0.032 & 0.192 \end{bmatrix}$$

Step 4 :- Dimension reduction

$Y = \omega^T x$