ML unit-3

Unsupervised learning
learning from unclassified data
clustering
Hierarchial agglomerative clustering
K-means partitional clustering
Density-Based clustering
Expectation maximization for soft clustering
Dimensionality Reduction
Linear Discriminant Analysis
Principal component Analysis

Unsupervised learning:-

Unsupervised learning is a type of machine learning where the algorithm is trained on data without explicit labels. The goal is to identify underlying patterns, structures or relationships within the data.

learning from unclassified data :-

Objective : Infer the natural structure present within a set of data points.

Key characteristics :

(i) No target labels.

(ii) Models attempt to group similar datapoints together.

Advantages :-

(ii) No need for labeled data
(iii) Discovery of hidden patterns
(iii) Versatility
(iv) Data exploration
(v) Handles complex data

Disadvantages:-

(i) Sensitive to initialization
(ii) Risk of overfitting
(iii) Assumption-dependant
(iv) Computational complexity
(v) Uncertainty in outputs

Applications:-

(i) Data clustering
(ii) Anomaly detection
(iii) Market trend analysis
(iv) Customer segmentation
(v) Dimensionality reduction

Clustering:-

Clustering is a key technique in unsupervised learning that involves grouping a set of objects such that objects in the same group are more similar to each other than those in different groups.

Hierarchical Clustering:-

Hierarchical clustering seeks to build a hierarchy of clusters, i.e a tree-type structure based on similarity.

Types — Agglomerative clustering
Divisive clustering

1. Agglomerative clustering :-
starts with each data point as a separate cluster and successively merges clusters until a single cluster or a stopping criterion is met.

Process :-
(i) Assign each datapoint to its own cluster.
(ii) Find the closest pair of clusters and merge them.
(iii) Repeat until a single cluster remains or the desired number of clusters is achieved.

Advantages :-
(i) Easy to implement and interpret.
(ii) Does not require the number of clusters to be specified in advance.

Disadvantages :-
(i) Sensitive to noise.
(ii) Computationally expensive for large datasets.

Applications :-
(i) Gene expression analysis
(ii) Social network analysis
(iii) Document clustering

Qus = Consider datapoints

	A	B	C	D	E
A	0	2	6	10	9
B	2	0	5	9	8
C	6	5	0	4	5
D	10	9	4	0	3
E	9	8	5	3	0

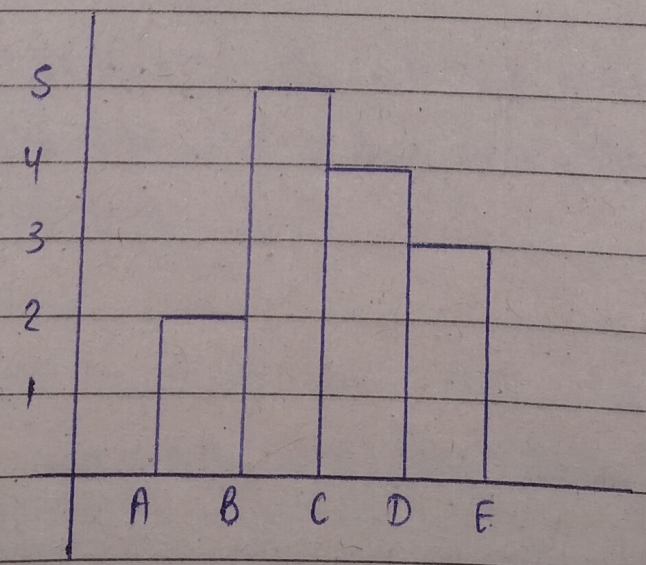
Start with each datapoint as its cluster
{A}, {B}, {C}, {D}, {E}

min distance between A-B = 2
{A, B}, {C}, {D}, {E}

min distance between D-E = 3
{A, B}, {C}, {D, E}

min distance between C-D = 4
{A, B}, {C, D, E}

min distance between B-C = 5
{A, B, C, D, E}



2. Divisive clustering:-
Starts with all data points in one cluster and splits them iteratively into smaller clusters based on centroids.

Centroid is mean or average of all points.

Process:-
(i) Assign all points to a single cluster.
(ii) Split the cluster into smaller clusters based on centroid distances.
(iii) Repeat until desired clusters are formed.

Advantages:-
(i) Provides a global view of the data.
(ii) Can handle different cluster shapes.

Disadvantages:-
(i) Computationally intensive.
(ii) Requires defining stopping criteria

Applications:-
(i) Image segmentation
(ii) Retail customer segmentation
(iii) Financial risk analysis.

Qwy consider datapoints — 2, 4, 5, 9, 12

All points in a single cluster
{2, 4, 5, 9, 12}

→ Let's assume centroids, c1 = 2
c2 = 12

minimum centroid distance

	2	12
2	0 ✓	10
4	2 ✓	8
5	3 ✓	7
9	7	3 ✓
12	10	0 ✓

clusters = {2, 4, 5}, {9, 12}

→ New clusters - centroids, $c1 = \frac{2+4+5}{3} = \frac{11}{3}$

$c2 = \frac{9+12}{2} = \frac{21}{2}$

minimum centroid distance

	11/3	21/2
2	5/3 ✓	8.5
4	1/3 ✓	6.5
5	4/3 ✓	5.5
9	16/3	1.5 ✓
12	25/3	1.5 ✓

clusters = {2, 4, 5}, {9, 12}
Stop as clusters didn't change.

k-means partitional clustering :-

k-means partitional clustering is a type of clustering that divides data into k-clusters where each data point belongs to the cluster with the nearest mean or centroid.

Process :-

(i) Initialize k-centroids randomly.
(ii) Assign each data point to the nearest centroid.
(iii) Recalculate the centroids based on points assigned to each cluster.
(iv) Repeat steps 2 and 3 until centroids no longer change or a maximum number of iterations is reached.

Advantages :-

(i) Simple and fast for large datasets.
(ii) Works well when clusters are spherical and equally sized.

Disadvantages :-

(i) Requires specifying the number of clusters (k) in advance.
(ii) Sensitive to initial centroid placement.
(iii) Poor performance with non-spherical clusters or varying densities.

Applications :-

(i) Market Segmentation
(ii) Image compression
(iii) Document clustering

Ques Consider datapoints
(2,10), (2,5), (8,4), (5,8), (7,5), (6,4)
and k = 2

→ Let is assume centroids, c1 = (2,10)
c2 = (5,8)

minimum centroid distance

	(2,10)	(5,8)
(2,10)	0 ✓	3.6
(2,5)	5	4.2 ✓
(8,4)	8.5	5 ✓
(5,8)	3.6	0 ✓
(7,5)	7.1	3.6 ✓
(6,4)	7.2	4.1 ✓

clusters = {(2,10)}, {(2,5), (8,4), (5,8), (7,5), (6,4)}

→ New clusters-centroids, c1 = (2,10)
c2 = (5.6, 5.2)
as $\left(\frac{2+8+5+7+6}{5} = 5.6 \;,\; \frac{5+4+8+5+4}{5} = 5.2\right)$

minimum centroid distance

	(2,10)	(5.6,5.2)
(2,10)	0 ✓	6
(2,5)	5	3.6 ✓
(8,4)	8.5	2.73 ✓
(5,8)	3.6	2.86 ✓
(7,5)	7.1	1.41 ✓
(6,4)	7.2	1.26 ✓

clusters = {(2,10)}, {(2,5), (8,4), (5,8), (7,5), (6,4)}
Stop, we have 2 clusters that didn't change

DELTA Pg No.

DB-SCAN clustering:-

Density Based Spatial clustering of Application with Noise

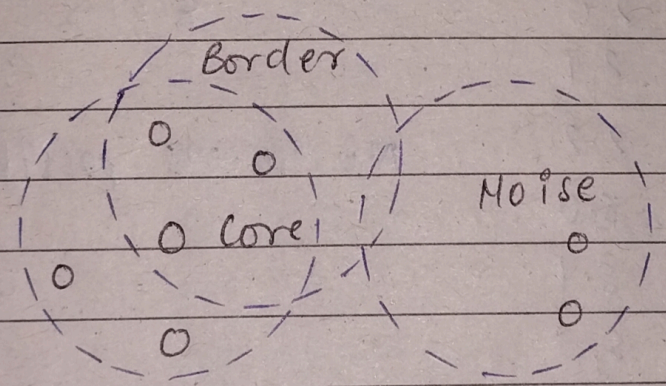
This is one of the most commonly used density based clustering. It forms clusters based on the density of points in a region, identifying 3-points —

(i) Core points
(ii) Border points
(iii) Noise

Inputs:-

(i) minimum-points (MinPts)
The minimum number of points required to form a dense area.

(ii) Epsilon (ε)
The maximum radius of the neighbourhood around a point.



Process:-

(i) Define MinPts and ε
(ii) Identify core points (points with atleast MinPts neighbours within ε).
(iii) Expand clusters from core points.
(iv) Mark points that are not part of any cluster as noise.

Advantages:-

(i) Can find arbitrarily shaped clusters.
(ii) Handles noise effectively.
(iii) Does not require specifying the number of clusters in advance.

Disadvantages:-

(i) Sensitive to the choice of ε and MinPts.
(ii) Struggles with varying density clusters.

Applications:-

(i) Geographic data analysis
(ii) Noise filtering
(iii) Image segmentation

Ques: Given a dataset with following points in 2D-space

Points	x	y
P1	2.0	2.0
P2	2.1	2.1
P3	1.9	2.2
P4	5.0	5.0
P5	5.2	5.1
P6	5.1	5.3
P7	8.0	1.0
P8	8.2	1.1
P9	7.9	1.2
P10	3.5	7.0
P11	3.6	7.1
P12	9.0	9.0

MinPts = 5

$\varepsilon = 1.5$

No clusters possible.

Ans: Given a dataset in the 2D-space

Point	X	Y
P1	1.0	1.0
P2	1.2	1.2
P3	0.8	0.9
P4	3.0	3.0
P5	3.1	3.2
P6	2.9	3.1
P7	8.0	8.0
P8	8.1	8.1
P9	7.9	8.0
P10	10.0	10.0

MinPts = 3
ε = 1.5

Calculate distances and identify the core and border points along with noise.

Core points	Border points
P1	P1, P2, P3
P2	P1, P2, P3
P3	P1, P2, P3
P4	P4, P5, P6
P5	P4, P5, P6
P6	P4, P5, P6
P7	P7, P8, P9
P8	P7, P8, P9
P9	P7, P8, P9
P10	—

{P1, P2, P3}, {P4, P5, P6}, {P7, P8, P9} forms dense clusters.

Core-points : P1, P2, P3, P4, P5, P6, P7, P8, P9

Noise : P10

Expectation maximization for soft clustering:

Soft clustering:
It is a clustering approach where each datapoint is assigned probabilities of belonging to all clusters rather than being assigned to a single cluster (as in hard clustering).

Expectation maximization (EM)
is a way to group datapoints into clusters while allowing each point to belong to multiple clusters to some extent instead of forcing it into just one cluster.

Process:-

(i) Start with an initial guess of cluster parameters (means, variances and weights).
Pretend we know where the clusters are.

(ii) Expectation, E-step
Calculate the probability of each datapoint belonging to each cluster.

(iii) maximization, M-step
Update the cluster parameters to maximize the likelihood of the data. It involves three steps:-

- Update mean –
  Moving the center of each cluster closer to the datapoints assigned to it.
- Update variance –
  Resizing the clusters based on how spread out the data is.

⊙ Update weights –
Adjusting how much weight each cluster has.
(mixing coefficients)

(iv) Repeat step 2 and 3 until convergence.

Advantages :-
(i) Handles overlapping clusters effectively.
(ii) Provides probabilistic cluster membership.

Disadvantages :-
(i) Requires specifying the number of clusters.
(ii) Computationally intensive for large datasets.

Applications :-
(i) Image segmentation
(ii) Anomaly detection
(iii) Bioinformatics

Dimensionality Reduction :-

Dimensionality reduction techniques reduce the number of features in a dataset while preserving as much information as possible.

Types – Linear Discriminant Analysis (LDA)
Principal component Analysis (PCA)

1. Linear Discriminant Analysis (LDA):-

It is a supervised machine learning algorithm used for classification and dimensionality reduction, also known as Normal Discriminant Analysis (NDA).

It is used to project the features from higher dimensional space to lower dimensional space.

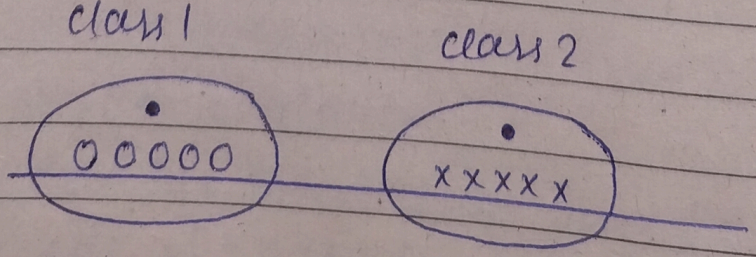
Suppose we have two different datapoints belonging to two different classes that we want to classify.

When the datapoints are plotted on the 2D-plane, there is no straight line that can separate the two classes of the datapoints completely.

LDA uses both x and y-axis to create a new axis and project data onto a new axis in a way to maximize the separation of two categories and hence reducing 2D-graph into a 1D-graph.

Two criterias are used by LDA to create a new axis -

(I) Maximize the distance between means of the two classes.
(II) Minimize the variance within each class.



Process :-

(II) Compute the class mean of dependent variable

$$\mu_1 = \frac{1}{N_1}\left(\sum_{n \in w_1} n\right) \qquad \mu_2 = \frac{1}{N_2}\left(\sum_{n \in w_2} n\right)$$

(III) Derive the covariance matrix of class variable

$$S_1 = \frac{1}{(N_1-1)} \sum_{n \in w_1} (n-\mu_1)(n-\mu_1)^T$$

$$S_2 = \frac{1}{(N_2-1)} \sum_{n \in w_2} (n-\mu_2)(n-\mu_2)^T$$

(IIII) Compute the within class scatter matrix

$$S_w = S_1 + S_2$$

(IV) Compute the between class scatter matrix

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

(V) Compute the eigen values and eigen vectors from the within class and between class

$$S_w + S_B = \lambda w$$

(VI) Sort the values of eigen values and select the top k-values

(VII) Find the eigen vectors corresponding to the top k-eigen values

$$(S_w + S_B - \lambda I)\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = 0$$

(VIII) Obtain the LDA by taking the dot product of eigen vectors and original data.

Advantages:-

(i) Effective for classification tasks.
(ii) Reduces overfitting by reducing dimensions.

Disadvantages:-

(i) Assumes linear separability.
(ii) Sensitive to outliers.

Applications:-

(i) Face recognition
(ii) Text classification
(iii) Medical diagnosis

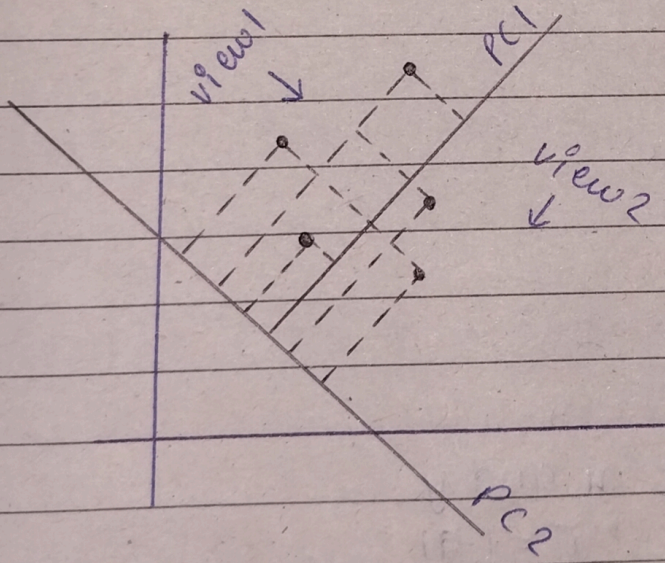
2. Principal Component Analysis (PCA):-

It is an unsupervised dimensionality reduction technique that transforms data into a new coordinate system by identifying the directions (principal components) of maximum variance.

It is used to transform the features from high dimensionality to low dimensionality.

Process:-

(i) Standardize the data (compute means).
(ii) Compute the covariance matrix
(iii) Calculate eigen-values and eigen-vectors of the covariance matrix.
(iv) Select the top k-eigen vectors to form the new feature space.
(v) To project the data onto the new feature space.



Number of Principal Components can be less than or equal to the number of features.

PC1 should be independent of PC2.

Advantages:-

(i) Captures the most important variance in the data.
(ii) Helps visualize high-dimensional data.

Disadvantages:-

(i) Sensitive to scaling of data.
(ii) May lose interpretability of features.

Applications:-

(i) Image compression
(ii) Gene expression analysis
(iii) Recommender systems

Ques

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

(i) Calculate mean of x and y

$N = 10 \quad \bar{x} = 1.81 \quad \bar{y} = 1.91$

(ii) Find covariance matrix

$$cov(x,y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(N-1)}$$

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix}$$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(x-\bar{x})$	$(y-\bar{y})(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
2.5	2.4	0.69	0.49	0.4761	0.2401	0.3381
0.5	0.7	-1.31	-1.21	1.7161	1.4641	1.5851
2.2	2.9	0.39	0.99	0.1521	0.9801	0.3861
1.9	2.2	0.09	0.29	0.0081	0.0841	0.0261
3.1	3.0	1.29	1.09	1.6641	1.1881	1.4061
2.3	2.7	0.49	0.79	0.2401	0.6241	0.3871
2	1.6	0.19	-0.31	0.0361	0.0961	-0.0589
1	1.1	-0.81	-0.81	0.6561	0.6561	0.6561
1.5	1.6	-0.31	-0.31	0.0961	0.0961	0.0961
1.1	0.9	-0.71	-1.01	0.5041	1.0201	0.7171

$$cov(x,x) = \frac{\Sigma(xi-\bar{x})(xi-\bar{x})}{N-1} = \frac{5.549}{9} = 0.6165$$

$$cov(y,y) = \frac{\Sigma(yi-\bar{y})(yi-\bar{y})}{N-1} = \frac{6.449}{9} = 0.7165$$

$$cov(x,y) = cov(y,x) = \frac{\Sigma(xi-\bar{x})(yi-\bar{y})}{N-1} = \frac{5.539}{9} = 0.6154$$

$$C = \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

(iii) Find eigen values for covariance matrix

$C - \lambda I = 0$ characteristic eqn

C = covariance matrix
λ = eigen value
I = identity matrix

$$\begin{bmatrix} 0.6165-\lambda & 0.6154 \\ 0.6154 & 0.7165-\lambda \end{bmatrix} = 0$$

find determinant

$$(0.6165-\lambda)(0.7165-\lambda) - (0.6154)^2 = 0$$
$$\lambda^2 - 1.333\lambda + 0.0630 = 0$$

eigen values, $\lambda_1 = 0.0490$
$\lambda_2 = \theta\ 1.2840$

(iv) Find eigen vector for both eigen values

$\lambda_1 = 0.0490$, $CV = \lambda_1 V$

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0.0490 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$0.6165x_1 + 0.6154y_1 = 0.0490x_1$
$0.6154x_1 + 0.7165y_1 = 0.0490y_1$

both yields same answer

$x_1 = -1.0844y_1$ $\quad x_1^2 + y_1^2 = 1$ (normalization)

$(-1.0844y_1)^2 + y_1^2 = 1$

$y_1^2 = 0.4595$

$y_1 = 0.6778$

$x_1 = -0.7350$

$\lambda_2 = 1.2840$

$CV = \lambda_2 V$

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 1.2840 \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

$0.6165 x_2 + 0.6154 y_2 = 1.2840 x_2$

$0.6154 x_2 + 0.7165 y_2 = 1.2840 y_2$

Both yields same answer

$x_2 = 0.9219 y_2$

$x_2^2 + y_2^2 = 1$ (normalization)

$(0.9219 y_2)^2 + y_2^2 = 1$

$y_2^2 = 0.5405$

$y_2 = 0.7351$

$x_2 = 0.6776$

Principal component with greater λ is considered,

i.e $\lambda_2 = 1.2840$

and principal component,

i.e $x_2 = 0.6776$

$y_2 = 0.7351$