📘 Machine Learning Unit IV

🎯 PYQ-Based Important Questions – Detailed Exam Answers

Q1. Explain K-Means Clustering

Definition

K-Means Clustering is an unsupervised learning algorithm used to divide data into **K number of clusters** based on similarity.

Easy Explanation

K-Means groups similar data points together.
Each group has a center point called **centroid**.

Example:

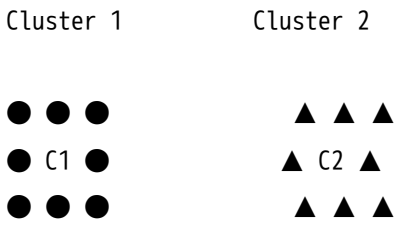
```
Students with similar marks → One cluster
Customers with similar shopping habits → One cluster
```

Working Steps

1. Choose the number of clusters `K`.
2. Select `K` random centroids.
3. Assign each data point to the nearest centroid.
4. Recalculate the centroid of each cluster.

5. Repeat the process until centroids do not change much.

Diagram



Advantages

- Simple and easy to understand
- Fast for large datasets
- Useful for grouping similar data
- Easy to implement

Disadvantages

- Need to choose value of K manually
- Sensitive to outliers
- Initial centroid selection affects result
- Works mainly for numerical data

Applications

- Customer segmentation
- Image compression
- Market analysis
- Document clustering
- Pattern recognition

Conclusion

K-Means is a simple and important clustering algorithm that groups similar data points around centroids.

Q2. Explain PCA

Full Form

PCA stands for **Principal Component Analysis**.

Definition

PCA is a dimensionality reduction technique used to reduce the number of features while keeping important information.

Easy Explanation

Sometimes data has many features. All features may not be useful.
PCA removes less important features and keeps the most important features.

Example:

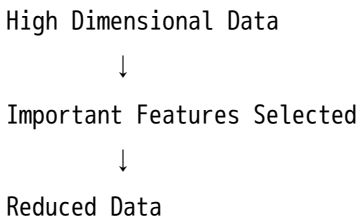
```
Original Features:
Height, Weight, Age, Marks, Attendance, Study Hours

After PCA:
Important Combined Features
```

Working of PCA

1. Take high-dimensional data.
2. Find important directions of maximum variance.
3. Select principal components.
4. Convert original data into fewer dimensions.
5. Use reduced data for analysis.

Diagram



Important Terms

Term	Meaning
Principal Component	Important direction in data
Variance	Spread of data
Dimensionality Reduction	Reducing number of features

Advantages

- Reduces data complexity
- Improves computation speed
- Removes redundant features
- Helps visualization
- Reduces storage requirement

Disadvantages

- Some information may be lost

- New features may be hard to understand
- Works better with numerical data

Applications

- Image compression
- Face recognition
- Data visualization
- Feature extraction
- Noise reduction

Conclusion

PCA is an important technique used to reduce features and improve machine learning performance.

Q3. Explain Hierarchical Clustering

Definition

Hierarchical Clustering is an unsupervised learning method that creates clusters in the form of a hierarchy or tree structure.

Easy Explanation

It does not directly divide data into fixed groups.
Instead, it builds clusters step by step.

Types of Hierarchical Clustering

Type	Meaning
Agglomerative	Bottom-up approach

Type	Meaning
Divisive	Top-down approach

1. Agglomerative Clustering

Each data point starts as a separate cluster.

Nearest clusters are merged step by step.

```
Small Clusters → Bigger Clusters → Final Cluster
```

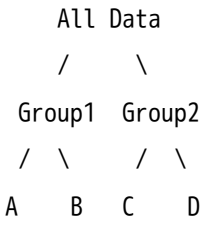
2. Divisive Clustering

All data points start in one big cluster.

Then the cluster is divided into smaller clusters.

```
One Big Cluster → Smaller Clusters → Final Groups
```

Diagram



Advantages

- Easy to understand
- No need to decide number of clusters initially
- Gives tree-like structure
- Useful for small datasets

Disadvantages

- Slow for large datasets
- Once merging/splitting is done, it cannot be easily changed
- Sensitive to noise

Applications

- Document clustering
- Biology classification
- Social network analysis
- Customer grouping

Conclusion

Hierarchical Clustering creates clusters in tree form and is useful for understanding data relationships.

Q4. Differentiate AGNES and DIANA

AGNES Definition

AGNES stands for **Agglomerative Nesting**.
It is a bottom-up hierarchical clustering method.

DIANA Definition

DIANA stands for **Divisive Analysis**.

It is a top-down hierarchical clustering method.

Difference Table

AGNES	DIANA
Full form is Agglomerative Nesting	Full form is Divisive Analysis
Bottom-up approach	Top-down approach
Starts with individual data points	Starts with one large cluster
Merges nearest clusters	Splits large cluster into smaller clusters
Simple and commonly used	More complex
Good for small datasets	Useful when natural large groups exist

AGNES Working

```
Point 1  Point 2  Point 3
    ↓
Small Clusters
    ↓
Merged Cluster
```

DIANA Working

```
One Large Cluster
        ↓
Two Smaller Clusters
        ↓
Final Clusters
```

Advantages of AGNES

- Simple to understand
- Easy to implement
- Good for small datasets

Advantages of DIANA

- Useful for top-level division
- Gives clear large group separation

Conclusion

AGNES starts from individual points and merges them, while DIANA starts from one large cluster and divides it into smaller clusters.

Q5. Explain EM Algorithm

Full Form

EM stands for **Expectation Maximization**.

Definition

Expectation Maximization is an iterative algorithm used to estimate unknown parameters in statistical models.

Easy Explanation

EM is used when some data or information is hidden.
It repeatedly guesses missing information and improves the model.

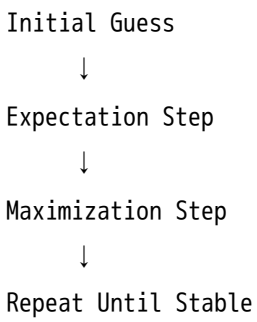
Main Steps of EM

Step	Meaning
E-Step	Estimate hidden values
M-Step	Update model parameters

Working

1. Start with initial guess of parameters.
2. E-Step: Estimate probability of hidden variables.
3. M-Step: Update parameters using estimated values.
4. Repeat until result becomes stable.

Diagram



Example

In clustering, we may not know which point belongs to which cluster.
EM estimates probability of each point belonging to different clusters.

Advantages

- Useful for incomplete data
- Works well in probabilistic models

- Used in soft clustering
- Can improve accuracy step by step

Disadvantages

- Sensitive to initial values
- May get stuck in local optimum
- Can be slow for large data

Applications

- Gaussian Mixture Models
- Image segmentation
- Speech recognition
- Missing data estimation
- Pattern recognition

Conclusion

EM is an important iterative method that estimates hidden information and improves model parameters step by step.

Q6. Explain GMM

Full Form

GMM stands for **Gaussian Mixture Model**.

Definition

Gaussian Mixture Model is a probabilistic clustering method that assumes data is generated from multiple Gaussian distributions.

Easy Explanation

K-Means assigns each point to only one cluster.

But GMM gives probability of a point belonging to each cluster.

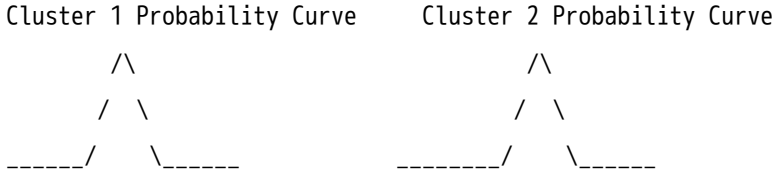
Example:

```
Data Point X:
70% belongs to Cluster 1
30% belongs to Cluster 2
```

Working of GMM

1. Assume data comes from multiple Gaussian distributions.
2. Initialize parameters.
3. Use EM algorithm to estimate cluster probabilities.
4. Assign points based on maximum probability.

Diagram



Advantages

- Supports soft clustering
- Handles overlapping clusters
- More flexible than K-Means

- Gives probability-based output

Disadvantages

- More complex than K-Means
- Requires number of components
- Sensitive to initialization
- Computationally expensive

Applications

- Speech recognition
- Image segmentation
- Pattern recognition
- Anomaly detection

Conclusion

GMM is a flexible probabilistic clustering model that assigns data points to clusters using probabilities.

Q7. Explain SOM

Full Form

SOM stands for **Self Organizing Map**.

Definition

SOM is an unsupervised neural network technique used for clustering and visualization of high-dimensional data.

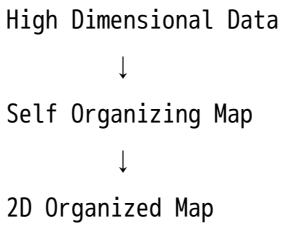
Easy Explanation

SOM converts complex high-dimensional data into a simple low-dimensional map while preserving similarity between data points.

Working of SOM

1. Input data is given to the network.
2. Neurons compete with each other.
3. The closest neuron becomes the winner.
4. Winner neuron and nearby neurons are updated.
5. Similar data points are placed near each other.

Diagram



Advantages

- Good for data visualization
- Handles high-dimensional data
- Useful for clustering
- Automatically organizes similar data

Disadvantages

- Training can be slow
- Difficult to choose map size

- Result interpretation can be difficult

Applications

- Image analysis
- Pattern recognition
- Customer segmentation
- Data visualization
- Medical data analysis

Conclusion

SOM is a neural-network-based clustering method used to organize and visualize high-dimensional data.

Q8. Explain LLE

Full Form

LLE stands for **Locally Linear Embedding**.

Definition

LLE is a non-linear dimensionality reduction technique that preserves local relationships between nearby data points.

Easy Explanation

LLE reduces dimensions, but it tries to keep nearby points close to each other.

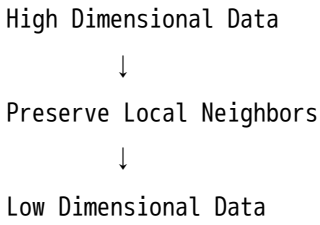
Example:

```
High-dimensional curved data → Low-dimensional simple representation
```

Working of LLE

1. Find nearest neighbors of each data point.
2. Represent each point using its neighbors.
3. Preserve local relationships.
4. Convert data into lower dimensions.

Diagram



Advantages

- Good for non-linear data
- Preserves local structure
- Useful for visualization
- Works better than PCA for curved data

Disadvantages

- Computationally expensive
- Sensitive to number of neighbors
- Not suitable for very noisy data

Applications

- Image recognition
- Face recognition
- Data visualization
- Feature extraction

Conclusion

LLE is a non-linear dimensionality reduction method that keeps nearby data relationships preserved.

Q9. Explain Factor Analysis

Definition

Factor Analysis is a statistical technique used to identify hidden factors that explain relationships among observed variables.

Easy Explanation

Sometimes many visible variables depend on a few hidden reasons.
Factor Analysis finds those hidden reasons.

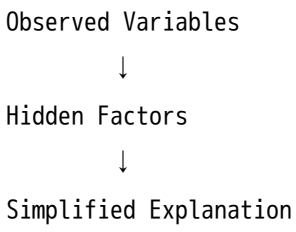
Example:

```
Student marks, attendance, assignments
        ↓
Hidden Factors:
Study habit, intelligence, motivation
```

Working

1. Take many related variables.
2. Find correlation between variables.
3. Identify hidden factors.
4. Represent variables using fewer factors.

Diagram



Advantages

- Reduces number of variables
- Finds hidden relationships
- Helps in data interpretation
- Useful for research analysis

Disadvantages

- Difficult to interpret factors
- Requires expert understanding
- Results depend on quality of data

Applications

- Psychology
- Market research

- Student performance analysis
- Social science research
- Data reduction

Conclusion

Factor Analysis is used to discover hidden factors behind observed variables and reduce data complexity.