Question 1: What is Dimensionality Reduction? Why is it important in Machine Learning? (20 Marks)

Answer:
Dimensionality Reduction refers to the process of reducing the number of input variables (features) in a dataset while preserving the essential structure, patterns, and relationships present in the data. In modern data environments, datasets often contain hundreds or thousands of features, many of which may be redundant, irrelevant, or noisy. Such high-dimensional data causes several problems in machine learning, known as the *"curse of dimensionality."*

Dimensionality reduction aims to transform high-dimensional data into a lower-dimensional representation through mathematical transformations or feature selection processes. It helps maintain key information while simplifying data, which is crucial for visualization, computation, and model performance.

Importance of Dimensionality Reduction:

1. **Reduces Overfitting:**
   High-dimensional datasets often contain redundant or irrelevant variables. Removing them reduces noise, helping models generalize better.
2. **Improves Model Accuracy:**
   With fewer features, models can focus on the most important attributes, improving performance.
3. **Decreases Computational Cost:**
   Lower dimensions require less memory, faster training time, and reduced algorithmic complexity.
4. **Helps in Data Visualization:**
   Dimensionality reduction techniques like PCA and t-SNE allow visualization of high-dimensional data in 2D or 3D form, helping analysts understand clusters and patterns.
5. **Removes Multicollinearity:**
   When features are highly correlated, they distort model learning. Dimensionality reduction resolves this by combining correlated variables.
6. **Enhances Storage Efficiency:**
   Storing reduced feature datasets saves significant space, especially in large-scale systems.
7. **Useful for Noise Reduction:**
   It extracts essential components and discards noise, resulting in cleaner datasets.

In summary, dimensionality reduction is a foundational technique in modern machine learning pipelines because it improves interpretability, efficiency, performance, and scalability.

Question 2: Name and briefly describe three common dimensionality reduction techniques. (20 Marks)

Answer:
Dimensionality reduction techniques fall into two categories—**feature selection** and **feature extraction**. Three widely used techniques include:

---

1. Principal Component Analysis (PCA)

PCA is a linear feature extraction technique that transforms data into new axes called **principal components**. These components capture maximum variance in the data.

Key points:

- Removes redundancy by combining correlated features.
- First component captures maximum variation, second captures remaining variation.
- Useful for visualization of high-dimensional data.
- Widely used in finance, image compression, and pattern recognition.

---

2. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique mainly used for visualization of high-dimensional data into 2D or 3D space.

Key points:

- Preserves local neighborhood structure.
- Creates visually separable clusters.
- Useful in image datasets, genomics, NLP embeddings.
- Captures complex nonlinear relationships unlike PCA.

---

3. Linear Discriminant Analysis (LDA)

LDA is a supervised dimensionality reduction method that focuses on maximizing class separability.

Key points:

- Works by projecting data onto a lower-dimensional space based on maximizing between-class variance and minimizing within-class variance.
- Used in classification tasks like face recognition, speech recognition.

These techniques together address visualization, performance improvement, feature extraction, and improved class separation based on the data type and requirement.

Question 3: What is clustering in unsupervised learning? Mention three popular clustering algorithms. (20 Marks)

Answer:
Clustering is an **unsupervised machine learning** technique that groups data points into clusters based on similarity without using labeled data. The main goal is to identify natural structures, hidden patterns, or meaningful groupings in datasets.

Clustering works by analyzing the distance or similarity between data points using measures like Euclidean distance, cosine similarity, or density estimation. It is widely used for customer segmentation, anomaly detection, image segmentation, market basket analysis, and biological classification.

Why Clustering is Needed:

- Helps detect patterns in unlabeled data.
- Assists businesses in understanding customer behavior.
- Supports data compression and summarization.
- Helps detect unusual observations (anomalies).
- Useful for exploratory data analysis (EDA).

Three Popular Clustering Algorithms:

---

1. K-Means Clustering

A partition-based clustering algorithm that divides data into *K* clusters.

Key features:

- Uses “centroids” representing each cluster.
- Iteratively assigns points to nearest centroid.
- Simple, efficient, scalable.
- Works best with spherical clusters.

---

2. Hierarchical Clustering

Creates a tree-like structure of clusters (dendrogram).

Types:

- **Agglomerative:** bottom-up approach.
- **Divisive:** top-down approach.

Key features:

- No need to specify number of clusters initially.
- Provides deep insights into cluster relationships.

---

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

A density-based algorithm that groups densely packed points.

Key features:

- Excellent for irregular shapes.
- Identifies noise/outliers effectively.
- Works well for spatial datasets.

Clustering remains one of the most fundamental techniques in unsupervised learning for pattern discovery and grouping similar observations.

Question 4: Explain the concept of anomaly detection and its significance. (20 Marks)

Answer:
Anomaly detection refers to identifying data points or events that deviate significantly

from the expected pattern or normal behavior. Anomalies are also called **outliers**, **novelties**, **exceptions**, or **rare events**.

Anomalies may indicate important and actionable insights such as fraudulent transactions, faults in industrial machines, cyber attacks, medical abnormalities, or unusual market behaviors.

Types of Anomalies:

1. **Point Anomalies:** A single value is abnormal.
2. **Contextual Anomalies:** Abnormal only in certain contexts (e.g., temperature in winter).
3. **Collective Anomalies:** A group of readings forms an anomaly (e.g., cyber attack patterns).

Significance of Anomaly Detection:

---

1. Fraud Detection:

Banks use anomaly detection to identify unusual credit card transactions, insurance fraud, or loan fraud.

---

2. Cybersecurity & Intrusion Detection:

Network anomalies can indicate:

- Malware
- Data breach attempts
- Unauthorized login patterns

---

3. Healthcare Monitoring:

Detecting abnormal ECG patterns, unusual symptoms, rare diseases.

---

4. Manufacturing & Predictive Maintenance:

Machines generate sensor readings. Sudden deviations indicate equipment failure, preventing breakdown.

---

5. Stock Market & Financial Alerts:

Price spikes, trading volume anomalies, liquidity issues.

---

6. Business Intelligence:

Unusual customer behavior, churn detection, unusual sales drops.

Due to its broad applicability and the critical impact of anomalies, anomaly detection is considered essential in every modern digital system. Early detection can save costs, prevent losses, and ensure operational safety.

Question 5: List and briefly describe three types of anomaly detection techniques. (20 Marks)

Answer:
Anomaly detection techniques can be grouped under statistical, distance-based, and machine learning categories. Three widely used techniques include:

---

1. Statistical (Probabilistic) Methods

These techniques assume normal data follows a statistical distribution (e.g., Gaussian). Any data point far from mean or below probability threshold is considered anomalous.

Examples:

- Z-score
- Boxplot (IQR method)
- Gaussian distribution models
- Time-series moving averages

Best used for: simple datasets, univariate analysis.

---

2. Distance-Based Methods

These methods use distances between data points. Points far away from the cluster or neighbors are anomalies.

Examples:

- **KNN (K-Nearest Neighbors)** anomaly detection
- **Clustering-based detection** (e.g., DBSCAN)
- **Local Outlier Factor (LOF)**

Advantages:

- Good for multidimensional data
- Works without distribution assumptions

---

3. Machine Learning Based Methods

Modern techniques that learn data patterns and identify unusual behaviors.

Examples:

- **Isolation Forest:** isolates anomalies faster
- **One-Class SVM:** learns boundary around normal data
- **Autoencoders:** reconstruct data and detect high reconstruction errors

Used in:

- Financial Systems
- Cybersecurity
- Industrial IoT sensors
- Healthcare

These three techniques together cover almost all anomaly detection scenarios: statistical distributions, spatial structures, and complex learned behaviors.

Question 6: What is time series analysis? Mention two key components of time series data. (20 Marks)

Answer:
Time series analysis is a method of analyzing data collected over regular intervals of time to identify patterns such as trends, seasonal effects, and cyclic behavior. Unlike regular datasets, time series have a temporal dependency — meaning past values influence future values.

Time series data examples:

- Stock prices
- Weather temperatures
- Electricity consumption
- Website traffic
- Sensor readings

Time series analysis helps in forecasting future values, detecting anomalies, and understanding temporal behavior.

Two Key Components of Time Series Data

1. Trend Component

Refers to the long-term upward or downward movement in data over time.
Examples:

- Long-term increase in population
- Gradual rise in stock index
- Increasing global temperatures

Trends show the overall direction of data movement.

---

2. Seasonal Component

Seasonality refers to repeated patterns or behavior that occur at regular intervals.

Examples:

- Increased retail sales during festivals
- Temperature changes between summer and winter
- Daily electricity usage patterns

Seasonality is predictable and repeats periodically.

Other components include cyclic patterns, irregular variations, and noise, all of which contribute to how time series models behave.

Question 7: Describe the difference between seasonality and cyclic behavior in time series. (20 Marks)

Answer:

1. Seasonality

Seasonality refers to patterns that repeat at *fixed and known intervals*. The period is constant and predictable.

Examples:

- Retail sales increase every December
- Daily website traffic peaks in morning
- Electricity usage increases every evening

Key characteristics:

- Fixed duration
- Predictable repetition
- Driven by weather, festivals, business cycles

---

2. Cyclic Behavior

Cyclic variations refer to long-term fluctuations that do *not follow a fixed pattern* and occur irregularly. These are typically influenced by economic or social factors.

Examples:

- Business cycles: recession → recovery → boom → slowdown
- Market cycles in real estate
- Long-term climate variations

Key characteristics:

- Duration is unpredictable
- Not fixed or regular
- Influenced by external economic/social factors

Difference Summary:

Feature	Seasonality	Cyclic Behavior
Repetition	Fixed & regular	Irregular
Duration	Known (daily/monthly/yearly)	Unknown, long-term
Predictability	High	Low
Caused by	Natural patterns, festivals	Economic & social forces

Understanding this difference is crucial for time series forecasting and anomaly detection.

Question 8: Python code to perform K-Means clustering on a sample dataset. (20 Marks)

Answer:

```
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
import matplotlib.pyplot as plt

# Step 1: Create sample data
X, y = make_blobs(
   n_samples=300,
   centers=4,
   cluster_std=1.2,
   random_state=42
)

# Step 2: Apply K-Means Clustering
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(X)

# Step 3: Cluster labels and centroids
```

```
labels = kmeans.labels_

centroids = kmeans.cluster_centers_


# Step 4: Visualizing the clusters

plt.figure(figsize=(8,6))

plt.scatter(X[:,0], X[:,1], c=labels, s=30)

plt.scatter(centroids[:,0], centroids[:,1], s=200, marker='X')

plt.title("K-Means Clustering Visualization")

plt.xlabel("Feature 1")

plt.ylabel("Feature 2")

plt.show()
```

This program:

- Generates random data using make_blobs
- Applies K-Means clustering
- Plots clusters and centroids
- Displays the final grouping visually

Question 9: What is inheritance in OOP? Provide an example in Python. (20 Marks)

Answer:
Inheritance is a fundamental Object-Oriented Programming concept where one class (child or derived class) acquires attributes and methods from another class (parent or base class). It promotes code reuse, reduces redundancy, and enhances modularity.

Types of Inheritance in Python:

1. Single Inheritance
2. Multiple Inheritance
3. Multilevel Inheritance
4. Hierarchical Inheritance
5. Hybrid Inheritance

Benefits of Inheritance:

- Code reuse
- Improved readability
- Easy maintainability
- Extension of existing functionality
- Supports polymorphism

Example in Python (Multilevel + Method Overriding)

```
# Parent class
class Animal:
    def sound(self):
        return "This animal makes a sound"

# Child class
class Dog(Animal):
    def sound(self):
        return "Dog barks"

# Further derived class
class Puppy(Dog):
    def sound(self):
        return "Puppy makes a soft bark"

# Demonstration
a = Animal()
d = Dog()
p = Puppy()

print(a.sound())  # Parent method
```

print(d.sound()) # Overridden method

print(p.sound()) # Multilevel inheritance

This demonstrates:

- Method overriding
- Multilevel inheritance
- Polymorphic behavior

Question 10: How can time series analysis be used for anomaly detection? (20 Marks)

Answer:
Time series anomaly detection involves identifying unusual patterns or data points that deviate from expected temporal behavior. These anomalies may indicate failures, fraud, cyber-attacks, or unexpected system behavior.

How Time Series Analysis Helps in Anomaly Detection:

---

1. Forecasting Models

Models like ARIMA, SARIMA, Prophet, and LSTM forecast future values.
If actual value deviates significantly from predicted value → anomaly.

Example:

- Electricity consumption suddenly rises
- Stock price spikes sharply compared to forecast

---

2. Seasonal Decomposition

Time series is broken into:

- Trend
- Seasonality
- Residual

Large spikes in residual = anomaly.

---

3. Moving Average & Control Charts

Control charts define upper and lower limits.
Readings outside limits indicate abnormal events.

---

4. Machine Learning Methods

- LSTM autoencoders
- Isolation Forest for time-lagged data
- One-Class SVM

These learn normal time-dependent patterns and detect deviations.

---

5. Change Point Detection

Identifies abrupt changes in mean or variance.
Used in stock markets, climate analysis, and security monitoring.

Applications of Time Series Anomaly Detection:

- Fraud detection in banking
- Server CPU, memory monitoring
- Industrial sensor fault detection
- Predicting machine failures
- Healthcare vital sign monitoring
- Rainfall, weather abnormality detection

Time series analysis is highly effective because it captures temporal dependencies and identifies even subtle irregular patterns that static anomaly detection techniques often miss.
