UNIT-5

Unsupervised Learning

Definition

Unsupervised Learning is a type of machine learning where the model learns patterns from unlabeled data.
There are no target labels, only input data is provided.

Goal:

- Find hidden structure or patterns in the data.
- Examples: grouping customers by purchase behavior, anomaly detection.

Key Types

1. **Clustering** → Group similar data points together.
2. **Association Rule Mining** → Discover relationships between variables in datasets.
3. **Dimensionality Reduction** → Reduce number of features while preserving information (e.g., PCA).

Association Rules

Association Rule Mining is a data mining technique used to discover interesting relationships or correlations between items in large datasets.

Common example: Market basket analysis — "Customers who buy bread also buy butter."

Rule format:

$$X \rightarrow Y$$

- X= Antecedent (if)
- Y= Consequent (then)

Key Metrics

1. **Support:**
   Frequency of occurrence of the itemset.

$$\text{Support}(X \to Y) = \frac{\text{Transactions containing X and Y}}{\text{Total transactions}}$$

2. **Confidence:**
   Likelihood of occurrence of Y when X occurs.

$$\text{Confidence}(X \to Y) = \frac{\text{Support(X} \cup \text{Y)}}{\text{Support(X)}}$$

3. **Lift:**
   Strength of rule over random chance.

$$\text{Lift}(X \to Y) = \frac{\text{Confidence(X} \to \text{Y)}}{\text{Support(Y)}}$$

Types of Association Algorithms

1. Apriori Algorithm

- Concept: Uses frequent itemset generation based on support.
- **Steps:**
    1. Find all frequent 1-itemsets above minimum support.
    2. Generate candidate 2-itemsets and prune those below support.
    3. Repeat until no new frequent itemsets.
    4. Generate rules from frequent itemsets using confidence.

---

2. **Eclat Algorithm**

- Concept: Uses vertical data format (transaction IDs for each item) instead of horizontal.
- **Steps:**

1. Represent itemsets as sets of transaction IDs.
2. Intersect sets to find frequent itemsets.

3. **FP-Growth (Frequent Pattern Growth)**

- Concept: Uses a compact tree structure (FP-tree) to store dataset without candidate generation.
- **Steps:**
    1. Scan dataset → build FP-tree.
    2. Recursively extract frequent patterns from tree.

Cluster Analysis (Clustering)

Definition

Clustering is the task of grouping a set of objects so that objects in the same cluster are more similar than objects in other clusters.

Popular Clustering Methods

1. **K-Means Clustering**
    - Divide data into K clusters.
    - Steps:
        1. Initialize K centroids randomly.
        2. Assign each point to nearest centroid.
        3. Update centroids as mean of points in cluster.
        4. Repeat until convergence.
    - **Distance metric:** Usually Euclidean.
    - **Limitation:** Must choose K, sensitive to outliers.
2. **Hierarchical Clustering**
    - Builds a tree of clusters (dendrogram).
    - Two types:
        - Agglomerative (bottom-up)
        - Divisive (top-down)
    - No need to pre-specify number of clusters.

3. **DBSCAN (Density-Based)**
    - Groups points based on density.
    - Can find arbitrary shaped clusters and identify outliers.

Applications of Clustering

- Customer segmentation
- Document clustering
- Image segmentation
- Anomaly detection

Principal Component Analysis (PCA)

Definition

PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space, preserving most of the variance.

Goal: Reduce features while retaining important information.

Steps in PCA

1. Standardize the data (mean=0, variance=1).
2. Compute the covariance matrix.
3. Calculate eigenvalues and eigenvectors.
4. Sort eigenvectors by descending eigenvalues (variance captured).
5. Select top k principal components.
6. Transform original data into new reduced feature space.

Use Case:

- Reducing dimensionality in image data (e.g., face recognition).
- Noise reduction in datasets.

Random Forests

Definition

Random Forest is an ensemble learning method using multiple decision trees to improve accuracy and reduce overfitting.

- Each tree is trained on a random subset of data (Bootstrap sampling).
- Final prediction = majority vote (classification) or average (regression).

Steps to Build Random Forest

1. **Bootstrap Sampling:** Randomly sample data with replacement.
2. **Tree Construction:** For each tree:
    - Choose a random subset of features at each split.
    - Grow tree fully (no pruning).
3. **Prediction:** Aggregate predictions from all trees.

Advantages

- Reduces overfitting compared to single decision tree.
- Works well on large datasets.
- Handles high-dimensional data.
- Can compute feature importance.

Disadvantages

- Can be slow for large forests.
- Less interpretable than a single decision tree.

Applications of Random Forest

- Credit scoring
- Fraud detection
- Customer churn prediction
- Medical diagnosis