Unsupervised Learning

Dr. Pruthwik Mishra
DoAI, SVNIT Surat

Unsupervised Learning: Overview

- Unsupervised learning finds patterns or structure in data without labeled outputs.
- Goal: Discover hidden representations, groupings, or summaries from raw data.
- Common tasks: Clustering, dimensionality reduction, density estimation.

Key Concepts and Settings

- No explicit supervision or labels; data is only inputs X = {x_1, x_2, ..., x_n}.
- Algorithms learn from intrinsic data structure, not target labels.
- Useful for exploratory analysis, data preprocessing, and feature learning.

Clustering: Definition

- Clustering groups data points based on similarity.
- Objects in same cluster are more similar than those in different clusters.
- Popular algorithms: K-Means, Hierarchical clustering, DBSCAN.

Clustering: K-Means Algorithm

- Objective: Partition n observations into k clusters so that within-cluster variance is minimized.
- Minimize J = sum_{i=1}^{k} sum_{x_j in S_i} lx_j - mu_i|^2, where mu_i is cluster centroid.
- Iterative steps: Assignment and update; repeat until convergence.

Dimensionality Reduction: Definition

- Transformation of high-dimensional data into lower dimensions, retaining structure.
- Uncovers latent variables, removes noise, and enables visualization.
- Main approaches: Principal Component Analysis (PCA), t-SNE, Autoencoders.

Principal Component Analysis (PCA)

- Linear technique projecting data onto directions of maximum variance.
- For data matrix X, compute covariance matrix and eigenvectors.
- Principal components v_1, v_2, ... are eigenvectors with largest eigenvalues.

Unsupervised Learning: Mathematical Setup

- Given data X = {x_1, ..., x_n}, learn a mapping f : X -> structure.
- No labels y; objective typically involves maximizing likelihood or minimizing reconstruction error.
- Metric-based goals, e.g., distance, similarity, or variance.

Applications of Unsupervised Learning

- Market segmentation, customer grouping, anomaly detection.
- Data visualization, compression, feature extraction.
- Preprocessing for downstream supervised tasks (semi-supervised learning).

Challenges and Limitations

- No ground truth: Validation, evaluation can be difficult.
- Algorithm sensitivity to parameters (e.g., number of clusters k).
- Interpretability of discovered patterns and clusters.

Summary: Unsupervised Learning

- Discovers patterns in unlabeled data via clustering, dimensionality reduction, density estimation.
- Opens up exploratory analysis and feature engineering for machine learning workflows.
- Provides foundations for advanced unsupervised and semi-supervised methods.