Unsupervised Learning

A Beginner’s Introduction

Outline

Introduction

What is Unsupervised Learning?

- Learning patterns from **unlabeled** data
- No predefined output or target variable
- Goal: discover hidden structures, groupings, or representations
- Common in exploratory data analysis

Why Unsupervised Learning?

- Labeling data is expensive or impossible
- Understand data distribution and relationships
- Useful for:
  - Clustering customers, documents, images
  - Reducing dimensionality for visualization or speed
  - Detecting anomalies or outliers

Main Tasks

Clustering

- Group similar data points into clusters
- Examples:
  - **k-Means**: partition data into k groups by minimizing within-cluster variance
  - **Hierarchical Clustering**: build a tree of clusters
  - **DBSCAN**: density-based clustering, detects noise

Dimensionality Reduction

- Reduce number of features while preserving structure
- Common methods:
  - **PCA** (Principal Component Analysis): finds directions of maximum variance
  - **t-SNE**, **UMAP**: nonlinear embeddings for visualization
  - **Autoencoders**: neural networks that learn compressed representations

Density Estimation and Anomaly Detection

- Estimate probability distribution of data
- Detect outliers as points in low-density regions
- Examples:
  - Gaussian Mixture Models (GMM)
  - Kernel Density Estimation (KDE)
  - One-class SVM

Popular Algorithms

k-Means Clustering

- Initialize k centroids randomly
- Assign points to nearest centroid
- Update centroids as mean of assigned points
- Repeat until convergence

Principal Component Analysis (PCA)

- Linear projection to lower dimension
- Finds orthogonal directions maximizing variance
- Useful for:
  - Noise reduction
  - Visualization
  - Feature extraction

Challenges and Considerations

Challenges in Unsupervised Learning

- No ground truth for evaluation
- Choosing number of clusters or components
- Sensitivity to initialization and parameters
- Scalability to large datasets
- Interpretability of results

Applications and Tools

Applications

- Customer segmentation
- Document and image organization
- Anomaly detection in fraud, network security
- Data compression and visualization

Popular Libraries

- **Scikit-learn**: clustering, PCA, GMM, DBSCAN
- **TensorFlow/PyTorch**: autoencoders and deep clustering
- **HDBSCAN**, **UMAP** packages for advanced clustering and visualization

Summary

Key Takeaways

- Unsupervised learning finds patterns without labels
- Key tasks: clustering, dimensionality reduction, density estimation
- Many algorithms exist; choice depends on data and goal
- Evaluation is often subjective or uses proxy metrics

Thank you!