Clustering

Course:
INFO-6145 Data Science and Machine Learning

Revised by:
Mohammad Noorchenarboo

October 31, 2024

Contents

Current Section

Supervised vs. Unsupervised Learning

- **Supervised Learning**: Uses labeled data to train a model, where each data point is paired with a known output label.
- **Unsupervised Learning**: Identifies patterns in data without any labels, grouping data based on inherent similarities or distributions.

Example of Supervised Learning

Image classification, where labeled images (e.g., "cat" or "dog") train the model to recognize categories.

Example of Unsupervised Learning

Clustering similar images together without predefined labels, like grouping animals by appearance.

What is Clustering?

Clustering is an unsupervised learning technique that groups data points based on similarity measures.

Key Concept

Clustering algorithms find natural groupings in data, creating clusters where members are more similar to each other than to those in other clusters.

Example

Clustering can group customers based on purchasing patterns to identify market segments without knowing customer profiles beforehand.

Difference from Classification

Unlike classification, clustering does not use labeled data and does not assign specific labels to data points; instead, it organizes data based on patterns or proximity.

Clustering vs. Classification

- **k-means** (Clustering): Partitions data into clusters by finding the nearest centroid. Suitable for unsupervised grouping.
- **k-nearest neighbors (KNN)** (Classification): Assigns labels to data points based on labeled examples, determining categories based on the "k" nearest neighbors.

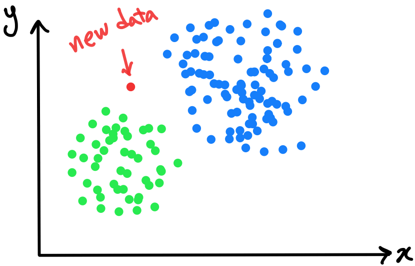
Example of Clustering

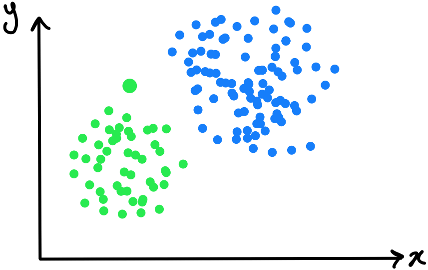
Using k-means to group unlabeled customers by purchasing behaviors.

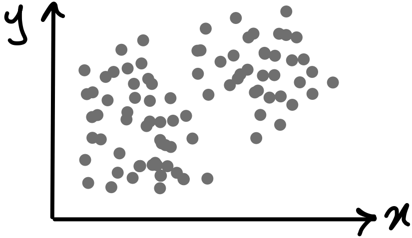
Example of Classification

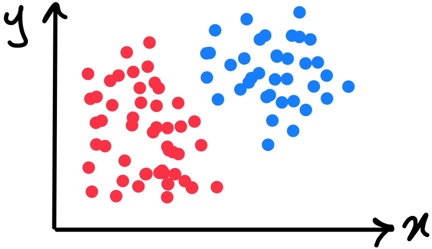
Using KNN to classify new customers as "high-value" or "low-value" based on previous labeled customer data.

Visualization









Applications of Clustering

Clustering is widely used across fields for organizing and analyzing data. Examples include:

- **Market Segmentation**: Identifying groups of customers with similar preferences.
- **Social Network Analysis**: Finding communities within large networks.
- **Search Result Grouping**: Grouping related search results for improved browsing.
- **Medical Imaging**: Segmenting tissues in images to assist in diagnosis.
- **Image Segmentation**: Identifying distinct objects or regions in images.
- **Anomaly Detection**: Detecting unusual patterns in network data.

Common Clustering Algorithms

Different clustering algorithms work well in various situations. Some popular algorithms include:

- **k-means**: Groups data based on nearest centroids, sensitive to the choice of "k" (number of clusters).
- **DBSCAN** (Density-Based Spatial Clustering): Finds clusters based on data density, handling noise well and avoiding assumptions about cluster shape.
- **Gaussian Mixture Modeling (GMM)**: Assumes data points follow a Gaussian distribution, good for clusters with complex, non-linear shapes.
- **Hierarchical Clustering**: Builds a tree of clusters (dendrogram) to reveal hierarchical structures within data, helpful for subcategory analysis.

k-means Clustering

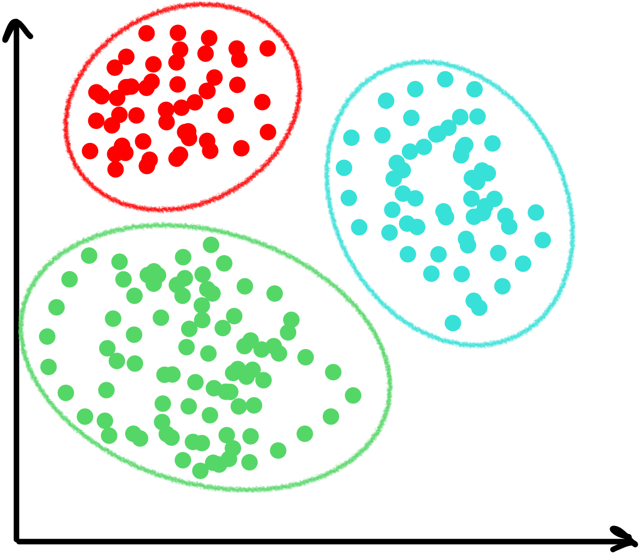
k-means is a popular clustering algorithm that:

- Divides data into a specified number of clusters (**k**).
- Assigns each data point to the nearest cluster center (centroid).
- Updates centroids iteratively until clusters are stable.

Limitations of k-means

- Sensitive to initial centroid positions, which can lead to different results.
- Requires choosing the correct number of clusters, **k.**
- May struggle with data containing outliers or irregular cluster shapes.

Visualization



DBSCAN Clustering

DBSCAN is a density-based clustering algorithm that:

- Groups points closely packed together, leaving outliers or noise unclustered.
- Suitable for clusters of varying shapes and densities.

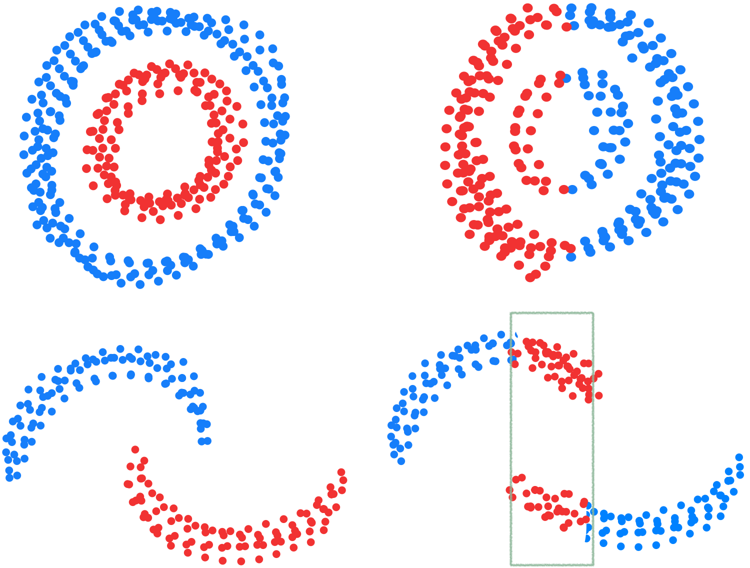
Advantages of DBSCAN

- Works well with noise and outliers.
- Does not require specifying the number of clusters in advance.

Limitations

- Struggles with data containing clusters of similar density.

Visualization



Gaussian Mixture Modeling (GMM)

GMM assumes data is generated from multiple Gaussian distributions, fitting data to these distributions rather than predefined shapes.

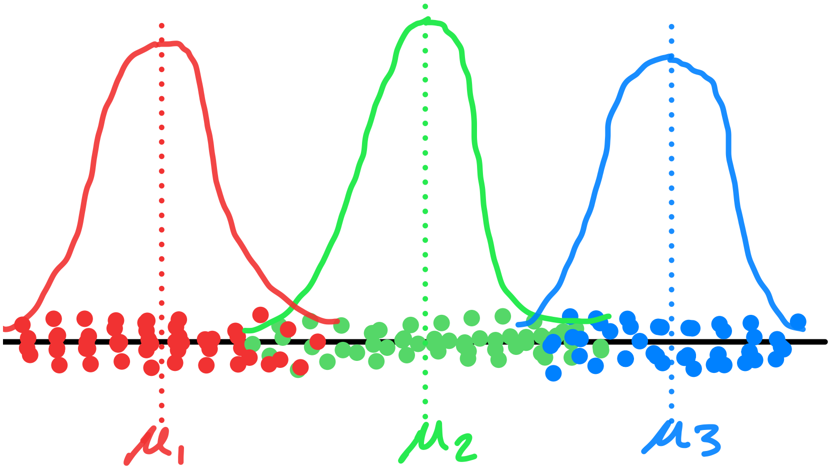
Advantages of GMM

- Flexible in handling clusters of different shapes.
- Suitable for clusters with overlapping areas or non-linear boundaries.

Limitations

- More complex to implement and requires estimating multiple parameters.

Visualization



Hierarchical Clustering

Hierarchical clustering builds clusters in a tree-like structure, which can be visualized with a dendrogram.

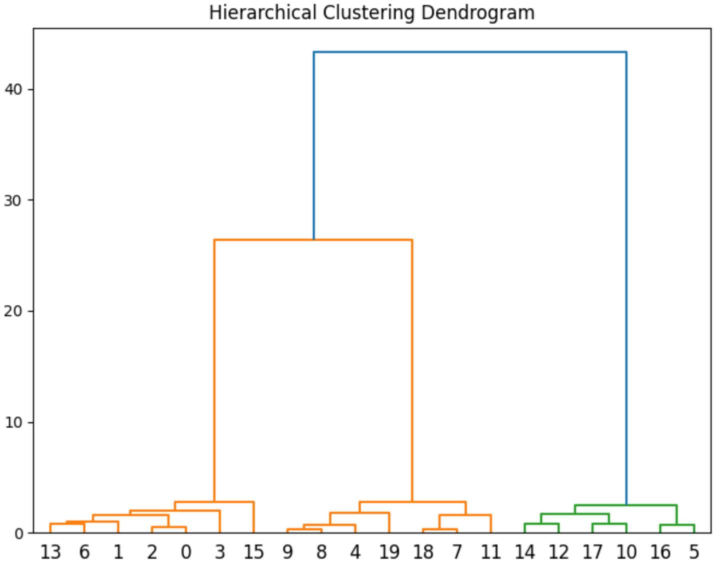
Advantages

- Allows for analyzing clusters within clusters, useful for subcategories.
- Does not require specifying the number of clusters in advance.

Limitations

- Computationally intensive for large datasets.
- Difficult to determine the optimal level to "cut" the dendrogram.

Visualization



Summary: The Power of Clustering

Clustering is a fundamental technique in unsupervised learning that provides:

- Flexible grouping of data without predefined labels.
- Valuable insights for exploratory data analysis and segmenting complex datasets.

Key Takeaway

Clustering enriches the machine learning toolkit by offering ways to understand patterns and organize unlabeled data, with applications across fields.