# Module 4

## Clustering

## 📌 Clustering Strategies and Real-World Applications

### 🔹 What is Clustering?

Clustering is an **unsupervised machine learning technique** that groups **similar data points** into clusters based on their **features and relationships**.

It’s a **pattern discovery technique** that finds **hidden structures** in datasets. It assigns **similar data points to groups (clusters) based on feature similarity**.

✔ Unlike classification, clustering does not require **labeled data**; instead, it **identifies patterns within the dataset** to form natural groupings.  
✔ **Clusters are formed naturally** based on proximity in an **N-dimensional feature space**.  
✔ It is used in applications where patterns exist but predefined labels are unavailable.

Real-World Applications of Clustering.

* **Customer Segmentation** → Grouping customers by **shopping habits** to personalize marketing.
* **Image Segmentation** → Identifying regions in medical imaging (e.g., tumor detection).
* **Anomaly Detection** → Detecting **fraudulent transactions** or **equipment failures**.
* **Genetic Data Analysis** → Grouping similar DNA sequences to study genetic relationships.
* **Data Summarization** → Reducing large datasets by summarizing them into **representative clusters**.

### 🔹 Clustering uses

🚀 Overall, clustering is used to simplify data, reveal hidden relationships, and enhance decision-making across industries.

It’s a powerful tool in data analysis because it reveals patterns and relationships that may not be apparent in raw data, it is widely used for:

✔In **exploratory data analysis**, clustering **uncovers natural groupings**, such as customer segmentation, for targeted marketing.

✔ Boosts **pattern recognition** by grouping similar objects and aiding in image segmentation, such as detecting medical abnormalities.

✔ Clustering helps **anomaly detection** by identifying outliers and detecting fraud or equipment malfunctions.

✔In **feature engineering**, clustering creates new features or reduces dimensionality, improving model performance and interpretability.

✔In **data summarization**, clustering simplifies data by summarizing it into a small number of representative clusters.

✔Clustering **reduces data size** by replacing data points with cluster centers, which is useful for image compression.

## 📌 Types of Clustering Methods

Clustering methods vary in their approach to grouping data points. The three main categories are:

|  |  |  |
| --- | --- | --- |
| Type | How It Works | Best Use Case |
| Partition-Based | Divides data into **K clusters** (fixed number). | Large, well-separated datasets. |
| Density-Based | Groups **high-density regions** and ignores noise. | Data with irregular patterns. |
| Hierarchical | Creates a **tree of nested clusters**. | Small to mid-sized datasets with hierarchical relationships. |

Support Vector Machines (SVMs) are supervised learning algorithms used for both classification and regression, with a primary focus on finding the optimal decision boundary that best separates different classes in a dataset.

✔ SVM is powerful for **high-dimensional** spaces where clear class separation is required.

✔ It **works well for both linearly and non-linearly separable data** by mapping data into higher dimensions using kernel functions.

✔ It aims to maximize the margin between data points belonging to different classes.

Unlike traditional classifiers like KNN or Logistic Regression, SVM does not rely on probability scores but instead finds the most optimal hyperplane that distinctly classifies data points.

### 1. Partition-Based Clustering (K-Means)

Partition-based clustering divides the dataset into **non-overlapping groups**, where **each data point belongs to only one cluster**.

✔ **K-Means** is the most commonly used **partition-based clustering algorithm**.  
✔ It **minimizes intra-cluster variance** by iteratively refining cluster assignments.

Example:

A retail company applies K-Means clustering to segment customers into three groups based on purchasing behavior.

🚀 **Strengths**: Efficient, scalable.  
⚠ **Limitations**: Requires pre-defining **K (number of clusters)**.

### 2. Density-Based Clustering (DBSCAN, HDBSCAN)

Density-based clustering identifies high-density regions and groups points that are closely packed together.

✔ Unlike K-Means, it does not require specifying K (number of clusters).

✔ **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) finds arbitrarily shaped clusters.

✔ **HDBSCAN** (Hierarchical DBSCAN) improves upon DBSCAN by adjusting density thresholds dynamically.

Example:

Detecting fraudulent transactions → Unusual spending patterns (outliers) form separate clusters.

🚀 **Strengths:** Handles noisy data and irregular cluster shapes.

⚠ **Limitations:** Struggles with varying density clusters.

### 3. Hierarchical Clustering (Agglomerative & Divisive)

Hierarchical clustering creates a tree of clusters called a dendrogram, showing the relationships between clusters.

🚀 **Strengths**: Provides interpretable cluster relationships.  
⚠ **Limitations**: Computationally expensive for **large datasets**.

✔ **Agglomerative Clustering (Bottom-Up Approach):**

* Each data point starts as an individual cluster.
* Clusters merge iteratively based on similarity.
* The process continues until a predefined number of clusters is reached.

Steps in Agglomerative Clustering:

1. Compute pairwise distances between data points.
2. Form a distance matrix and merge the closest clusters.
3. Update the distance matrix with new cluster distances.
4. Repeat until the required number of clusters is reached.

✔ **Divisive Clustering (Top-Down Approach):**

* The entire dataset starts as one large cluster.
* Clusters are split recursively until a stopping condition is met.

Steps in Divisive Clustering:

1. Start with one cluster containing all data points.
2. Identify the most dissimilar points and split the cluster.
3. Repeat recursively, continuing to split clusters.
4. Stop when clusters meet a predefined threshold.

✔ **More computationally expensive** than Agglomerative clustering.

✔ **Distance Metrics for Clustering:**

Hierarchical clustering needs two thing in order for be trained:

* A way to measure distance, in order determine the proximity of data points.
* A criterion to merge clusters, this criterion is names **linkage**, and in simple words, is how we obtain the distance to measure the proximity between data points. The distance metrics are:
  + **Single Linkage** → Uses the **closest** data points in two clusters to measure the distance.
  + **Complete Linkage** → Uses the **farthest** points in two clusters.
  + **Average Linkage** → Uses the **mean** distance between clusters.

## 📌 K-Means

K-Means Clustering is a **centroid-based, unsupervised learning algorithm** that partitions a dataset into **K distinct clusters** based on feature similarity. Unlike **supervised learning**, where models learn from labeled data, K-Means identifies **inherent structures** in unlabeled datasets, grouping similar data points together.

### 🔹 Understanding K-Means Clustering

K-Means **divides data into K non-overlapping clusters**, aiming to **minimize intra-cluster variance** while ensuring **maximum dissimilarity between clusters**. Each cluster is represented by a **centroid**, which is iteratively updated to refine cluster assignments.

✔ **Key concepts:**

* + **Centroid** → The **mean position** of all points in a cluster.
  + **Cluster Assignment** → Each data point is assigned to the **nearest centroid** based on distance.
  + **K Value** → The number of clusters, **must be defined before running the algorithm**.

### 🔹 K-Means Algorithm: Step-by-Step

The K-Means algorithm follows an **iterative process** to optimize cluster assignments.

1. Initialize Centroids
   * Choose the **number of clusters (K)**.
   * Randomly **select K initial centroids** from the dataset.
2. Assign Data Points to Nearest Centroids

* Compute **distances from each data point to all centroids** (commonly using **Euclidean distance**).
* Assign each point to the **nearest centroid**.

1. Update Centroids

* Recalculate **the centroid of each cluster** as the mean of all assigned points.

1. Repeat Until Convergence

* Reassign points to the new nearest centroids.
* Continue until centroids no longer move or a maximum number of iterations is reached.

✔ The algorithm converges when centroid positions stabilize or maximum iterations are reached.

### 🔹 Distance Metrics Used Clustering

K-Means Clustering relies on a **distance metric** to determine **which data points belong to which cluster**. The most commonly used metric is **Euclidean distance**, but **other distance measures** can also be used depending on the dataset and its characteristics.

#### ✔ Euclidean Distance (Most Common in K-Means)

Simple and computationally efficient.

Works well when **data is normalized and clusters are spherical**.

Directly measures how close a point is to its centroid.

⚠ **Limitations:**

* Assumes all features contribute **equally** to distance calculations.
* **Sensitive to feature scaling** (features with large values dominate).

**Solution:** Standardize features (e.g., Z-score normalization) before applying K-Means.

**Formula:**

Where:

* *x* and *y* are two points in **n-dimensional space**.
* and are their respective coordinates.
* The sum calculates **squared differences for each feature**, and the square root gives the final distance.

#### ✔ Manhattan Distance (L1 Norm)

* Measures distance as the sum of absolute differences between feature values.
* Preferred when features are highly independent (e.g., city block/grid data).

⚠ **Why K-Means Doesn’t Use It Often?**

* Creates **diamond-shaped clusters** instead of circular ones.
* Less sensitive to outliers but **not ideal for centroid-based algorithms**.

**When to use** -> If data is high-dimensional but sparse (e.g., text data with word counts).

**Formula:**

#### ✔ Cosine Distance (Similarity Measure)

Used for **high-dimensional data** (e.g., text, embeddings, gene expressions).

Measures how "aligned" two vectors are rather than their absolute difference.

Often used in document clustering (e.g., NLP applications).

⚠ **Why K-Means Doesn’t Use by default?**

* Requires **vector normalization** before use.

**When to use** -> If feature values represent **angles/directions** (e.g., word frequency vectors).

**Formula (Cosine similarity):**

Where:

* is the **dot product** of two vectors.
* and are the magnitudes (lengths) of the vectors.

#### ✔ Mahalanobis Distance (Accounts for Feature Correlation)

**Accounts for correlations between features** (unlike Euclidean distance).

Creates **elliptical** rather than circular clusters, making it useful when features are dependent.

⚠ **Why K-Means Doesn’t Use It Often?**

* Requires computing the **covariance matrix**, which is expensive for large datasets.
* Works **poorly with small sample sizes** or highly imbalanced data.

**When to use** -> If **features are correlated** (e.g., financial data, chemical compositions).

**Formula:**

Where:

* is the inverse covariance matrix, adjusting for feature correlations.

#### ✔ Choosing the Right Distance Metric

🚀 **Key Takeaways:**

* **Euclidean Distance** is the default for K-Means because it is computationally simple and aligns well with centroid-based clustering.
* **Manhattan and Cosine Distances** are better for non-Euclidean feature spaces, but they are **not ideal for centroid-based clustering**.
* **Mahalanobis Distance** is powerful for correlated features, but **computationally expensive**.

|  |  |
| --- | --- |
| **Distance Metric** | **Best for...** |
| Euclidean Distance | Well-separated, spherical clusters |
| Manhattan Distance | Grid-like/sparse data (e.g., city distances) |
| Cosine Distance | Text or high-dimensional sparse data |
| Mahalanobis Distance | Correlated features, elliptical clusters |

### 🔹 Challenges and Limitations of K-Means

While K-Means is efficient, it **relies on specific assumptions** that may not hold in all datasets.

✔ **Clusters are Convex and Isotropic**

* K-Means assumes clusters are **spherical** and **well-separated**.
* If clusters have **irregular shapes**, K-Means struggles to separate them properly.

✔ **Equal Cluster Sizes**

* The algorithm assumes **clusters contain a similar number of points**.
* When clusters are **imbalanced**, K-Means may assign more points to the larger cluster, **distorting results**.

✔ **Sensitivity to Initial Centroids**

* Since centroids are **randomly initialized**, different runs may produce **different cluster assignments**.
* **Solution** → The **K-Means++ initialization method** reduces this issue by choosing centroids **far apart** initially.

✔ **Sensitivity to Noise and Outliers**

* K-Means uses **variance as a criterion**, making it sensitive to **noisy data**.
* **Solution →** Preprocess data by **removing outliers** or **scaling features**.

### 🔹 How to Choose the Optimal Number of Clusters (K)?

One major challenge in K-Means is determining the **best K value**. Several techniques help identify the most suitable number of clusters.

#### Elbow Method

The **Elbow Method** plots **inertia (within-cluster sum of squares, WCSS)** as a function of K.

The goal is to **find the "elbow point"**, where adding more clusters **yields diminishing returns**.

✔ **Steps to Use the Elbow Method:**

1. Compute K-Means for **multiple values of K** (e.g., K=1 to K=10).
2. Record the **inertia (sum of squared distances between points and their centroids)**.
3. Plot the **K values vs. inertia** and **identify the "elbow"**, where inertia **stabilizes**.

✔ **Interpretation:**

* + If **K is too small** → Clusters **contain highly dissimilar points**, increasing WCSS.
  + If **K is too large** → Clusters become **too small**, increasing variance unnecessarily.
  + The **best K is where the curve bends** (elbow point).

#### Silhouette Score

Measures how well-separated clusters are by comparing:

* + **Cohesion** (how close a point is to others in clusters).
  + **Separation** (how far it is from points in the nearest cluster).

✔ **Formula:**

Where:

* + **a** = Mean intra-cluster distance (cohesion).
  + **b** = Mean nearest-cluster distance (separation).

✔ **Ranges from -1 to 1:**

* + **Higher scores (close to 1)** → Well-separated clusters.
  + **Low scores (near 0)** → Overlapping clusters.
  + **Negative scores** → Points assigned to the wrong cluster.

#### Davies-Bouldin Index

Measures the **average similarity ratio between each cluster and its most similar cluster.**

Lower values indicate **better clustering performance**

✔ Used in cases where the Elbow Method does not provide a clear K.

## 📌 DBSCAN and HDBSCAN, Density-Based Clustering

Density-based clustering is a powerful method used to **detect clusters of arbitrary shapes** and **identify noise (outliers)** in datasets. Unlike **K-Means**, which assumes spherical clusters and requires a predefined **K**, **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** and **HDBSCAN (Hierarchical DBSCAN)** dynamically adapt to **density variations** in data, making them useful for complex real-world clustering problems.

✔ Why use density-based clustering?

* Handles arbitrarily shaped clusters (e.g., spirals, curves).
* Works well when the number of clusters is unknown.
* Automatically identifies outliers (noise points).
* More robust than K-Means for real-world, noisy datasets.

### 🔹 How DBSCAN works?

DBSCAN groups points into clusters based on **local density** instead of distance from a centroid. It relies on **two parameters** to define cluster formation:

✔ **ε (Epsilon - Neighborhood Radius)** → Defines the **maximum distance** within which points are considered **neighbors**.

✔ **MinPts (Minimum Points)** → The **minimum number of points** required to form a **dense region (core point)**.

DBSCAN labels data points into **three categories**:

* **Core Points ->** Have at least MinPts neighbors within ε radius.
* **Border Points ->** Fall within the neighborhood of a core point but have fewer than MinPts neighbors.
* **Noise (outlier) Points ->** Do not belong to any cluster because they lack enough nearby points.

#### ✔ DBSCAN Algorithm Steps

1. Select an **unvisited point** and check if it is a **core point** (has at least MinPts neighbors).
2. If it’s a **core point**, create a **new cluster** and expand it by adding all density-connected points.
3. If it’s a **border point**, assign it to an existing cluster.
4. If it’s a **noise point**, mark it as an **outlier**.
5. Repeat until all points are **processed**.

⚠ **DBSCAN is a single-pass algorithm** → Once a point is labeled, it does **not change** (unlike K-Means, which iterates).

**Example: Clustering with DBSCAN**:

Imagine a **dataset with two crescent-shaped clusters**. K-Means would fail because **it assumes spherical clusters**, but DBSCAN successfully separates them because it recognizes **dense regions, regardless of shape**

🚀 **Strengths of DBSCAN**

✔ Finds **clusters of arbitrary shapes**.

✔ **No need to predefine K** (like in K-Means).

✔ **Handles outliers** by labeling them as noise.

⚠ **Limitations of DBSCAN**

✖ **Struggles with varying densities** → The same **ε** may be **too small for sparse clusters and too large for dense ones**.

✖ **Sensitive to parameter tuning** → Requires careful selection of **ε and MinPts.**

### 🔹 How HDBSCAN works?

HDBSCAN enhances DBSCAN by **dynamically adjusting the density threshold**, making it more flexible for **datasets with clusters of varying density**.

✔ **No need to predefine ε (radius)** → HDBSCAN **automatically finds the optimal density threshold**.

✔ **Uses cluster stability** → A stable cluster **remains unchanged** when density parameters are adjusted.

✔ **Creates a hierarchy of clusters** → It builds a **tree structure (dendrogram)** and extracts the **most stable clusters**.

#### ✔ HDBSCAN Algorithm Steps

* + 1. Treat each point as a separate cluster (initial state).
    2. **Merge clusters based on density connectivity** (reducing the density threshold).
    3. Build a hierarchy (dendrogram) of clusters.
    4. **Extract the most stable clusters** (those that persist across different density thresholds).

🚀 **Strengths of HDBSCAN**

✔ **No need to specify ε** (better for real-world applications).

✔ **Handles varying densities** in different regions of data.

✔ **Finds more meaningful clusters than DBSCAN**.

⚠ **Limitations of HDBSCAN**

✖ More **computationally expensive** than DBSCAN.

✖ **Cluster stability tuning requires expert interpretation**.

### 🔹 Comparing DBSCAN vs. HDBSCAN

|  |  |  |
| --- | --- | --- |
| Feature | DBSCAN | HDBSCAN |
| Predefined Parameters? | Requires **ε & MinPts** | **No predefined ε** (adaptive density) |
| Handles Noise? | ✅ Yes | ✅ Yes |
| Handles Varying Densities? | ❌ No (fixed density threshold) | ✅ Yes (dynamic density threshold) |
| Cluster Shape Flexibility | ✅ Arbitrary | ✅ Arbitrary |
| Performance | Fast | Slower but more accurate |

🚀 **When to Use Which?**

✔ **Use DBSCAN** when clusters have **similar density** and you can define **ε & MinPts** manually.

✔ **Use HDBSCAN** when clusters have **varying densities** and need **automatic tuning**.