Clustering Algorithm

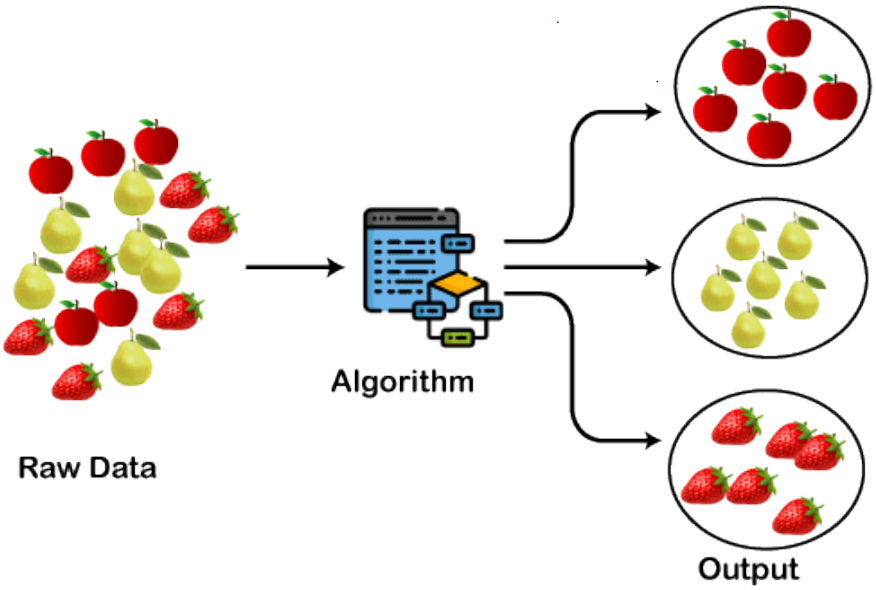
07 October 2025 17:15

1. What is Clustering?

- **Clustering** is an **unsupervised machine learning technique** that groups similar data points into clusters based on their features.
- Unlike classification, **no labels** are provided; the algorithm must **discover hidden patterns or group structures**.

Clustering is the process of grouping similar objects into clusters such that:

- Objects **within** a cluster are **very similar**.
- Objects **across** clusters are **very different**.



2. Applications/use cases:

- Customer segmentation (marketing)
- Image segmentation
- Document/topic grouping
- Anomaly detection
- Gene expression analysis

3. Criterion Functions for Clustering

Criterion functions measure the *quality of clusters*.

A good clustering minimizes:

1. **Intra-cluster distance** (points are close within cluster)
2. **Inter-cluster distance** (clusters are far apart)

Most common criterion function:

Sum of Squared Errors (SSE)

Used in K-Means and other partitional clustering.

For K clusters:

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

Where:

- x_i = data point
- μ_k = centroid of cluster C_k

Goal:

Minimize SSE ⇒ better tightness ⇒ better clustering.

3. 🕸 Types of Clustering

Category	Examples	Description
Partitioning-based	K-Means,	Divide data into *k* disjoint clusters
Hierarchical	Agglomerative, Divisive	Build a tree (dendrogram) of clusters
Density-based	DBSCAN, HDBSCAN	Group points close in dense regions
Model-based	Gaussian Mixture Model (GMM)	Assume data is generated from multiple probability distributions
Graph-based / Spectral	Spectral Clustering	Use graph Laplacian and eigenvectors to form clusters