Understanding Hierarchical Clustering: Comprehensive Overview

1. Model Overview

Hierarchical Clustering is an **unsupervised learning algorithm** that builds a hierarchy (tree-like structure) of clusters. Instead of predefining the number of clusters, it merges or splits clusters step by step to form a **dendrogram**, which visually represents the cluster relationships.

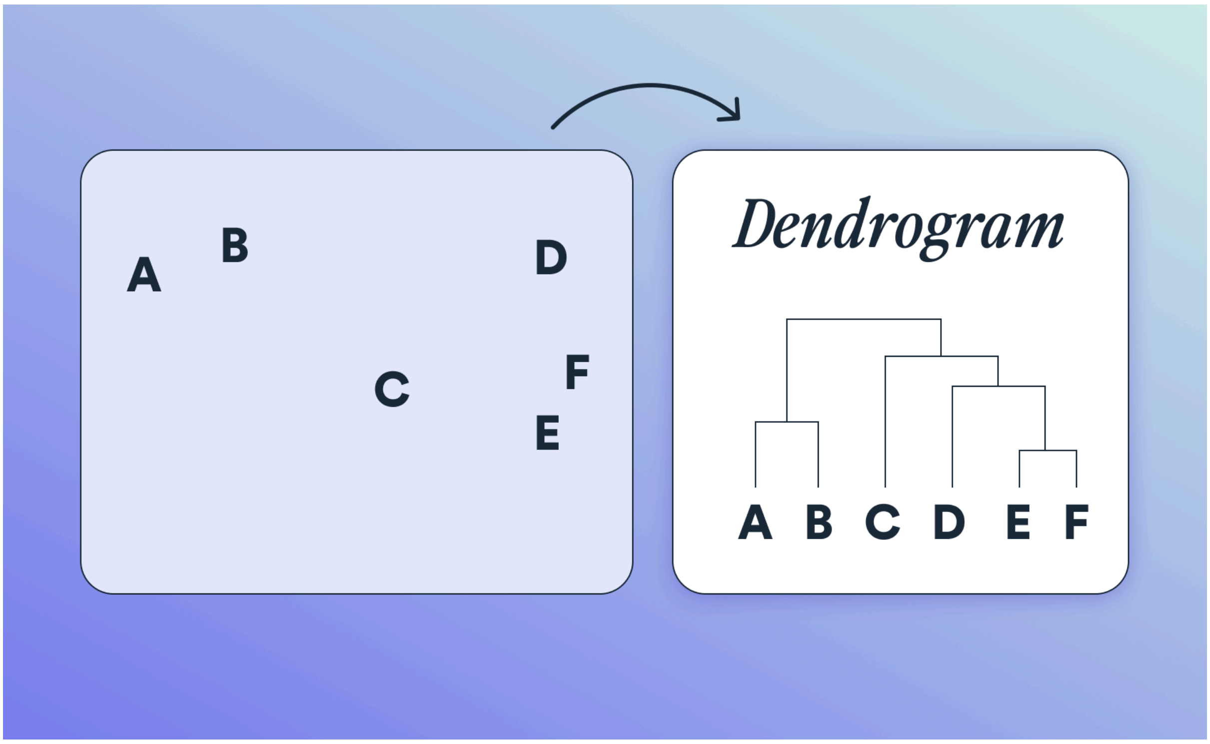
It can be used for **exploratory data analysis** to understand natural groupings and relationships between data points.

---

2. Key Aspects to Analyze for Hierarchical Clustering

A. Assumptions

1. Data contains **hierarchical or nested structures** suitable for tree representation.
2. Similar objects are **close in the feature space** under the chosen distance metric.
3. Clusters can be identified based on **distance thresholds** in the dendrogram.
4. The distance measure and linkage method meaningfully capture similarity.
5. Data should be **scaled**, as features with large magnitudes can dominate distance calculations.



B. Limitations

1. **Computationally expensive** ($O(n^2)$ time complexity).
2. **Not scalable** to very large datasets.
3. **Sensitive to noise and outliers** — a single outlier can affect multiple clusters.
4. **Choice of distance metric and linkage method** greatly affects results.
5. Once a merge or split occurs, it **cannot be undone** (no backtracking).
6. Difficult to handle **high-dimensional data** effectively.

C. Attributes / Input Features

1. Works best with **continuous, numerical** data.
2. Categorical data must be converted to numerical form using **encoding methods**.
3. Requires **feature scaling** (standardization or normalization).

4. Distance metrics assume **comparable feature influence**.

5. Outliers and irrelevant features should be handled before clustering.

D. Internal Model Variations / Subtypes

Hierarchical Clustering can be **Agglomerative** or **Divisive**:

1. **Agglomerative (Bottom-Up)** — starts with each point as its own cluster and merges clusters iteratively until one cluster remains. *(Most common)*

2. **Divisive (Top-Down)** — starts with one large cluster and splits recursively into smaller clusters.

Common **Linkage Methods** (for computing inter-cluster distances):

- **Single Linkage:** Minimum distance between points of two clusters.

- **Complete Linkage:** Maximum distance between points of two clusters.

- **Average Linkage:** Average distance between all pairs of points.

- **Ward's Method:** Minimizes variance within clusters (most robust and widely used).

E. Hyperparameters

1. **Linkage Criterion:** single, complete, average, or ward.

2. **Distance Metric:** Euclidean (default), Manhattan, cosine, etc.

3. **Number of Clusters (cut-off):** chosen by analyzing the dendrogram.

4. **Affinity:** type of distance measure (depends on the linkage method).

5. **Threshold:** maximum allowed distance between merged clusters.

F. Performance Evaluation Metrics

As it's **unsupervised**, internal metrics are used:

- **Silhouette Score** — measures how similar a point is to its cluster vs others.
- **Davies–Bouldin Index** — lower value indicates better clustering.
- **Calinski–Harabasz Index** — higher value indicates better separation.
 If true labels exist:
- **Adjusted Rand Index (ARI)**
- **Normalized Mutual Information (NMI)**

G. Use Cases

- Customer or market segmentation
- Gene expression analysis in bioinformatics
- Document/topic clustering
- Image segmentation
- Social network analysis

H. Optimization Tips

1. **Standardize data** before computing distances.
2. Use **Ward linkage** with Euclidean distance for compact, spherical clusters.
3. Perform **PCA** or feature selection to reduce dimensionality.
4. **Visualize dendrograms** to determine the optimal number of clusters.
5. **Truncate dendrograms** at different heights to analyze cluster granularity.
6. Remove **outliers** that distort hierarchical structure.

I. Model Interpretation

- The **dendrogram** is the core interpretation tool.
- The **height at which clusters merge** reflects dissimilarity between them.
- A **horizontal cut** through the dendrogram defines the number of clusters.
- **Tighter, lower merges** indicate more similar data points.

J. Summary Table

Aspect	Description
Type	Unsupervised Clustering
Approach	Hierarchical (tree-based)
Variants	Agglomerative (bottom-up), Divisive (top-down)
Linkage Methods	Single, Complete, Average, Ward
Key Parameter	Linkage criterion and distance threshold
Distance Metric	Euclidean (default)
Handles Outliers	Poorly
Requires Scaling	Yes
Ideal Dataset Size	Small to medium
Visualization	Dendrogram
Output	Hierarchy of clusters

K. Comparison with K-Means

Feature	K-Means	Hierarchical Clustering
Type	Partitional	Hierarchical
Number of Clusters	Must be pre-specified	Determined from dendrogram
Shape of Clusters	Spherical	Any shape
Scalability	High (large datasets)	Low (small datasets)

Robustness to Noise	Low	Low
Visualization	Centroid plots	Dendrogram
Reproducibility	Depends on initialization	Deterministic (if linkage fixed)