HIERARCHICAL CLUSTERING

1. Introduction

Hierarchical Clustering is an unsupervised machine learning technique used to group similar data points into clusters. It creates a hierarchy (tree-like arrangement) that shows how individual points or groups are related. The tree diagram is called a dendrogram.

- It groups data based on similarity or distance.
- The number of clusters does not have to be fixed before building the hierarchy.
- It is useful for exploring patterns and relationships in data.

2. Types of Hierarchical Clustering

There are two main approaches:

A. Agglomerative Clustering (Bottom-Up)

- Initially, every data point is treated as a separate cluster.
- The distances between clusters are calculated.
- The two closest clusters are merged into one cluster.
- Distances are updated and the merging process is repeated.
- It continues until one cluster remains or the required stopping condition is reached.

B. Divisive Clustering (Top-Down)

- Initially, all data points are placed in one large cluster.
- The large cluster is divided into smaller clusters.
- The splitting process is repeated on the resulting groups.
- It continues until each point is separate or a stopping condition is reached.

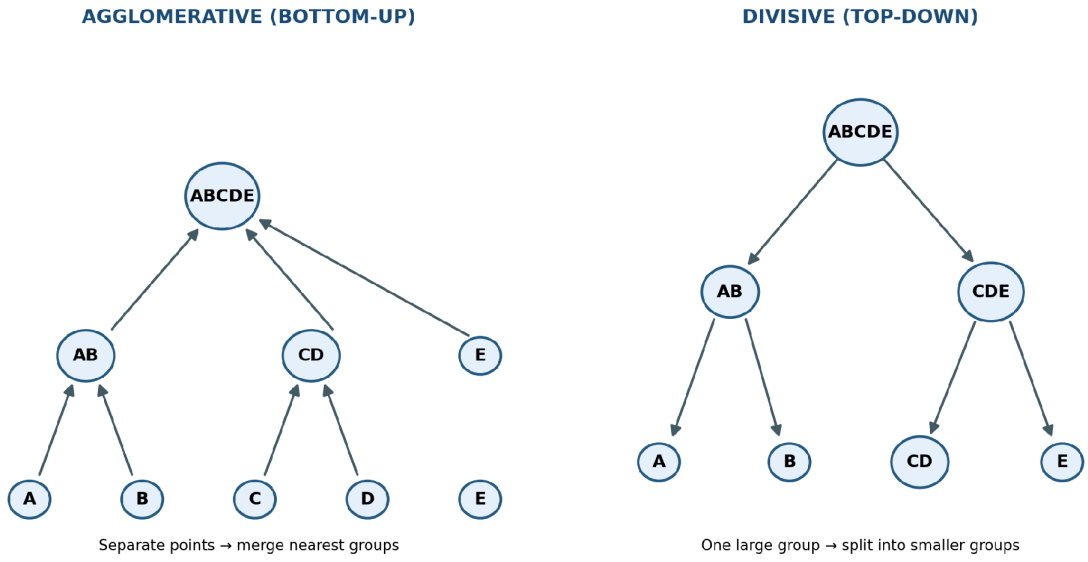


Diagram 1: Agglomerative merges groups from bottom to top; divisive splits from top to bottom.

3. Steps in Agglomerative Clustering

1. Start with each data point as an individual cluster.
2. Calculate the distance between every pair of clusters.
3. Find and merge the two closest clusters.
4. Update the distance matrix using the selected linkage method.
5. Repeat the merging and updating steps.
6. Stop when one cluster remains or the desired number of clusters is obtained.

4. Dendrogram

A dendrogram is a tree-like diagram that displays the order in which clusters are merged or divided. Each leaf represents an original data point. The height of a joining line represents the distance (or dissimilarity) at which clusters are joined.

- A low joining height indicates that groups are more similar.
- A high joining height indicates that groups are less similar.
- A horizontal cut across the dendrogram can be used to obtain a chosen set of clusters.

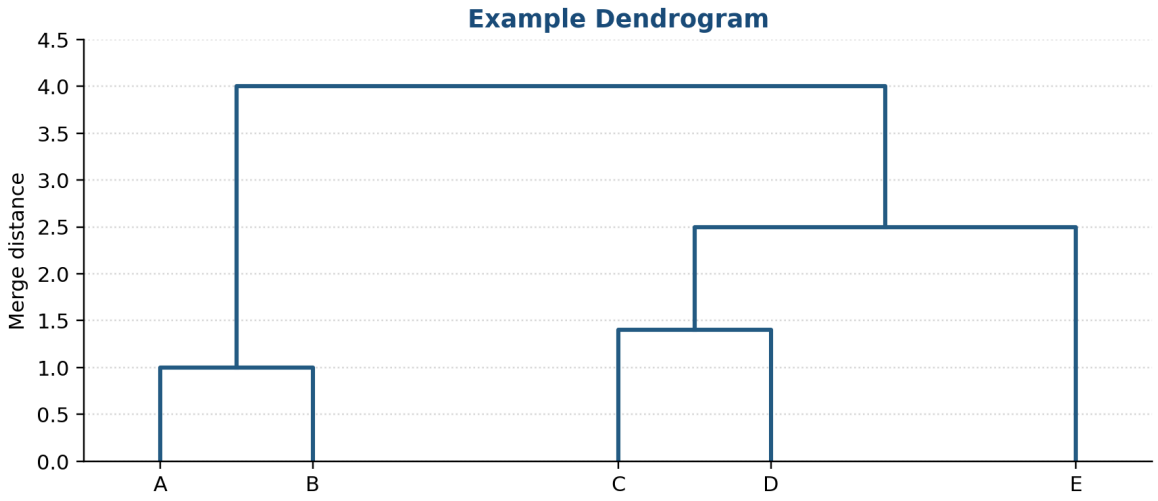


Diagram 2: Example dendrogram showing A and B joining first, followed by C and D.

5. Distance and Linkage Methods

To decide which clusters should be joined, the algorithm needs a way to measure the distance between two clusters. A linkage method defines how this inter-cluster distance is calculated.

Method	Meaning
Single Linkage (Min)	Uses the minimum distance between any point in one cluster and any point in the other cluster.
Complete Linkage (Max)	Uses the maximum distance between any point in one cluster and any point in the other cluster.
Average Linkage	Uses the average of all pairwise distances between points in the two clusters.
Ward's Method	Merges clusters in a way that gives the smallest increase in within-cluster variation.

6. Simple Example

Suppose five fruits are represented by their weights: apple (100 g), banana (120 g), cherry (50 g), grape (30 g) and orange (110 g). In agglomerative clustering, each fruit starts alone. The closest fruits are grouped first. The algorithm then joins the closest groups repeatedly until all fruits belong to one hierarchy. The dendrogram records the order of joining.

7. Applications

- Customer segmentation: grouping customers with similar buying habits.
- Document and topic grouping: organizing similar articles or documents.
- Image segmentation: grouping pixels with similar colour or intensity.
- Bioinformatics: studying similarities among genes or biological samples.
- Pattern recognition: discovering related patterns in datasets.

8. Advantages

- The number of clusters need not be selected in advance.
- The dendrogram provides a clear visual representation of cluster relationships.
- Different distance and linkage methods can be used.
- The hierarchy can be examined at different levels to choose clusters.

9. Limitations

- It can be computationally expensive for large datasets.
- The result may change depending on the distance metric and linkage method.
- Once a merge or split is made, many standard hierarchical methods do not reverse that decision.
- Noise and outliers can affect the grouping process.

10. Agglomerative vs Divisive

Feature	Agglomerative	Divisive
Direction	Bottom-up	Top-down
Starting point	Each point is a separate cluster	All points are in one cluster
Main operation	Merges closest clusters	Splits clusters into smaller groups
Ending point	One cluster or a stopping condition	Individual points or a stopping condition