

Hierarchical Clustering

Builds hierarchy of clusters (dendrogram)



Agglomerative

Bottom-Up

Start with n clusters (each point)
Merge iteratively



Divisive

Top-Down

Start with 1 cluster (all points)
Split iteratively

Linkage Criteria

Single

Minimum distance

Complete

Maximum distance

Average

Mean distance

Ward

Minimizes variance



No k Required

Cut dendrogram at desired level



$O(n^3)$ Complexity

Not scalable to large datasets



Exploratory Analysis

Understanding data structure



Agglomerative Clustering Algorithm

1

Initialize

Start with n clusters (each data point as individual cluster)

2

Compute Distance Matrix

Calculate distances between all pairs of clusters to create distance matrix

3

Find Closest Pair

Find the two closest clusters from the distance matrix

4

Merge Clusters

Merge the two closest clusters into one and record the merge distance

5

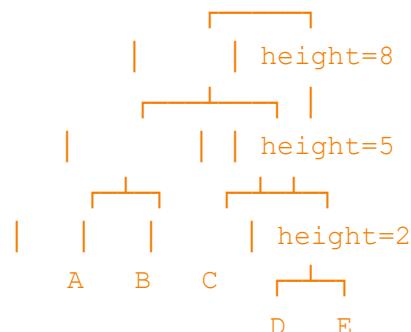
Update Distance Matrix

Recalculate distances between the newly formed cluster and other clusters

6

Repeat

Repeat steps 3-5 until all points form a single cluster

**Dendrogram Structure****Y-axis (Height)**

Represents the distance at which clusters merge. Higher values indicate merging of more dissimilar clusters

**Horizontal Lines**

Represent the merging of two clusters. The height of the line indicates the distance at merge

**Cutting Dendrogram**

Visualizes the height (distance) at which merges occur

Cutting horizontally at a desired height determines the number of clusters at that level

Optimal Cut Point

Cutting where there are large height differences reveals natural cluster boundaries



Phylogenetics

Essential for constructing phylogenetic trees representing evolutionary relationships among species



Data Exploration

Visually understand natural groupings and hierarchical structure of data through dendrograms



Intuitive Interpretation

Easy to understand cluster similarity by viewing merge order and distances at a glance



Flexible Clustering

More flexible than k-means as you can choose the desired number of clusters after analysis