Hierarchical Clustering Algorithm

Jerome Chou

February 1, 2025

Introduction to Hierarchical Clustering

- Hierarchical clustering is an unsupervised learning algorithm used to build a hierarchy of clusters.
- It does not require a predefined number of clusters.
- Two main types:
  - Agglomerative (bottom-up approach)
  - Divisive (top-down approach)

Mathematical Formulation of Hierarchical Clustering

- Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$, the algorithm iteratively merges or splits clusters based on a distance metric.
- Distance between two clusters C_i and C_j can be defined as:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \quad \text{(Single Linkage)} \tag{1}$$

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \quad \text{(Complete Linkage)} \tag{2}$$

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y) \quad \text{(Average Linkage)} \tag{3}$$

Hierarchical Clustering Algorithm Steps

- Compute pairwise distance matrix for all points.
- Repeat until one cluster remains:
    - Merge the two closest clusters based on a linkage criterion.
    - Update the distance matrix.
- For divisive clustering, start with one cluster and recursively split it.

Machine Learning Applications of Hierarchical Clustering

- Genomic data analysis and biological taxonomy.
- Customer segmentation and recommendation systems.
- Anomaly detection in cybersecurity.
- Image segmentation and pattern recognition.