Chapter 14: Hierarchical Clustering
Hierarchical Clustering

The goal of hierarchical clustering is to create a sequence of nested partitions, which can be conveniently visualized via a tree or hierarchy of clusters, also called the cluster *dendrogram*.

The clusters in the hierarchy range from the fine-grained to the coarse-grained – the lowest level of the tree (the leaves) consists of each point in its own cluster, whereas the highest level (the root) consists of all points in one cluster.

Agglomerative hierarchical clustering methods work in a bottom-up manner. Starting with each of the $n$ points in a separate cluster, they repeatedly merge the most similar pair of clusters until all points are members of the same cluster.
Hierarchical Clustering: Nested Partitions

Given a dataset \( D = \{x_1, \ldots, x_n\} \), where \( x_i \in \mathbb{R}^d \), a clustering \( C = \{C_1, \ldots, C_k\} \) is a partition of \( D \).

A clustering \( A = \{A_1, \ldots, A_r\} \) is said to be nested in another clustering \( B = \{B_1, \ldots, B_s\} \) if and only if \( r > s \), and for each cluster \( A_i \in A \), there exists a cluster \( B_j \in B \), such that \( A_i \subseteq B_j \).

Hierarchical clustering yields a sequence of \( n \) nested partitions \( C_1, \ldots, C_n \). The clustering \( C_{t-1} \) is nested in the clustering \( C_t \). The cluster dendrogram is a rooted binary tree that captures this nesting structure, with edges between cluster \( C_i \in C_{t-1} \) and cluster \( C_j \in C_t \) if \( C_i \) is nested in \( C_j \), that is, if \( C_i \subset C_j \).
The dendrogram represents the following sequence of nested partitions:

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>${A}, {B}, {C}, {D}, {E}$</td>
</tr>
<tr>
<td>$C_2$</td>
<td>${AB}, {C}, {D}, {E}$</td>
</tr>
<tr>
<td>$C_3$</td>
<td>${AB}, {CD}, {E}$</td>
</tr>
<tr>
<td>$C_4$</td>
<td>${ABCD}, {E}$</td>
</tr>
<tr>
<td>$C_5$</td>
<td>${ABCDE}$</td>
</tr>
</tbody>
</table>

with $C_{t-1} \subset C_t$ for $t = 2, \ldots, 5$. We assume that $A$ and $B$ are merged before $C$ and $D$. 
The total number of different dendrograms with \( n \) leaves is given as:

\[
\prod_{m=1}^{n-1} (2m - 1) = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n - 3) = (2n - 3)!!
\]
In agglomerative hierarchical clustering, we begin with each of the $n$ points in a separate cluster. We repeatedly merge the two closest clusters until all points are members of the same cluster.

Given a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, we find the closest pair of clusters $C_i$ and $C_j$ and merge them into a new cluster $C_{ij} = C_i \cup C_j$.

Next, we update the set of clusters by removing $C_i$ and $C_j$ and adding $C_{ij}$, as follows $\mathcal{C} = (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$.

We repeat the process until $\mathcal{C}$ contains only one cluster. If specified, we can stop the merging process when there are exactly $k$ clusters remaining.
**Agglomerative Clustering Algorithm**

**Algorithm 14.1: Agglomerative Hierarchical Clustering Algorithm**

\[
\text{AGGLOMERATIVE CLUSTERING}(D, k):
\]

1. \(C \leftarrow \{C_i = \{x_i\} | x_i \in D\} \quad \text{// Each point in separate cluster}\)
2. \(\Delta \leftarrow \{\delta(x_i, x_j) : x_i, x_j \in D\} \quad \text{// Compute distance matrix}\)
3. repeat
4. \hspace{1em} Find the closest pair of clusters \(C_i, C_j \in C\)
5. \hspace{1em} \(C_{ij} \leftarrow C_i \cup C_j \quad \text{// Merge the clusters}\)
6. \hspace{1em} \(C \leftarrow (C \setminus \{C_i, C_j\}) \cup \{C_{ij}\} \quad \text{// Update the clustering}\)
7. \hspace{1em} Update distance matrix \(\Delta\) to reflect new clustering
8. until \(|C| = k\)
Distance between Clusters: Single, Complete and Average

The distance between two points is typically computed using the Euclidean distance or \( L_2 \)-norm

\[
\delta(x, y) = \|x - y\|_2 = \left( \sum_{i=1}^{d} (x_i - y_i)^2 \right)^{1/2}
\]

The between-cluster distances are obtained as follows.

**Single Link:** The minimum distance between a point in \( C_i \) and a point in \( C_j \)

\[
\delta(C_i, C_j) = \min \{ \delta(x, y) \mid x \in C_i, y \in C_j \}
\]

**Complete Link:** The maximum distance between points in the two clusters:

\[
\delta(C_i, C_j) = \max \{ \delta(x, y) \mid x \in C_i, y \in C_j \}
\]

**Group Average:** The average pairwise distance between points in \( C_i \) and \( C_j \):

\[
\delta(C_i, C_j) = \frac{\sum_{x \in C_i} \sum_{y \in C_j} \delta(x, y)}{n_i \cdot n_j}
\]
Distance between Clusters: Mean and Ward’s

**Mean Distance:** The distance between two clusters is defined as the distance between the means or centroids of the two clusters:

$$\delta(C_i, C_j) = \delta(\mu_i, \mu_j)$$

**Minimum Variance or Ward’s Method:** The distance between two clusters is defined as the increase in the sum of squared errors (SSE) when the two clusters are merged, where the SSE for a given cluster $C_i$ is given as

$$\delta(C_i, C_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j$$

where $SSE_i = \sum_{x \in C_i} \|x - \mu_i\|^2$. After simplification, we get:

$$\delta(C_i, C_j) = \left(\frac{n_in_j}{n_i + n_j}\right) \|\mu_i - \mu_j\|^2$$

Ward’s measure is therefore a weighted version of the mean distance measure.
Whenever two clusters $C_i$ and $C_j$ are merged into $C_{ij}$, we need to update the distance matrix by recomputing the distances from the newly created cluster $C_{ij}$ to all other clusters $C_r$ ($r \neq i$ and $r \neq j$).

The Lance–Williams formula provides a general equation to recompute the distances for all of the cluster proximity measures

$$\delta(C_{ij}, C_r) = \alpha_i \cdot \delta(C_i, C_r) + \alpha_j \cdot \delta(C_j, C_r) + \beta \cdot \delta(C_i, C_j) + \gamma \cdot |\delta(C_i, C_r) - \delta(C_j, C_r)|$$

The coefficients $\alpha_i, \alpha_j, \beta,$ and $\gamma$ differ from one measure to another.
Lance–Williams Formulas for Cluster Proximity

<table>
<thead>
<tr>
<th>Measure</th>
<th>$\alpha_i$</th>
<th>$\alpha_j$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single link</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>$-\frac{1}{2}$</td>
</tr>
<tr>
<td>Complete link</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>Group average</td>
<td>$\frac{n_i}{n_i+n_j}$</td>
<td>$\frac{n_j}{n_i+n_j}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean distance</td>
<td>$\frac{n_i}{n_i+n_j}$</td>
<td>$\frac{n_j}{n_i+n_j}$</td>
<td>$-\frac{n_i \cdot n_j}{(n_i+n_j)^2}$</td>
<td>0</td>
</tr>
<tr>
<td>Ward’s measure</td>
<td>$\frac{n_i+n_r}{n_i+n_j+n_r}$</td>
<td>$\frac{n_j+n_r}{n_i+n_j+n_r}$</td>
<td>$-\frac{n_r}{n_i+n_j+n_r}$</td>
<td>0</td>
</tr>
</tbody>
</table>
Iris Dataset: Complete Link Clustering

Contingency Table:

<table>
<thead>
<tr>
<th></th>
<th>iris-setosa</th>
<th>iris-virginica</th>
<th>iris-versicolor</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁ (circle)</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C₂ (triangle)</td>
<td>0</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>C₃ (square)</td>
<td>0</td>
<td>49</td>
<td>14</td>
</tr>
</tbody>
</table>