Context

We consider that we have N data points in a simple D-dimensional Euclidean space

\[ \{x_1, x_2, ..., x_N\} \]

and we assume a given distance d in that space, that can be for example usual Euclidean distance (L_2), Manhattan distance (L_1) or Maximum distance (L_\infty).

Linkage Criteria

In that space, we also consider a linkage criterion l so that for any two clusters of points A and B

\[ A = \{a_1, a_2, ..., a_{|A|}\} \quad B = \{b_1, b_2, ..., b_{|B|}\} \]

l(A,B) is a measure of the similarity between the two clusters A and B. Some usual linkage criteria are for example minimum, maximum, average or ward’s criteria.

Minimum Linkage

\[ l(A, B) = \min\{d(a, b) : a \in A, b \in B\} \]

Maximum Linkage

\[ l(A, B) = \max\{d(a, b) : a \in A, b \in B\} \]

Average Linkage

\[ l(A, B) = \frac{1}{|A|.|B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \]

Ward’s Criterion (for Euclidean distance)

\[ \sum_{x \in A \cup B} \|x - m_{A \cup B}\|^2 - \left( \sum_{a \in A} \|a - m_A\|^2 + \sum_{b \in B} \|b - m_B\|^2 \right) \]
**GENERATE A HIERARCHY OF CLUSTERINGS**

As indicated by its name, hierarchical clustering is a method designed to find a suitable clustering among a generated hierarchy of clusterings. The generated hierarchy depends on the linkage criterion and can be bottom-up, we will then talk about agglomerative clustering, or top-down, we will then talk about divisive clustering.

Agglomerative clustering consists in setting an initial clustering with \( N \) clusters containing a single point each and defining iteratively “hierarchically higher” clusterings. At each iteration, we take in the current clustering the two “closest” clusters according to the chosen linkage criterion and we merge these two clusters together so that to obtain a new clustering with one less cluster.

On the contrary, divisive clustering consists in setting an initial clustering with a single cluster containing the \( N \) points and defining iteratively the “hierarchically lower” clusterings.

**WHEN TO STOP MERGING CLUSTERS?**

The agglomerative and divisive processes we just described give a way to generate a hierarchy of clusterings. However we still need to find a way to pick one final clustering among all those that have been generated.

A common approach consists in plotting the dendrogram of this hierarchy and in identifying the “larger gaps” as possible candidates for cuts. Let’s illustrate all this with a schema.

---

**DENDROGRAM**

- The value of the linkage criterion between \( \{1,2\} \) and \( \{3,4\} \) is \( C \), they are merged third.
- The value of the linkage criterion between 3 and 4 is \( B \), they are merged second.
- The value of the linkage criterion between 1 and 2 is \( A \), they are merged first.
- The gap between \( B \) and \( C \) is the largest (much larger than the gap between \( 0 \) and \( A \) or the gap between \( A \) and \( B \)).
- The “cut” associated with the largest gap generates two clusters: \( \{1,2\} \) and \( \{3,4\} \).
EXAMPLE

Let's illustrate this notion of hierarchical clustering with a simple example for which we consider the natural Euclidean distance and the minimum linkage criterion.

REMARKS

The agglomerative process has a $O(N^3)$ time complexity and a $O(N^2)$ memory complexity that makes it not tractable for large datasets.

The divisive process requires at each iteration to search for the best split, implying a $O(2^N)$ time complexity that has to be tackled with some heuristics.