Hierarchical Clustering Algorithms

Given n items to be clustered, and an $n \times n$ distance matrix, do:

① Start with n clusters, each containing 1 item. Let the distance between the clusters be equal to the distances between the items.

② Find the closest pair of clusters and merge them, so that you have one fewer clusters.

③ Compute "distances" between the new cluster and each of the remaining old clusters.

④ Repeat ② and ③ until there is just a single cluster with n items.

⑤ Return the output as a dendrogram or barcode.

There are various ways to define "distance" between clusters in step ③.

Let X, Y be two clusters, each containing one or more data points. The distance $d(X, Y)$ can be defined as:

Single-linkage: $d(X, Y) = \min_{x \in X, y \in Y} d(x, y)$

Complete-linkage: $d(X, Y) = \max_{x \in X, y \in Y} d(x, y)$

Average-linkage: $\frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x, y)$

Algorithm for single-linkage

(1) Set $m := 0$

$L[0] = 0$

(We'll say the level is $L[m]$ at stage m)

$D_{(0)} = (d_0(i,j))$ The initial $n \times n$ matrix of distances.

(2) while $m < n$ do:

(3) find $1 \le t, s \le n-m$ such that

$$d_m(t,s) = \min_{1 \le i,j \le n-m} d_m(i,j)$$

set $L[m+1] = d_m(t,s)$

(4) Let $D_{(m+1)}(d_{m+1}(i,j))$ be the $(n-m-1) \times (n-m-1)$ matrix obtained from $D_{(m)}$

- delete row s, row t, column s column t
- re-index so that old row i'/ column i' becomes the new row i / column i
- add a new final row and column with

The shape of a data set $S \subseteq \mathbb{R}^n$ can be analysed using cluster analysis.

Compute

$$c = \frac{1}{|S|} \sum_{v \in S} v \qquad \text{(centre)}$$

and for $r > 0$

$$S_r = \{ v \in S : \|v - c\| \geq r \}$$

and for $\varepsilon > 0$ the graph

$G_\varepsilon(S_r)$: vertices $= S_r$

edge $u \text{ —— } v$ if $\|u - v\| \leq \varepsilon$.

For fixed $\varepsilon > 0$ and for

$$r_1 > r_2 > r_3 > \cdots$$

we get a chain of inclusions of graphs

$$G_\varepsilon(S_{r_1}) \subseteq G_\varepsilon(S_{r_2}) \subseteq G_\varepsilon(S_{r_3}) \subseteq \cdots$$

For ε in a "stable range" the resulting bar codes for the connected components may be informative.

$$d_{m+1}(n-m-1, j) = d_{m+1}(j, n-m-1)$$
$$= \min(d_m(s, j'), d_m(t, j'))$$

⑤ Return the list

$L[0], L[1], \ldots, L[n-1]$

as a barcode.