Agglomerative Clustering algorithm

Agglomerative clustering is a type of hierarchical clustering that uses a *bottoms-up* approach of constructing a dendrogram while repeatedly merging separate clusters into one.

1. Start with every point in its own cluster
2. Compute the distance between all pairs of clusters
3. Merge the two nearest clusters
4. Repeat steps 2 and 3 until every point belongs to the same cluster

The *bottoms-up* approach refers to the fact that the dendrogram is **built up** by iteratively combining data points from individual singleton[1] clusters into one maximal cluster.[2]

Distance Functions for Agglomerative Clustering

Single-Link Distance

The **min** of all pairwise distances between a point from one cluster and a point from the other cluster.

- Can handle clusters of varying sizes
- Sensitive to noise points
- Tends to create elongated clusters

$$D_{SL}(C_1, C_2) = \min\{d(p_1, p_2) | p_1 \in C_1, p_2 \in C_2\}$$

Complete-Link Distance

The **max** of all pairwise distances between a point from one cluster and a point from the other cluster.

- Less susceptible to noise than single-link
- Creates more balanced clusters (equal diameter)
- Tends to split up large clusters

$$D_{CL}(C_1, C_2) = \max\{d(p_1, p_2) | p_1 \in C_1, p_2 \in C_2\}$$

[1] informally, a cluster that contains a single data point.
[2] summarized from *An Introduction to Statistical Learning with Applications in Python*, 525-526

Average-Link Distance

The **average** of all pairwise distances between a point from one cluster and a point from the other cluster.

- Less susceptible to noise and outliers
- Biased towards globular clusters

$$D_{AL}(C_1, C_2) = \frac{1}{|C_1| \times |C_2|} \times \sum_{p_1 \in C_1, p_2 \in C_2} d(p_1, p_2)$$

Centroid Distance

The distance between the centers of separate clusters.

$$D_C(C_1, C_2) = d(\mu_1, \mu_2)$$

Ward's Distance

The difference between the spread of points in the merged cluster and the unmerged clusters.

$$D_{WD}(C_1, C_2) = \sum_{p \in C_{12}} d(p, \mu_{12}) - \sum_{p_1 \in C_1} d(p_1, \mu_2) - \sum_{p_2 \in C_2} d(p_2, \mu_2)$$