

## Hierarchical Clustering

Now begins our foray into unsupervised learning. Most unsupervised learning problems deal only with *input* data, typically considered as a matrix  $X \in \mathbb{R}^{N \times D}$  over  $N$  data points and  $D$  dimensions. There are no labels.

Clustering is the problem of grouping data according to similarity.

Typically one uses a *distance metric* to judge similarity (though this is not always required). A metric  $d$  must satisfy:

1.  $d(x, y) = 0$  iff  $x = y$  – isolation
2.  $d(x, y) = d(y, x)$  – distance is symmetric
3.  $d(x, y) + d(y, z) \geq d(x, z)$  – triangle inequality

If (1) doesn't hold, it is a pseudo-metric.

All of our favorite norms *are* metrics. Non-real metrics include things like string edit distance.

Basically two types of clustering:

- Hierarchical
- Flat

Hierarchical tries to build a tree over the data ( $N$  leaves). Typically binary. Flat just tries to find  $K$  clusters of the data.

### Hierarchical Clustering

Bottom-up strategy:

1. Put every element into its own cluster
2. Merge two most similar clusters
3. If  $> 1$  cluster exists, go to (2)

How to measure “similar.” For individual data points, just use the metric  $d$ . What about  $\text{sim}(\{x_1, x_2\}, \{x_3, x_4, x_5\})$ ?

- Min-link (aka “single link”):

$$\text{sim}(R, S) = \min_{x_R \in R, x_S \in S} d(x_R, x_S)$$

- Max-link (aka “complete”):

$$\text{sim}(R, S) = \max_{x_R \in R, x_S \in S} d(x_R, x_S)$$

- Average-link:

$$\text{sim}(R, S) = \frac{1}{|R||S|} \sum_{x_R \in R, x_S \in S} d(x_R, x_S)$$

These have very different properties. Min-link results in chaining. Max-link results in very “round” clusters. Average-link is an intermediary.

Top-down strategy:

1. Put every element into a single cluster
2. Remove the “outsider” (or outsiders) from the group
3. If a cluster of size  $> 1$  exists, go to (2)

Same story as bottom up with linkage. Here, if we remove a single element, we get a funny tree. If we want to split clusters in half, it can be *very* computationally expensive. Typically bottom-up is preferred.