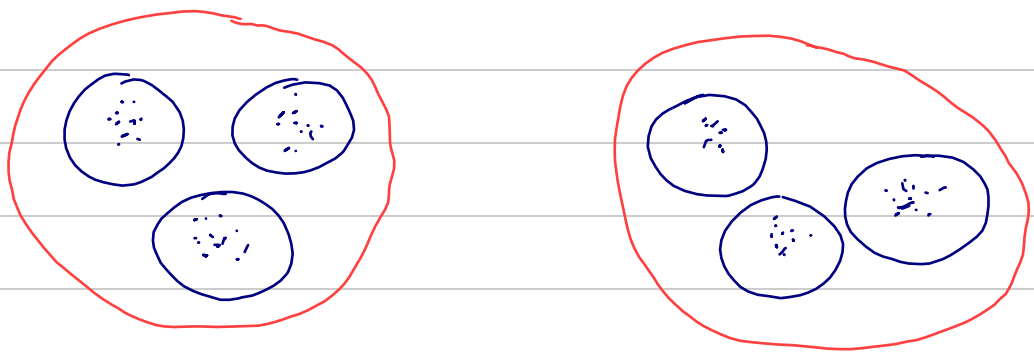
Hierarchical clustering

① Motivation:

certain datasets have a multiscale structure.



→ clustering result is not unique, depends on scale.

→ idea: provide a hierarchy of clusterings, indexed by scale (dendrogram).

② Dendrograms:

Def: A dendrogram on a set $P = \{p_1, \dots, p_n\}$ is a map $\sigma : \mathbb{R}^+ \longrightarrow \text{Partitions}(P)$ such that:

(leaves) * $\sigma(0) = \{\{p_1\}, \dots, \{p_n\}\}$.

(root) * $\exists\, t_0 \geq 0$ s.t. $\sigma(t) = \sigma(t_0) = \{P\}$ $\forall t \geq t_0$.

(monotonicity) * $\forall\, 0 \leq t \leq t'$, $\sigma(t)$ refines $\sigma(t')$, i.e. $\forall\, C \in \sigma(t)$, $\exists!\, C' \in \sigma(t')$ s.t. $C \subseteq C'$.

(continuity) * σ is right-continuous, i.e. $\forall t \geq 0\ \exists\, \varepsilon > 0$ s.t. $\sigma(t') = \sigma(t)$ $\forall\, t' \in [t, t+\varepsilon]$.

Thus, a dendrogram provides a hierarchy of clustering, from the collection of singleton clusters to the full cluster.

→ 2 approaches to build dendrograms:

a) Top-down: recursively partition the dataset.

b) Bottom-up: iteratively agglomerate the clusters.

$\boxed{\text{Note:}}$ approach b) is usually preferred as each step (iteration) has a bounded combinatorial complexity: $O(n^2)$ pairs of clusters to be considered for fusion, versus $\Omega(2^n)$ possible partitions into two subgroups.

→ fusion criteria (approach b):

* single-linkage: merge 2 clusters C, C' as soon as the scale parameter reaches their minimum distance $\min_{x \in C, y \in C'} d(x,y)$.

* complete linkage: same as above but with the maximum distance $\max_{x \in C, y \in C'} d(x,y)$.

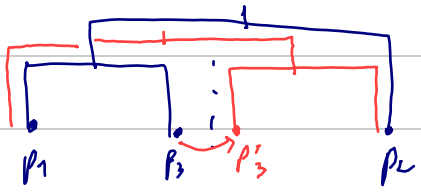
* average linkage: same as above but with the average distance $\frac{1}{|C|\,|C'|} \sum_{\substack{x \in C \\ y \in C'}} d(x,y)$.

* Ward's criterion: same as above but with the weighted squared distance between centroids $\frac{|C|\,|C'|}{|C|+|C'|} \|\bar{C} - \bar{C}'\|_2^2$

centroids of C, C'

Note: Dendrograms are not stable.

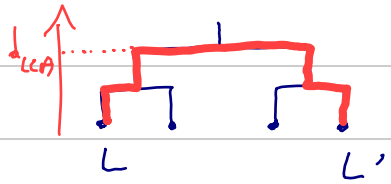
↳ example (with single-linkage):



Q How to express stability? Which part of the dendrogram is stable?

A the merging heights → least common ancestor (LCA)

Def: Given leaves L, L' in a tree T, the LCA height of L, L' is denoted by $d_{LCA}(L, L')$ and defined as the minimum maximum height achieved by continuous paths between L and L' in T.



③ Connection to ultrametrics:

Prop: In a dendrogram, d_{LCA} is an ultrametric, i.e. it is a metric s.t. $d_{LCA}(L, L'') \leq \max\{d_{LCA}(L, L'), d_{LCA}(L', L'')\}$ for all leaves L, L', L''.

→ proof:
- $d_{LCA}(\{p_i\}, \{p_j\}) = 0$ if $i = j$ and > 0 if $i \neq j$ by axioms (leaves) and (continuity).
- $d_{LCA}(\{p_i\}, \{p_j\}) = d_{LCA}(\{p_j\}, \{p_i\})$ by definition (paths are reversible).
- $d_{LCA}(\{p_i\}, \{p_k\}) \leq \max\{d_{LCA}(\{p_i\}, \{p_j\}), d_{LCA}(\{p_j\}, \{p_k\})\}$ because any paths $\{p_i\} \to \{p_j\}$ and $\{p_j\} \to \{p_k\}$ can be concatenated into a path $\{p_i\} \to \{p_j\} \to \{p_k\}$ with maximum height equal to the max. of the maximum heights of $\{p_i\} \to \{p_j\}$ and $\{p_j\} \to \{p_k\}$. ∎

Note: In an ultrametric space (X,d), all triangles are isoceles.

→ proof: let $x,y,z \in X$, and assume wlog that
$$d(x,y) \le d(y,z) \le d(x,z).$$
Then, $d(x,z) \le \max\{d(x,y), d(y,z)\} = d(y,z)$,
and so $d(x,z) = d(y,z)$. ∎

Thm: [see Jain and Dubes: "Algorithms for Clustering Data", 1988]
Given $P = \{p_1, \dots, p_n\}$, there is a bijection $\Psi : \mathcal{D}(P) \to \mathcal{U}(P)$ between the set $\mathcal{D}(P)$ of dendrograms on P and the set $\mathcal{U}(P)$ of ultrametrics on P.
Moreover, $\forall \sigma \in \mathcal{D}(P)$, the ultrametric $\Psi(\sigma)$ yields the same hierarchical decomposition of P as σ:
$$\forall t \ge 0, \quad p_i, p_j \in \text{same cluster in } \sigma(t) \iff \Psi(\sigma)(p_i,p_j) \le t.$$

→ proof: Given $\sigma \in \mathcal{D}(P)$, define $\Psi(\sigma)$ as the LCA height in the dendrogram σ:
$$\forall i,j, \quad \Psi(\sigma)(p_i,p_j) = d_{LCA}(p_i,p_j) = \inf\{t \ge 0 \mid p_i, p_j \in \text{same cluster in } \sigma(t)\}.$$

infimum is a minimum by the axiom (continuity)

As mentioned previously, $\Psi(\sigma)$ is indeed an ultrametric.
Reciprocally, given $d \in \mathcal{U}(P)$, define a dendrogram $\varphi(d)$ as follows:
$\forall t \ge 0$, $\varphi(d)(t)$ is the set of equivalence classes of the relation "$p_i \sim p_j$ iff $d(p_i,p_j) \le t$".

equivalence relation (transitive) by the ultrametric axiom:
$$p_i \sim p_j \text{ and } p_j \sim p_k \Rightarrow d(p_i,p_k) \le \max\{d(p_i,p_j), d(p_j,p_k)\} \le t.$$

Then:

- $\forall \sigma \in \mathcal{D}(P)$, $\varphi \circ \Psi(\sigma) = \sigma$:

  $\forall t \geq 0$, $p_i, p_j \in$ same cluster in $\varphi \circ \Psi(\sigma)(t) \iff \Psi(\sigma)(p_i, p_j) \leq t$

  $\Updownarrow$

  $p_i, p_j \in$ same cluster in $\sigma(t) \iff$ height of LCA of p_i, p_j in σ is $\leq t$

- $\forall d \in \mathcal{U}(P)$, $\Psi \circ \varphi(d) = d$:

  $\forall t \geq 0$, $\Psi \circ \varphi(d)(p_i, p_j) \leq t \iff$ height of LCA of p_i, p_j in $\varphi(d)$ is $\leq t$

  $\Updownarrow$

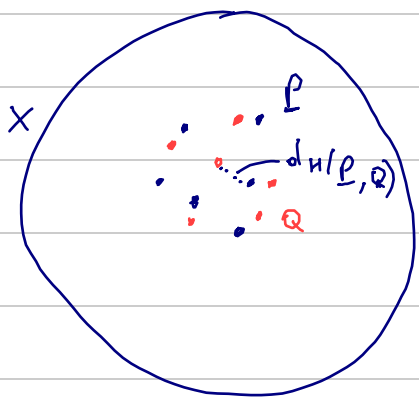
  $d(p_i, p_j) \leq t \iff p_i, p_j \in$ same cluster in $\varphi(d)(t)$

Thus, Ψ and φ are mutual inverses, so Ψ is a bijection.
Moreover, by construction it satisfies:
$\forall t \geq 0$, $\Psi(\sigma)(p_i, p_j) \leq t \iff p_i, p_j \in$ same cluster in $\sigma(t)$. ∎

(4) Stability:

Def: Let P, Q be compact subsets of a metric space (X, d).
The <u>Hausdorff distance</u> $d_H(P, Q)$ is defined by:

$$d_H(P, Q) = \max\left\{ \max_{p \in P} \min_{q \in Q} d(p, q), \max_{q \in Q} \min_{p \in P} d(p, q) \right\}.$$



Def: Let (P, d_P) and (Q, d_Q) be compact metric spaces. The <u>Gromov-Hausdorff distance</u> $d_{GH}(P,Q)$ is defined by:

$$d_{GH}(P,Q) = \inf_{\substack{(X,d_X) \\ \gamma_P : P \to X \\ \gamma_Q : Q \to X}} d_H^{d_X}(\gamma_P(P), \gamma_Q(Q)).$$

γ_P, γ_Q isometries

Example:

$$\begin{cases} d_H(P,Q) > 0. \\ d_{GH}(P,Q) = 0 \end{cases}$$

Note: d_{GH} is NP-hard to compute, even between metric trees [Agarwal et al. 2015]

Thm: [Carlson, Mémoli 2010]

Let (P, d_P) and (Q, d_Q) be finite metric spaces (possibly with distinct cardinalities). Let u_P and u_Q be the ultrametrics induced on P, Q respectively by the dendrogram of single-linkage clustering.

Then:

$$d_{GH}((P, u_P), (Q, u_Q)) \leq d_{GH}((P, d_P), (Q, d_Q)).$$

Thus, the height of LCA in the dendrogram is stable w.r.t. small perturbations of the original metric in the GH-distance.

→ proof in the special case where $P = Q$ (same ingredients as but fewer technicalities than in the general case):

Let d, d' be two metrics on P, and let

$$\varepsilon := \max_{p,q \in P} |d(p,q) - d'(p,q)| .$$

Denote by $\sigma^{SL}(d)$ and $\sigma^{SL}(d')$ the corresponding dendrograms given by single-linkage clustering. Then:

$$\forall p,q \in P,\ \forall t \geq 0,\ \Psi(\sigma^{SL}(d))(p,q) \leq t \iff p,q \in \text{same cluster in } \sigma^{SL}(d)(t)$$

$$\Updownarrow$$

$p,q \in$ same cluster in $\sigma^{SL}(d')(t+\varepsilon)$ $\Leftarrow$ $d'(p_i, p_{i+1}) \leq t+\varepsilon$ $\forall i = 0, \dots, r-1$ $\Leftarrow$ $\exists\, p = p_0, p_1, \dots, p_r = q$ such that $d(p_i, p_{i+1}) \leq t$ $\forall i = 0, \dots, r-1$.

$$\Downarrow$$

$$\Psi(\sigma^{SL}(d')) \leq t + \varepsilon .$$

By symmetry, we obtain $\Psi(\sigma^{SL}(d'))(p,q) \leq t \Rightarrow \Psi(\sigma^{SL}(d))(p,q) \leq t+\varepsilon$,

and so $|\Psi(\sigma^{SL}(d))(p,q) - \Psi(\sigma^{SL}(d'))(p,q)| \leq \varepsilon$.

∎

Note: This only proves stability for the dendrograms of single-linkage clustering under GH-noise.

↳ does not handle outliers ⇝ chaining effect:

$d_{LCA}(Q) \ll d_{LCA}(P)$.

↳ does not say anything about complete-linkage, etc.