AMAT 583 Lec 21 11/12/19

Today: Single linkage clustering, continued
Dendrograms

Review of Single Linkage

Input: A finite metric space (X, d) (we'll assume d is $\mathbb{N}$-valued)
Output: A hierarchical partition (assumed discrete for simplicity).

Def: A hierarchical partition of X is a collection $\{P_\alpha\}_{\alpha \in [0,\infty)}$ of partions of X such that if $\alpha \leq \beta$ and $A \in P_\alpha$, then $A \subset B$ for some $B \in P_\beta$.

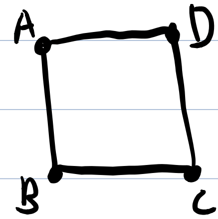
the following variant will be convenient for expository purposes.

Def: A discrete hierarchical partition of X is a collection $P = \{P_\alpha\}_{\alpha \in \mathbb{N}}$ of partions of X such that if $\alpha \leq \beta$ and $A \in P_\alpha$, then $A \subset B$ for some $B \in P_\beta$.

Recall: An undirected graph G is a pair $G = (V, E)$
- V is a set
- E is a set of two-element subsets of V.

For $G=(V,E)$ an undirected graph and $v,w \in V$, a path γ from v to w is a sequence of $n \geq 1$ vertices $v=v_1, v_2, \ldots, v_n = w$ such that for $1 \leq i \leq n-1$, $[v_i, v_{i+1}] \in E$.

If $v=w$ and all edges are distinct (i.e. $[v_i, v_{i+1}] \neq [v_j, v_{j+1}]$ for $i \neq j$), we call γ a cycle.

Example 

$\gamma = \{A, B, C, D, A\}$ is a cycle.

Def: If G has no cycles, it is called a forest.

Define a relation $\sim$ on V by taking $v \sim w$ iff $\exists$ a path from v to w.

Prop: $\sim$ is an equivalence relation.

A subgraph of a graph $G=(V,E)$ is a graph $G'=(V',E')$ with $V' \subset V$, $E' \subset E$.

Def: A connected component of G is a subgraph $G'=(V',E')$ such that
1) V' is an equivalence class of $\sim$
2) $E' = \{(v,w) \in E \mid v,w \in V'\}$. ← That is, every edge in G between vertices in V' is included in G'.

Def: If G is called **connected** if it has one path component.

Def: A connected graph with no cycles is called a **tree**

For X a finite metric space and $z \in \mathbb{N}$, let $N_z(X)$ be the graph with:

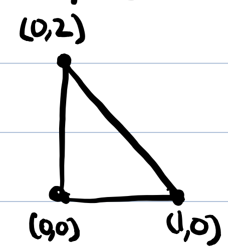
- Vertex set X
- An edge $[x,y]$ included iff $d(x,y) \leq z$.

This is called the neighborhood graph of X at scale z.

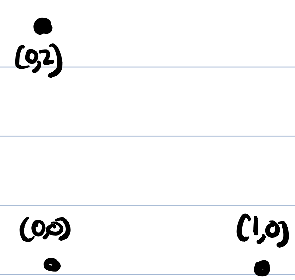
Definition: The single linkage clustering of X is the discrete hierarchical partition $SL(X) = \{SL(X)_z\}_{z \in \mathbb{N}}$

$$SL(X)_z = \{X' \subset X \mid X' \text{ is the vertex set of a connected component of } N_z(X)\}.$$

Example: $X = \{(0,0), (0,2), (,0)\}$, $d = d_1$, the manhattan distance.

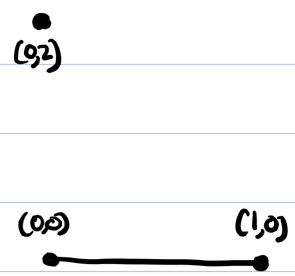


$N_0(X) =$



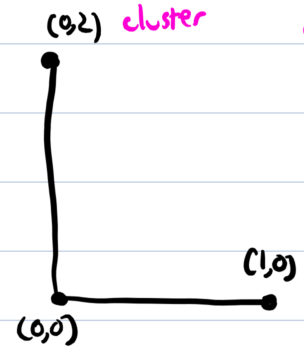
$$SL(X)_0 = \{\underbrace{\{(0,0)\}}_{\text{cluster}}, \underbrace{\{(1,0)\}}_{\text{cluster}}, \underbrace{\{(0,2)\}}_{\text{cluster}}\}.$$

$N_1(X) =$



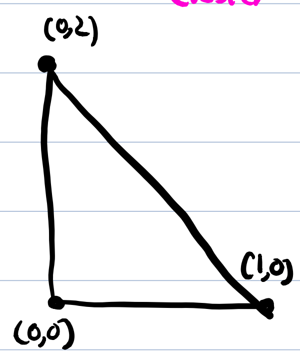
$$SL(X)_1 = \{ \underbrace{\{(0,0),(1,0)\}}_{\text{cluster}}, \underbrace{\{(0,2)\}}_{\text{cluster}} \}.$$

$N_2(X) =$



$$SL(X)_2 = \{ \underbrace{\{(0,0),(1,0)\},(0,2)}_{\text{cluster}} \}.$$

$N_3(X) =$



$$SL(X)_3 = SL_2(X) = \{ \underbrace{\{(0,0),(1,0)\},(0,2)}_{\text{cluster}} \}.$$

In fact $SL(X)_z = SL_2(X)$ $\forall z \geq 2$.

Summarizing,

$$SL(x)_z = \begin{cases} \{\{(0,0)\}, \{(1,0)\}, \{(0,2)\}\} & \text{if } z=0 \\ \{\{(0,0),(1,0)\}, \{(0,2)\}\} & \text{if } z=1 \\ \{\{(0,0),(1,0)\},(0,2)\}\} & \text{if } z\geq 2. \end{cases}$$

Dendrograms

- A standard way of visualizing a hierarchical clustering.

Let $P=\{P_z\}_{z\in\mathbb{N}}$ be a discrete hierarchical partition.

The (unlabeled) dendrogram of P consists of:

- An (infinite) graph $D(P)=(V,E)$
- A function $L: V\to \mathbb{N}$

Specifically, $V=\{(S,z) \mid z\in\mathbb{N},\ S\in P_z\}$

so every element of every partition P_z corresponds to one vertex in the graph.

$$E=\{[(S,z),(T,z+1)] \mid z\in\mathbb{N},\ S\subset T\}.$$

L is defined by $L(S,z)=z$.

Proposition: (1) In general, $D(P)$ is a forest.
(2) If X is a finite metric space, $D(SL(X))$ is a tree.

I'll skip the proof, but some examples should give a fell for this.

Trimming the dendrogram

For any finite metric space X there will be some smallest $z_{top} \in \mathbb{N}$ such that $|SL(X)_z| = 1$ for all $z \geq z_{top}$.

- We usually only plot the subgraph of $D(P)$ consisting of vertices (S, z) with $z \leq z_{top}$, and all edges between such vertices.

- We also usually remove vertices of degree 2.