AMAT 583 Lec 24

<u>Today</u>: More on clustering

<u>Single linkage review</u>

Finite Metric Space X

$\downarrow$

Neighborhood Graphs
$N_0(X) \subset N_1(X) \subset N_3(X) \subset \dots$

$\downarrow$

Discrete hierarchical Partition $SL(X)$

$\downarrow$

Trimmed dendrogram

<u>Recall</u>: I assumed that the metric d on X was integer valued.

But single linkage can be defined for arbitrary finite metric spaces.

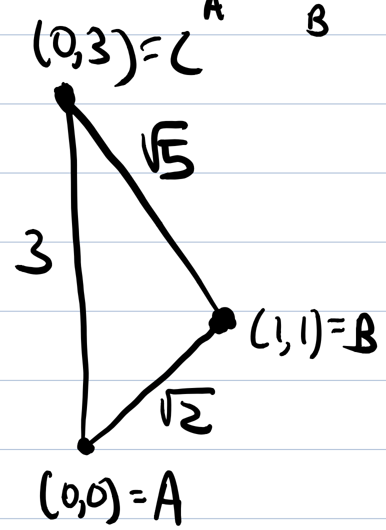
<u>Recall:</u> For X a finite metric space with an integer-valued metric d, and $z \in \mathbb{N}$, we defined $N_z(X)$ by

- $V = X$
- $[v,w] \in E$ iff $d(v,w) \leq z$

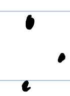
metric on V.

<u>Note:</u> This definition makes equal sense for d not necessarily integer valued and z not necessarily integer-valued.

<u>Example:</u> Consider $X = (S, d_2)$, where $S = \{\underset{A}{(0,0)}, \underset{B}{(1,1)}, \underset{C}{(0,3)}\}$.



$N_r(X) =$

-  for $r \in [0, \sqrt{2})$
-  for $r \in [\sqrt{2}, \sqrt{5})$
-  for $r \in [\sqrt{5}, 3)$
-  for $r \in [3, \infty)$.

Let $N(X) = \{N_r(X)\}_{r \in [0,\infty)}$.

As in the discrete case, for each $N_r(X)$, we obtain a partition of X:

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

These give us a hierachical partition

$$SL(X) = \{SL(X)_r\}_{r \in [0,\infty)}$$

Summery of the Single Linkage Pipeline so far in the case of $[0,\infty)$-valued metrics

Finite Metric Space X

$\downarrow$

Neighborhood Graphs $N(X) = \{N_r(X)\}_{r \in [0,\infty)}$

$\downarrow$

Hierarchical Partition $SL(X) = \{SL(X)_r\}_{r \in [0,\infty)}$

$\downarrow$

???

Question: How do we define the dendrogram of such a (non-discrete) hierarchical partition.

(recall that our definition) of the dendrogram of a discrete hierarchical partition used the discreteness in an essential way.

Key observation: $SL(X)$ "changes" only at finitely many values.

$$0 = r_0 < r_1 < r_2 < r_3 < \cdots < r_n.$$

More precisely, $SL(X)_a = SL(X)_b$

whenever $a, b \in [r_i, r_{i+1})$ for $i \in \{0, 1, \ldots, n-1\}$

or

$a, b \in [r_n, \infty)$.

Example: for $X = (S, d_2)$ as above

$n = 2, \quad r_1 = \sqrt{2}, \quad r_2 = \sqrt{5}.$

(we don't inclue 3 in the r_i because $SL(X)_3 = SL(X)_{\sqrt{5}}$)

Define a <u>discrete</u> hierarchical partition $Q(X)$

By $Q(X)_z = SL(X)_{r_{\min(z,n)}}$.

<u>Example</u>: for $X=(S,d_2)$ as above,

$$Q(X)_0 = SL(X)_0 = \{\{A\},\{B\},\{C\}\}$$
$$Q(X)_1 = SL(X)_{\sqrt{2}} = \{\{A,B\},\{C\}\}$$
$$Q(X)_2 = SL(X)_{\sqrt{5}} = \{\{A,B,C\}\}$$
$$\shortparallel$$
$$Q(X)_3$$
$$\shortparallel$$
$$Q(X)_4$$
$$\vdots$$

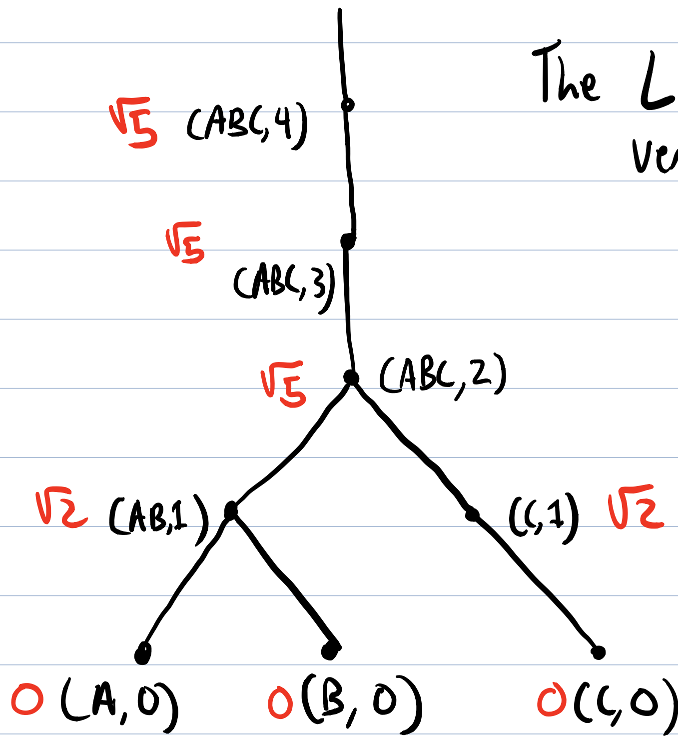
<u>Def</u>: The (untrimmed) <u>dendrogram</u> of $SL(X)$ consists of

- The graph $\underbrace{D(Q(X))}_{\text{The graph underlying the dendrogram of } Q(X).} = (V,E)$

- A function L on vertices $L: V \to [0,\infty)$ given by $L(S,z) = r_{\min(z,n)}$.

<u>Example</u>: For $X=(S,d)$ as above, $D(Q(X))$ is

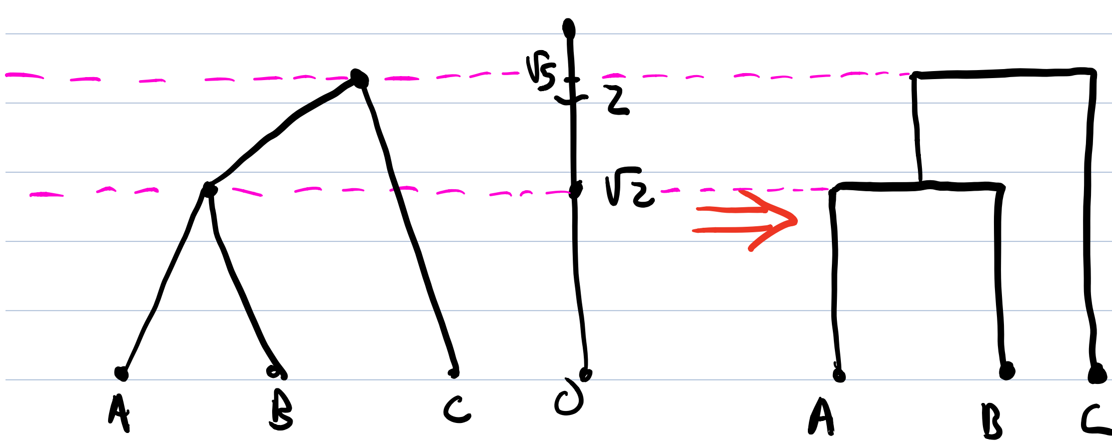
as follows



The L-value of each vertex is plotted in red.

We trim the dendrogram of SL(X) exactly as in the discrete case, and plot vertices at the height of their labels.

Example.



Summary: The definition of the single-linkage dendrogram in the non-discrete case is a simple extension of the definition in the discrete case.

Remark: We have defined the dendrogram of the (non-discrete) hierarchical partition coming from single linkage.

But this generalizes to define the dendrogram of any hierarchical (sub)partition $P = \{P_r\}_{r \in [0,\infty)}$ with the property that P changes at finitely many values $0 = r_0 < r_1 < \dots < r_n$, in the sense above.

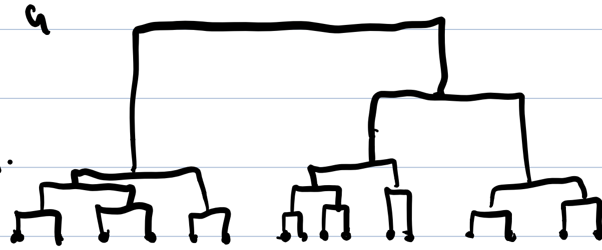
$$\left(\text{i.e. where } P_a = P_b \text{ for } \begin{array}{l} a, b \in [r_i, r_{i+1}) \text{ or} \\ a, b \in [r_n, \infty). \end{array} \right)$$

I call such a subpartition "Essentially discrete."

Remark: The algorithm we outlined for computing a single-linkage dendrogram extends immediately to metric spaces with $[0,\infty)$-valued metrics.

How we actually use dendrograms in practice

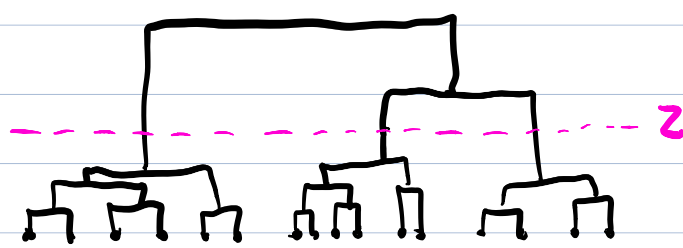
Suppose we have a single-linkage dendrogram like this for a data set X.



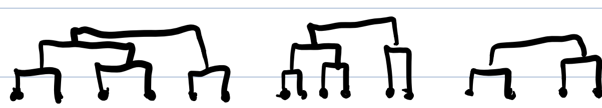
The dendrogram is a visual guide tells how to choose a specific clustering from the family of clusterings $SL(X)_z$.

That is, the dendrogram helps us choose z.

The choice of z can be thought of as a cutting of the dendrogram



Choosing this z corresponds to cutting the dendrogram at height z and keeping only those edges and vertices below the cut



This gives a forest, and the vertices of each tree in the forest is a cluster in $SL(X)$.

Generally, we try to choose z to avoid having many branch points of the dendrogram near level z...