AMAT 583, Lec 25 11/26/19

Today: Single linkage + Topology
Average linkage Clustering
k-means clustering

Business: Homework out ASAP
Due tuesday.
- Exam will cover up to this HW
- Practice final will be provided.
- (actual exam will be very similar)
- No office hours later this week
- Special office hours Monday.

Single Linkage + Topology

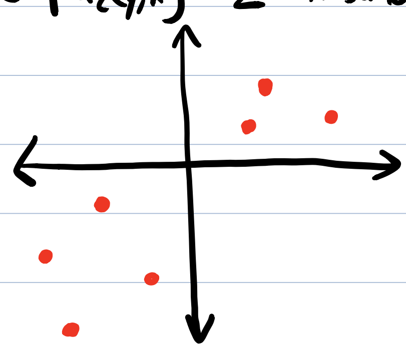
Definition: A filtration is a collection of topological spaces $F = \{F_r\}_{r \geq 0}$ such that $F_r \subset F_s$ whenever $r \leq s$.

Example

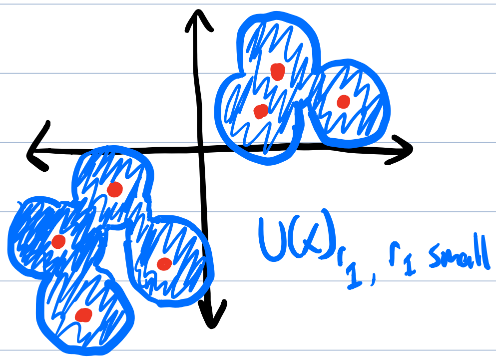
Let X be a finite subset of $\mathbb{R}^n$. For $r \geq 0$, define the union-of-balls filtration $U(X)$ by

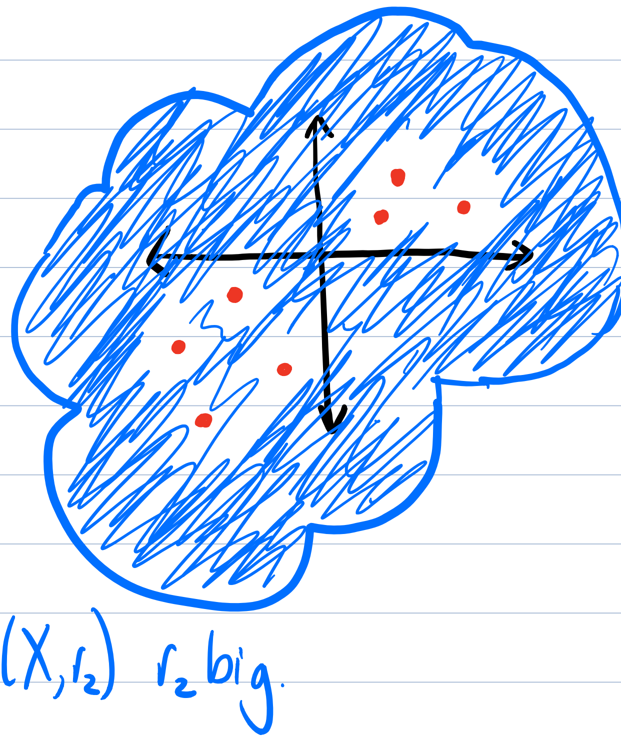
$$U(X)_r = \{y \in \mathbb{R}^n \mid d_2(y,x) \leq \tfrac{r}{2} \text{ for some } x \in X\}.$$

Illustration



$X = U(X)_0$

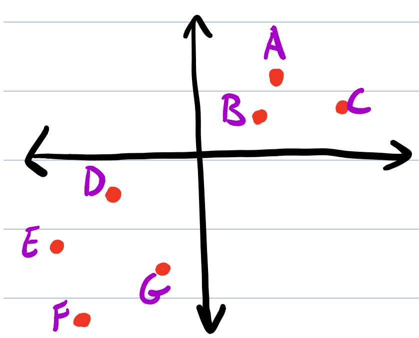




Define a hierarchical partition $SL(X)$ of X by taking

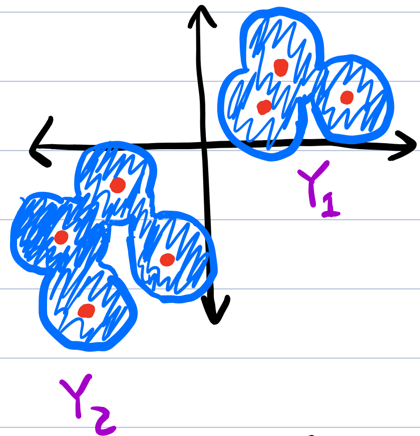
$$C(X)_r = \{X' \subset X \mid X' = X \cap Y, \text{ for } Y \text{ a path component of } U(X,r)\}.$$

<u>Example</u>: Returning to the illustration above,



$U(X)_0$ has 7 path components, one for each point, so

$$C(X)_0 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{G\}\}.$$



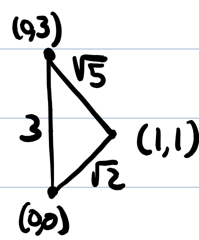
$U(X)_{r_1}$ has 2 path components Y_1, Y_2.

$Y_1 \cap X = \{A, B, C\}$
$Y_2 \cap X = \{D, E, F, G\}$.

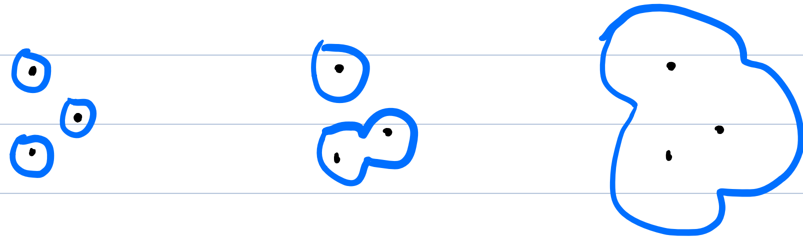
$\Rightarrow C(X)_{r_1} = \{\{A,B,C\},\{D,E,F,G\}\}$.

$U(X)_{r_2}$ has one path component Y. $Y \cap X = X$
$\Rightarrow C(X)_{r_2} = \{X\} = \{\{A,B,C,D,E,F,G\}\}$.

Proposition: For any $X \subset \mathbb{R}^n$, $C(X) = SL(X)$, where $SL(X)$ is the single linkage hierarchical partition defined in terms of neighborhood graphs.

Example Consider $X =$ 

$$U(X)_r \text{ has } \begin{cases} 3 \text{ path components if } 0 \le r < \sqrt{2} \\ 2 \text{ path components if } \sqrt{2} \le r < \sqrt{5} \\ 1 \text{ path component if } \sqrt{5} \le r \end{cases}$$



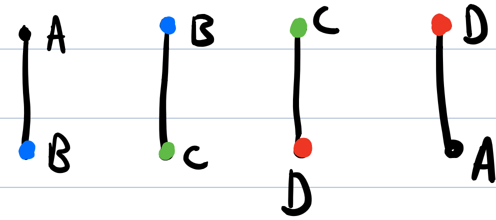
Another (related) connection between single linkage and path components has been implicit in our study of single linkage.

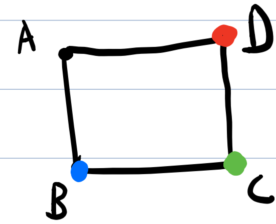
To explain this I need to explain how to regard a graph as a topological space.

For any graph $G=(V,E)$ we construct the geometric realization of G, denoted $|G|$ as follows.

- Let T be the topological space cosisting of 1 copy I_e of I for each edge $e \in E$.
- Label the endpoints of I_e by the corresponding vertices of E.
- $|G|$ is obtained from T by gluing endpoints with the same label together, via the quotient space construction introduced a few weeks ago in class. [Some low level details of this omitted.]

<u>Example</u>: $V=\{A,B,C,D\}$ $E=\{[A,B],[B,C],[C,D],[D,A]\}$.

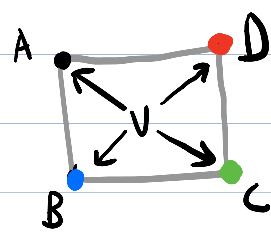
$T =$ 

$|G| =$ 

In this case, $|G|$ embeds into $\mathbb{R}^2$, but that is not always the case.

(However $|G|$ always embeds into $\mathbb{R}^3$.)

Note: We may regard V as a subset of $|G|$, as in the example above:



Now, here's another (equivalent) way to define single linkage clustering:

For any finite metric space (X, d), define the hierarchical partition $C(X)$ by

$$C(X)_r = \{X' \subset X \mid X' = X \cap Y, \text{ for } Y \text{ a path component of } |N_r(X)|\}.$$

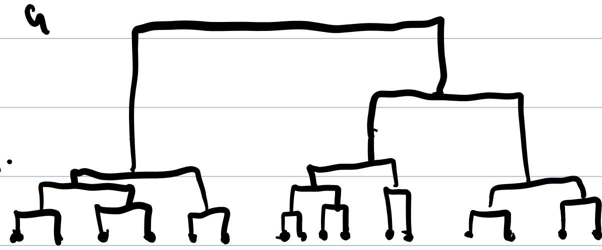
<u>Proposition</u> $C(X) = SL(X)$.

Thus, single linkage can be defined in terms of the path components of the geometric realization.

In this sense, single linkage is a topological clustering method.

<u>How we actually use dendrograms in practice</u>

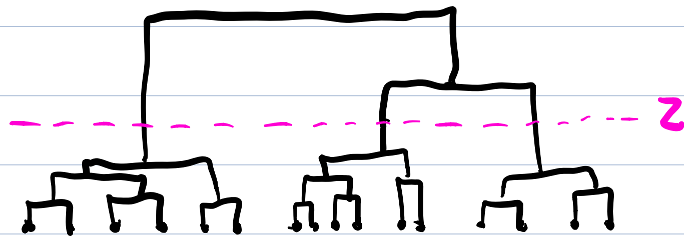
Suppose we have a single-linkage dendrogram like this for a data set X.



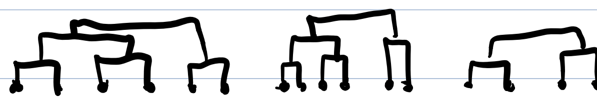
The dendrogram is a visual guide; tells how to choose a specific clustering from the family of clusterings $SL(X)_z$.

That is, the dendrogram helps us choose z.

The choice of z can be thought of as a <u>cutting</u> of the dendrogram

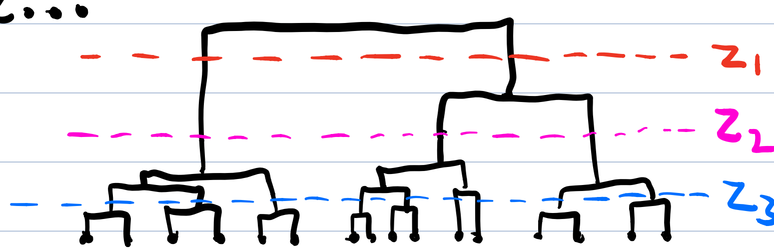


Choosing this z corresponds to cutting the dendrogram at height z and keeping only those edges and vertices below the cut



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

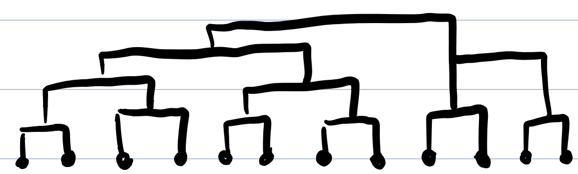
Generally, we try to choose z to avoid having many branch points of the dendrogram near level z...



z_1 or z_2 would be seen as good choices of the parameter z, because when the parameter is perturbed, the clustering does not change.

z_3 is not a good choice; perturbing the parameter will cause clusters to merge or split.

Note that in general, there might not be any "good" choice of z at all:



In this case we may conclude that the data has no clear cluster structure.

Average Linkage Clustering

A popular clustering method

Input is a finite metric space (X, d)

Yields a hierarchical partition (hence a dendrogram).

Motivation: Dendrograms of single linkage are too sensitive to outliers

Easiest to describe algorithmically (as computation of trimmed dendrogram)

Idea: Maintain a collection of clusters and distances between them. Iteratively merge them, and add a node in the dendrogram each time a cluster is merged.

Initially, at $r=0$, each $x \in X$ is in its own cluster $\{x\}$.

Place one vertex at $r=0$ for each cluster

Do the following until there is just 1 cluster:

- Find two different clusters C_1, C_2 s.t.

$$\underbrace{d = \frac{1}{|C_1||C_2|} \sum_{\substack{x \in C_1 \\ y \in C_2}} d(x,y)}_{\substack{\text{avg distance} \\ \text{between points} \\ \text{in } C_1 \text{ and points} \\ \text{in } C_2}}$$ is as small as possible

- Merge C_1 and C_2 to create a new cluster C.
- Add a vertex C to the dendrogram at level d.
- Add the edges $[C_1, C]$ and $[C_2, C]$ to the dendrogram.