Lecture 21: Hierarchical Clustering

What do we do for K-means w/ non-numeric data?

Just have D?

K-mediodis

Step 1: (Update means for K-means)

Find a mediod i.e a representative pt in the cluster - say i_k^* is the element "closest" to all other pts in the cluster

$$i_k^{*(t)} = \underset{i \in G_k^{(t)}}{\operatorname{argmin}} \sum_{i' \in G_k^{(t)}} D_{ii'}$$

Step 2: (update cluster assignments)

assign object i to cluster $G_k^{(t+1)}$ if the closest mediod is $i_k^{*(t)}$ i.e.

$$D_{i i_k^{*(t)}} \leq D_{i i_{k'}^{*(t)}} \quad \forall k'$$

Nice fact: only need D

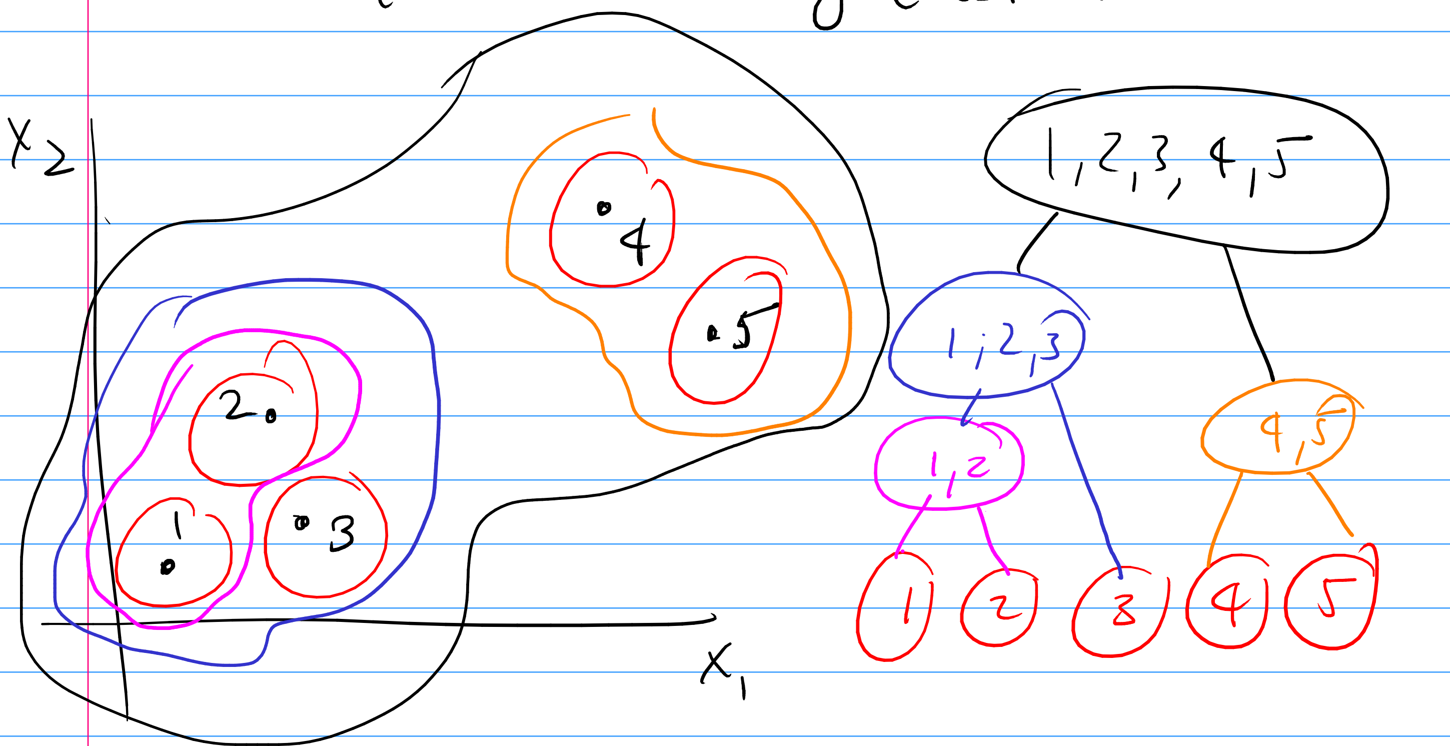
Bad fact: more comp. expensive.

Hierarchical Clustering

Build up a collection (hierarchy) of nested clusters.

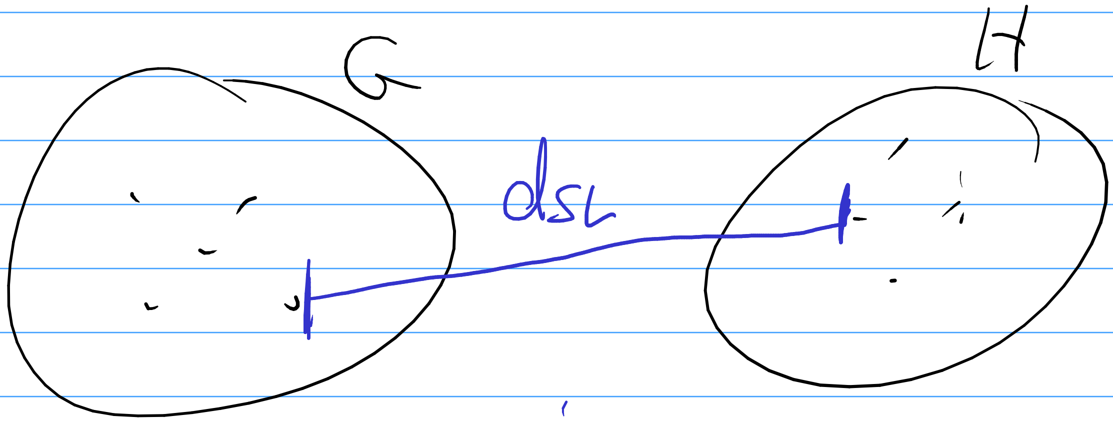
Agglomerative clustering: bottom-up

1) Start w/ each pt being individual cluster
2) merge clusters that are "closer"
3) recursive do step 2 until everything is in one big cluster.



To do this clustering need metric of "closeness" among clusters.

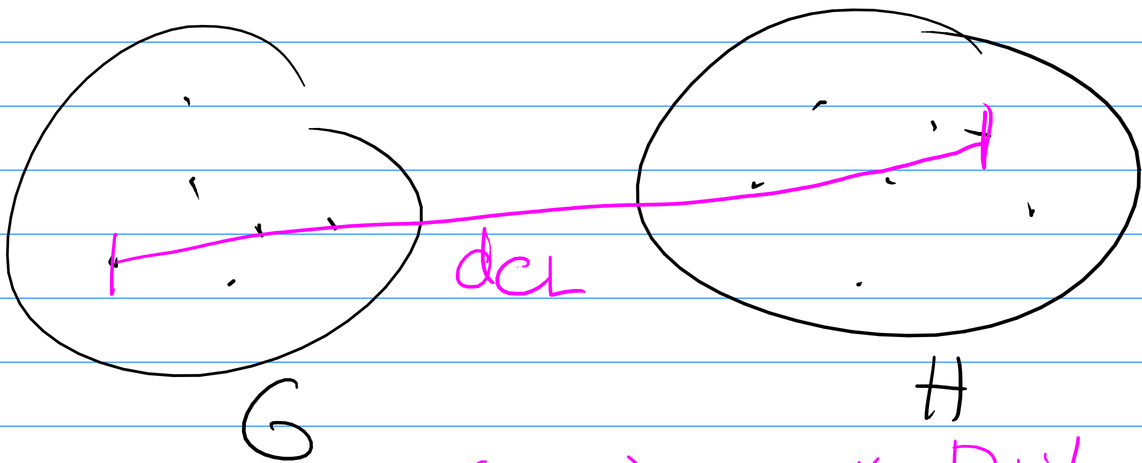
① **Single-Linkage:** dist. btw G and H is the min dissim of any pair one in G and one in H



$$d_{SL}(G,H) = \min_{\substack{i \in G \\ i' \in H}} D_{ii'}$$

② **Complete linkage**

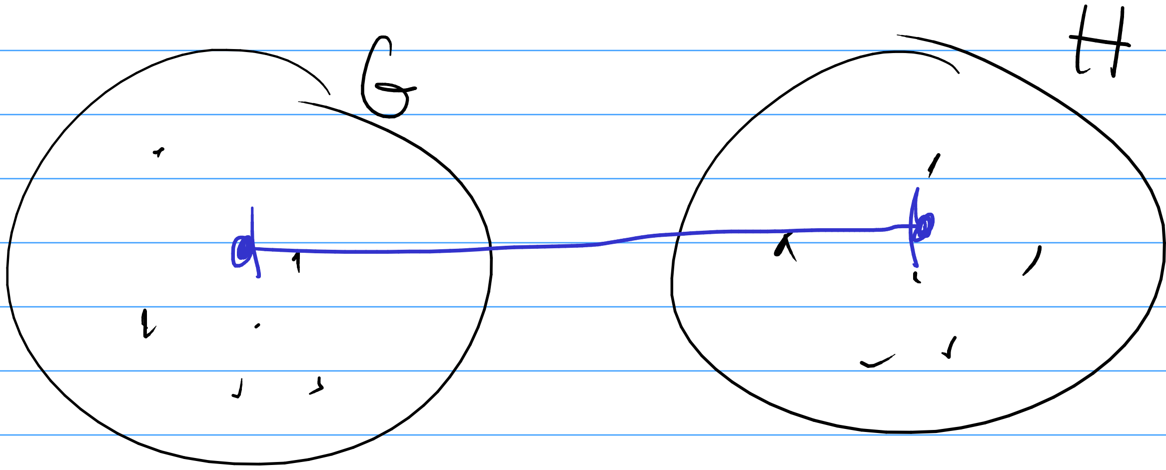
G and H are close if the max pairwise dissim is small



$$d_{CL}(G,H) = \max_{\substack{i \in G \\ i' \in H}} D_{ii'}$$

3) Avg. Linkage: avg. dissim. btwn clusters

$$d_{AVG}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} D_{ii'}$$



Dendogram