Last time: K-means clustering

If X is a $N \times P$ mtx of numeric covariates and I define my dissim mtx D so that

$$D_{ii'} = \|x_i - x_{i'}\|^2$$

Combinatorical Clustering Problem:

Choose Gs $(G_1, \ldots, G_K)$ so that

$$W = W(G_1, \ldots, G_K) = \frac{1}{2}\sum_{k=1}^{K} \sum_{i,i' \in G_k} D_{ii'}$$

is minimized, i.e.

$$\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_K = \underset{G_1, \ldots, G_K}{\arg\min}\, W(G_1, \ldots, G_K)$$

Can show: Equivalent to

obs in K^{th} cluster

$$W = \sum_{k=1}^{K} N_k \sum_{i \in G_k} \|x_i - \bar{x}_k\|^2$$

$\bar{x}_k \in \mathbb{R}^P$

$$W = \frac{1}{2}\sum_{k=1}^{K} \sum_{i,i' \in G_k} D_{ii'} = \frac{1}{2}\sum_{k=1}^{K} \sum_{i,i' \in G_k} \|x_i - x_{i'}\|^2$$

$$W = \frac{1}{2}\sum_{k=1}\sum_{i,i'\in G_k} D_{ii'} = \frac{1}{2}\sum_{k=1}\sum_{i,i'\in G_k} \underbrace{\|x_i - x_{i'}\|^2}$$

$$\|a\|^2 = a^T a$$

$$\|x_i - x_{i'}\|^2 = \|x_i - \bar{x}_k + \bar{x}_k - x_{i'}\|^2$$

$$= \big((x_i - \bar{x}_k) + (\bar{x}_k - x_{i'})\big)^T(\cdots)$$

$$\underbrace{(x_i - \bar{x}_k)^T(x_i - \bar{x}_k)}_{\|x_i - \bar{x}_k\|^2} + \underbrace{(\bar{x}_k - x_{i'})^T(\bar{x}_k - x_{i'})}_{\|x_{i'} - \bar{x}_k\|^2}$$

$$+ 2(x_i - \bar{x}_k)^T(\bar{x}_k - x_{i'})$$

$$= \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i,i'\in G_k}\|x_i - \bar{x}_k\|^2 + \|x_{i'} - \bar{x}_k\|^2 + \underbrace{2\sum_{i,i'\in G_k}(x_i - \bar{x})^T(\bar{x} - x_{i'})}_{0}\right]$$

$$\sum(x_i - \bar{x}) = 0$$

$$= \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in G_k}\sum_{i'\in G_k}\left(\|x_i - \bar{x}_k\|^2 + \|x_{i'} - \bar{x}_k\|^2\right)$$

$$= \frac{1}{2}\sum_{k=1}^{K}\left[\left(\sum_{i}\sum_{i'}^{N_k}\|x_i - \bar{x}_k\|^2\right) + \left(\sum_{i'}\sum_{i}^{N_k}\|x_{i'} - \bar{x}_k\|^2\right)\right]$$

$$= \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i} N_k\|x_i - \bar{x}_k\|^2 + \sum_{i}\|x_i - \bar{x}_k\|^2\right]$$

$$\boxed{= \sum_{k=1}^{K} N_k \sum_{i\in G_k}\|x_i - \bar{x}_k\|^2} = \text{TWCSS}$$

Proposed Lloyd's Algorithm:

$t = 1, 2, 3, \ldots$

Update ① $\mu_k^{(t)} = \frac{1}{N_k} \sum_{i \in G_k^{(t)}} x_i$

Assign ② $G_k^{(t)} = \{ x_i \mid \|x_i - \mu_k^{(t)}\| \le \|x_i - \mu_{k'}^{(t)}\| \ \forall k' \}$

Why does Lloyds work?

Problem: $\hat{G}_1, \ldots, \hat{G}_K = \underset{G_1, \ldots, G_K}{\operatorname{argmin}} \sum_k N_k \sum_{i \in G_k} \|x_i - \bar{x}_k\|^2$

Generalize Problem

→ $\min_{G_1, \ldots, G_K, \underbrace{m_1, \ldots, m_K}_{\text{centers}}} \underbrace{\boxed{\sum_k N_k \sum_{i \in G_k} \|x_i - m_k\|^2}}_{W}$

→ Fact: $\bar{x} = \underset{m}{\operatorname{argmin}} \|x_i - m\|^2$

Step ① Given Gs, set $m_k = \bar{x}_k$

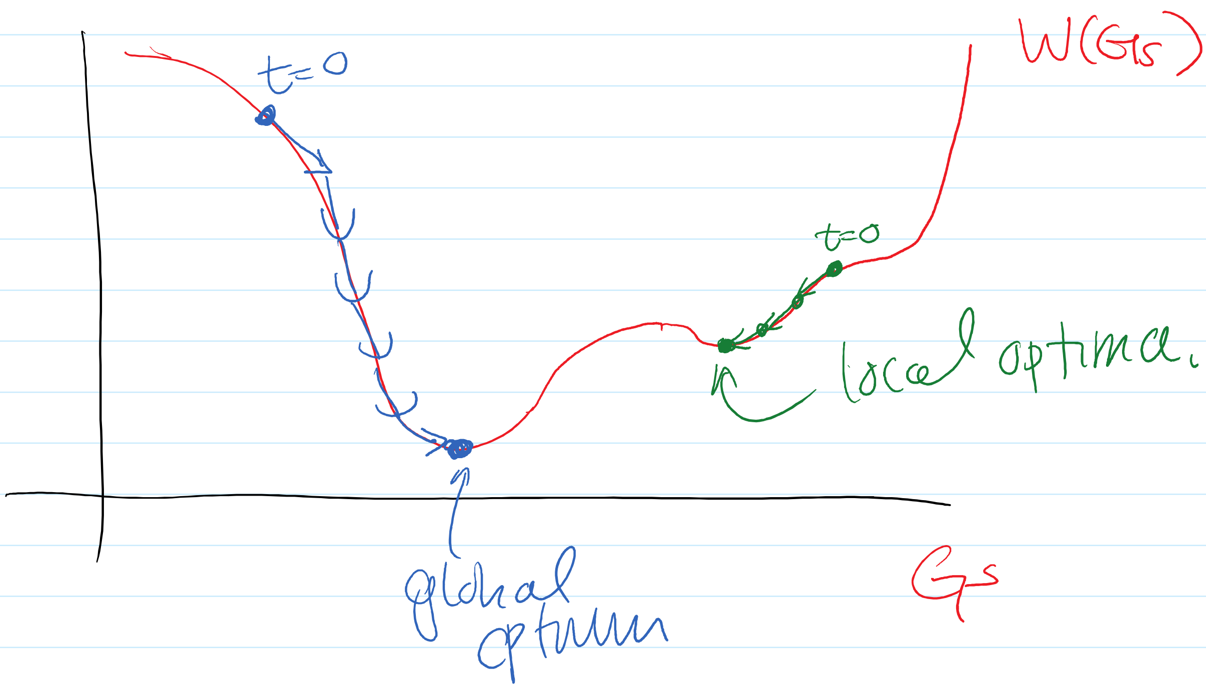
this reduces W

Step 2 Given ms, choose Gs as pts closest to ms, this reduces W

So at <u>each step</u> this <u>makes W smaller</u>.

<u>HOWEVER</u>

We don't nec. get to globally optimum soln.



<u>Soln:</u> Try several random initializations take soln w/ lowest W.

What about non-numeric data?

Just have D? ← non-euclidean dist/dissim.

K-Mediods

Soln: Replace step ① (update means) by setting m_k = pt in cluster closest to everything else in cluster

Step 1^* Find obs in cluster closest to others

$$\rightarrow \quad i_k^* = \underset{i \in G_k}{\operatorname{argmin}} \sum_{i' \in G_k} D_{ii'}$$

← explict optim problem

② Assignment step.

Object i goes in Group G_k if

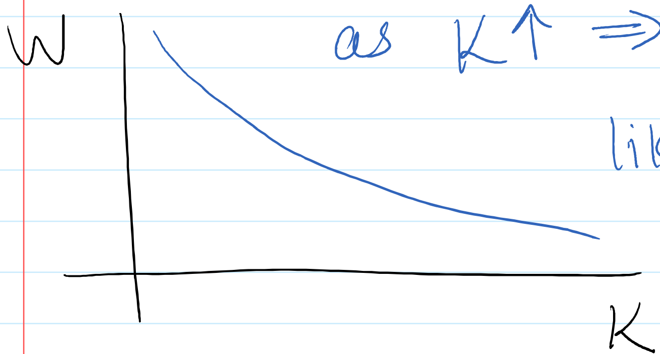
$$D_{ii_k^*} \leq D_{ii_{k'}^*} \quad \forall k'$$

Nice fact: No X, just D.

Nice fact: No X, just D.

Bad fact: more comp. intensive

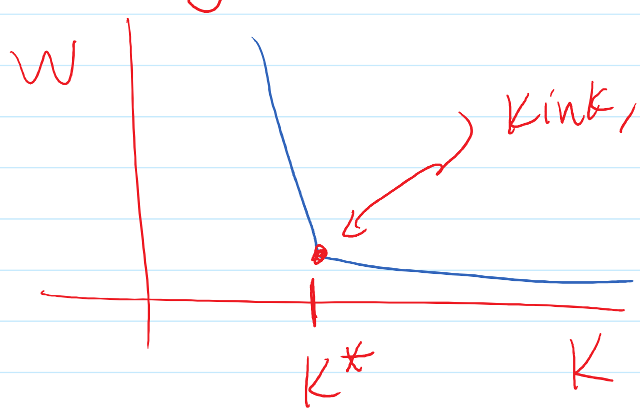
How do I choose K?



as $K \uparrow \Rightarrow W \downarrow$ (monotonically)

like KNN can't choose K in this way.

One way:



Hierarchical Clustering

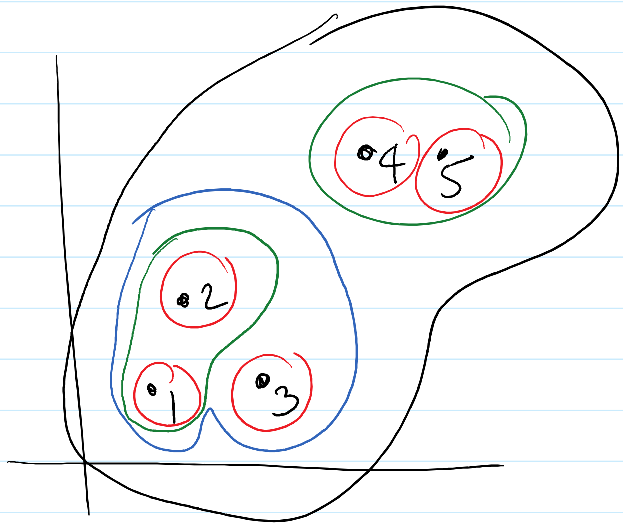
Build a collection (hierarchy) of nested clusters.

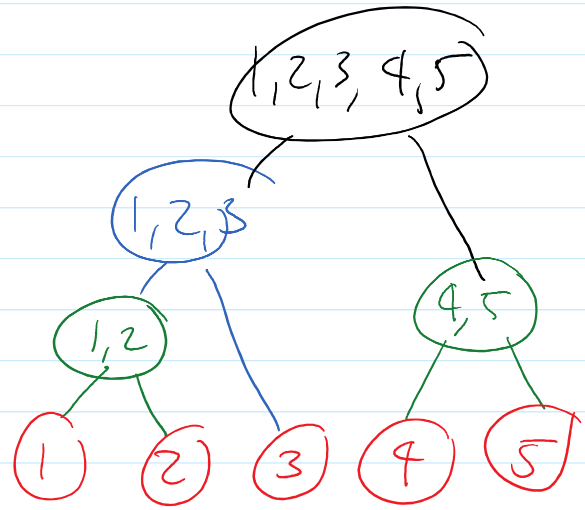
(1) Agglomerative Clustering

(i) start w/ clusters as individual pts

(ii) merge clusters that a "similar" or "close"

(iii) recursively repeat until everything is in one cluster.





(2) <u>Divisive Clustering</u> (opposite)

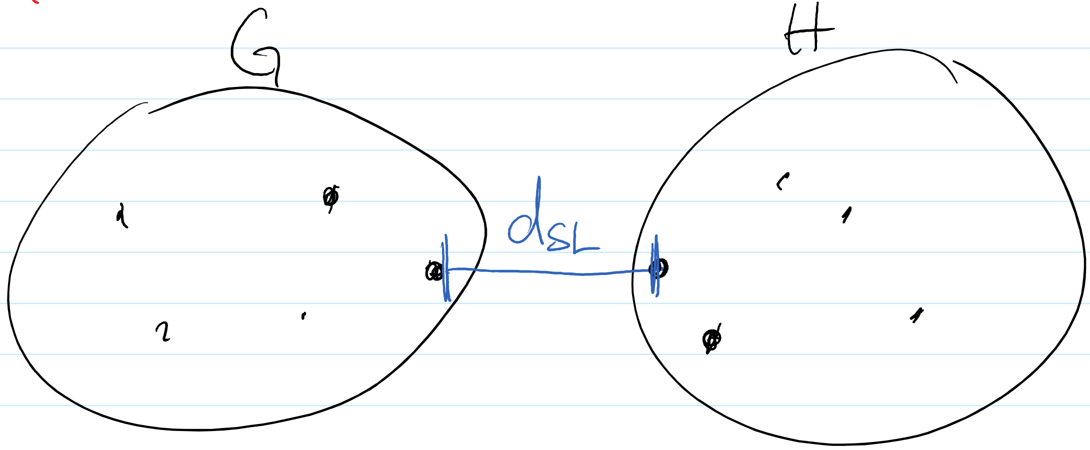
(1) start w/ 1 large cluster

(2) recursively break into smaller clusters

To do <u>agglomerative clustering</u> we need a measure of "closeness" between clusters.
"similarity"

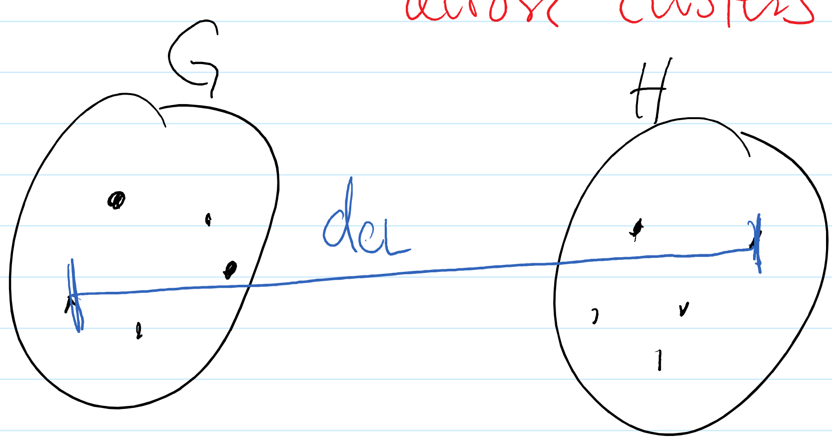
(1) <u>Single Linkage</u>: dist. btwn clusters G and H is the min dissim btwn any 2 pts:

one in G and one in H.



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

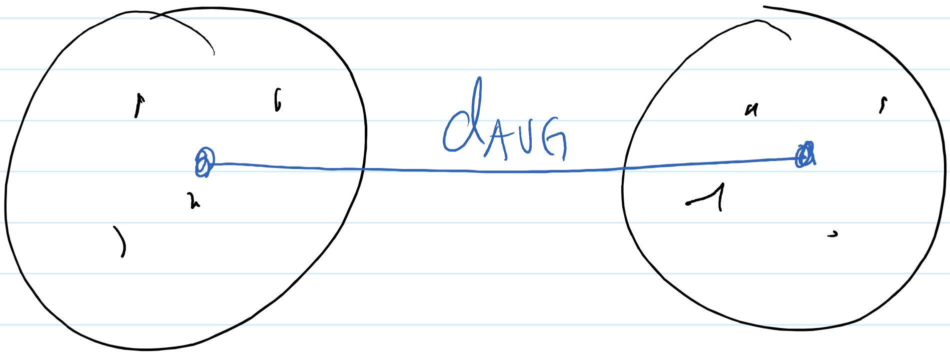
② Complete Linkage: max dissim btwn 2 pts across clusters



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

③ Average Linkage: avg dissim btwn clust.

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



Dendrogram