Lecture 16 -- k-means

Tuesday, October 29, 2024 3:28 PM

Why is this reasonable? (Lloyd's Algo)

$$\hat{G}_1, \dots, \hat{G}_K = \underset{G_1, \dots, G_K}{\operatorname{argmin}} \underbrace{\sum_k N_k \sum_{i \in G_k} \|X_i - \bar{X}_k\|^2}_{W}$$

Generalize:

$$\hat{G}_1, \dots, \hat{G}_K, \hat{m}_1, \dots, \hat{m}_k = \boxed{\underset{G_k, m_k}{\operatorname{argmin}} \boxed{\sum_k N_k \sum_{i \in G_k} \|X_k - m_k\|^2}}$$

Fact: ① given G_ks the best values for m_k is $m_k = \bar{X}_k$

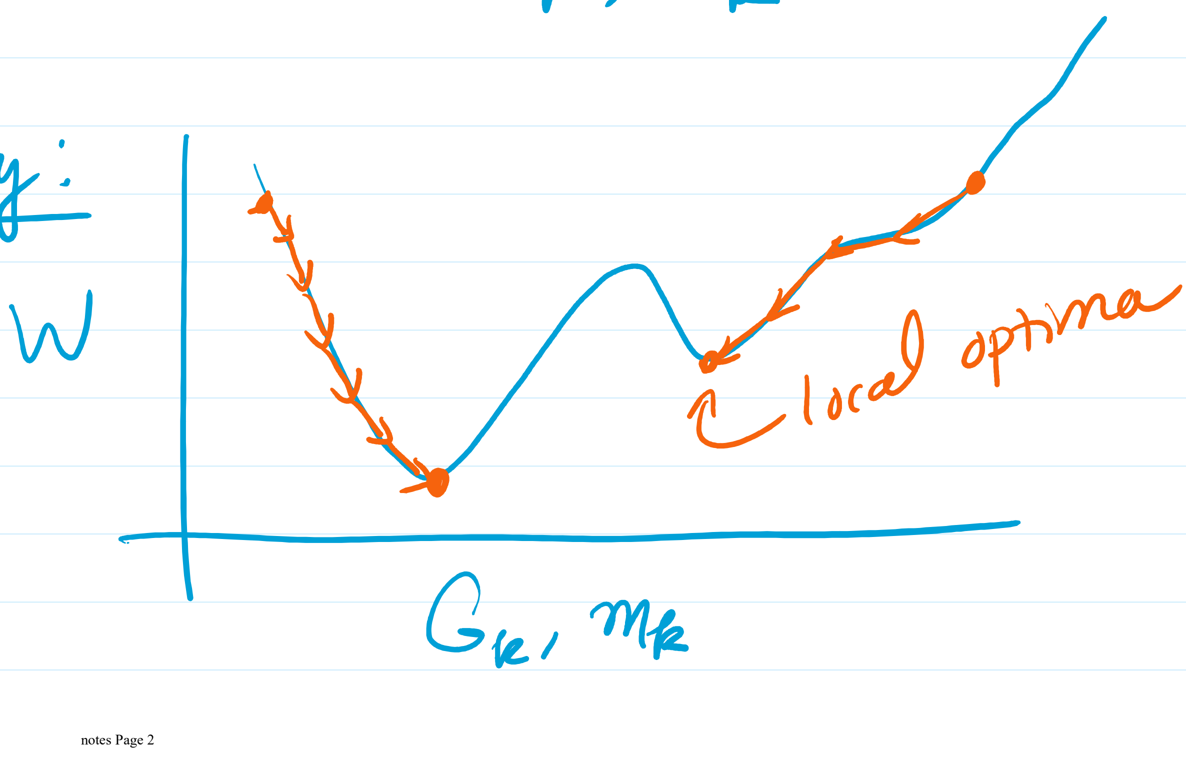
② Given fixed values for m_k, the best G_k is to assign pts to cluster w/ closest m_k

closest m_k

Because of these two facts, each step of Lloyd's also will decrease W

Ideally:

Reality:

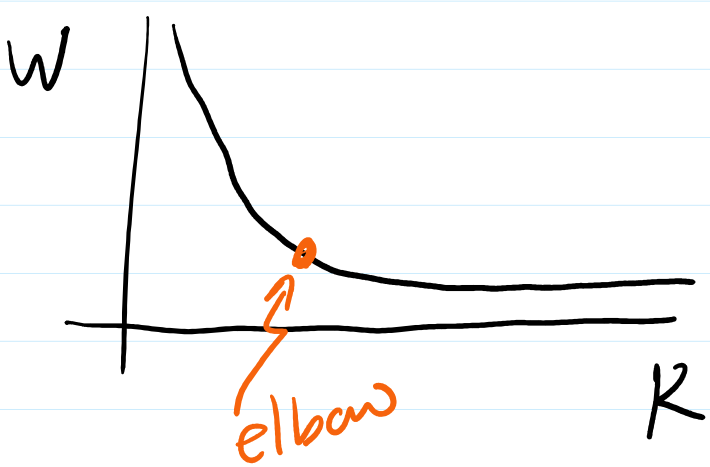


C_k, m_k

To avoid local minima try multiple random initializations and choose clustering w/ lowest val of W.

How do I choose K?

Can't choose K to minimize W, increasing K will always decrease W:



What about non-numeric data or non-euclidean dissim metric.

All you need is D

K-mediods

Step 0: initialization

choose K points i_k^* $k=1,\dots,K$
randomly in my data

"representatives" of my clusters - called mediods

For $t=1,2,3,\dots$

Step 1: assignment

assign each data point i to cluster G_k
if the dissim btwn i and i_k^* is the
smallest among all choice of k

assign i to G_k if $D_{i i_k^*} \leq D_{i i_{k'}^*}$ $\forall k'$

Step 2 : Update

Choose new medoids for each group as pt w/ least total dissim to all other pts in cluster

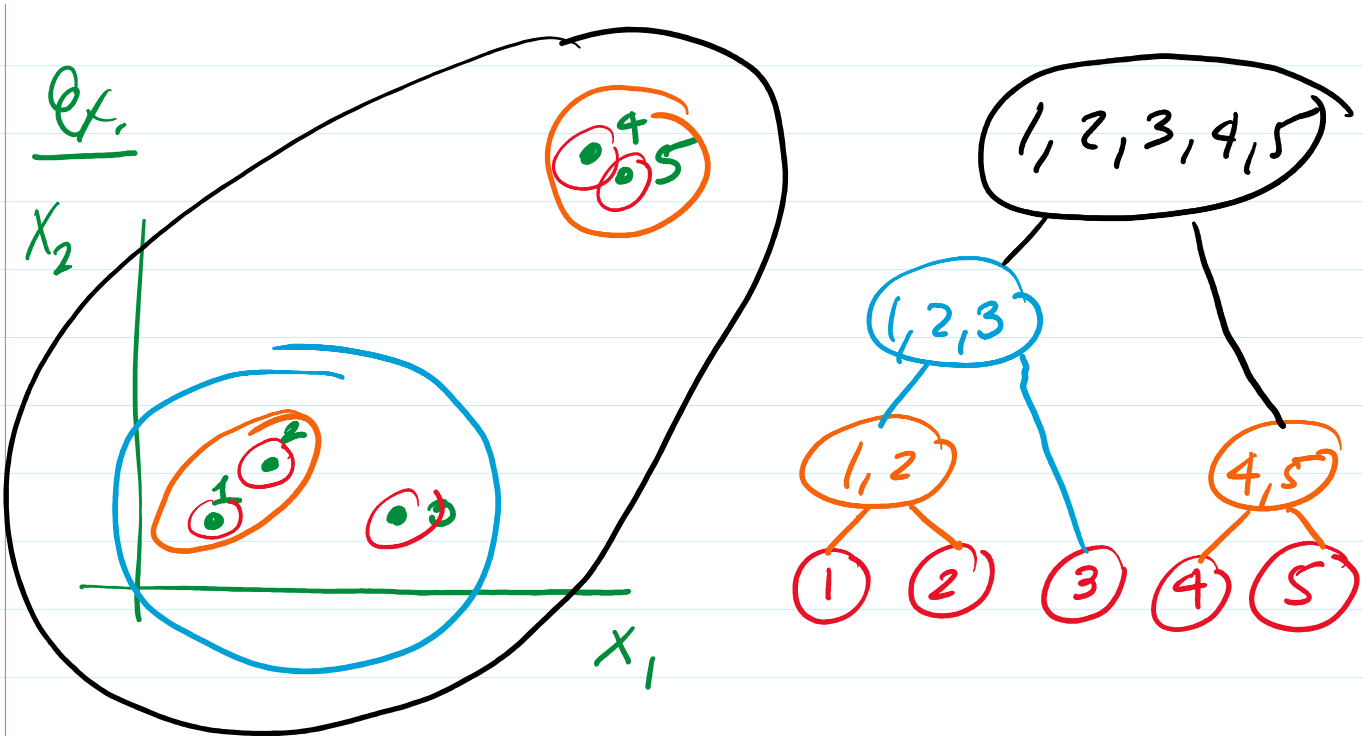
$$i_k^* = \underset{i \in G_k}{\operatorname{argmin}} \sum_{i' \in G_k} D_{ii'}$$

Hierarchical Clustering

Build up a collection (hierarchy) of nested clusters.

Agglomerative clustering: bottom-up

1. start w/ each pt as individual cluster
2. merge clusters that are close
3. recursively do 2 until everything is in a single cluster



To do this, need some measure of "closeness" of clusters

<u>Many ways to do:</u>

<u>Single-linkage</u>: dist. btwn G and H is the min dissim btwn pts

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

<u>Complete-linkage</u>: dist is max dissim

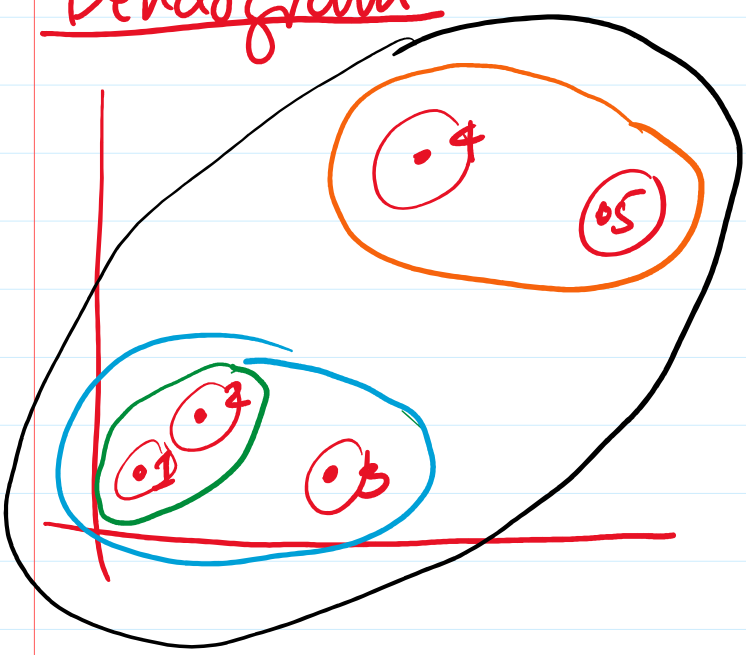
Complete-linkage : dist is max dissim

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

Average-Linkage

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

Dendrogram



dist btwn clusters when merged

Dendo