Lecture 20: K-means clustering

Assume that data comes from one of K clusters

$$G_1, G_2, \ldots, G_K$$

Want to assign each point to some G_k

How: do assignment to minimize some measure of not being well clustered ("Loss")

Classic loss/measurement of clustered-ness is

$$W = \text{total w/in cluster dissim.} = \sum_{k=1}^{K} \sum_{i,i' \in G_k} D_{ii'}$$

↑ should be large if poorly clustered

↑ should be small if well clustered

$$T = \text{total dissim} = \sum_{i,i'} D_{ii'}$$

$$B = \text{total between cluster dissim} = \sum_{k,k'} \sum_{i \in G_k} \sum_{i' \in G_{k'}} D_{ii'}$$

One can show that

$$T = W + B$$

So to find $G_1, \ldots, G_K$ we should

(1) minimize W

or (2) maximize B

Q: how? Ideally, I should try all possible cluster assignments and find/choose one w/ smallest W

Problem: not typically computationally tractable,

Ex. $N = 19$, $K = 4$

$\sim 10^{10}$ ways to do this.

Sh: Greed. (Greedy optim approach)

K-means: all features are numeric

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

Want to find G_ks to minimize W

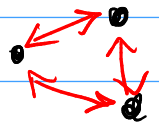
Can show:

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

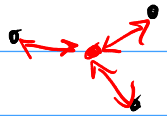
N_k: # items in cluster k

$\bar{x}_k$: mean of xs in G_k

Defn. W



claim



also W

$$W = \sum_k \sum_{i \in G_k} \sum_{i' \in G_k} \|x_i - x_{i'}\|^2$$

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

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

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

$\|x_i - \bar{x}_k\|^2$

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

$\sum_i = 0$

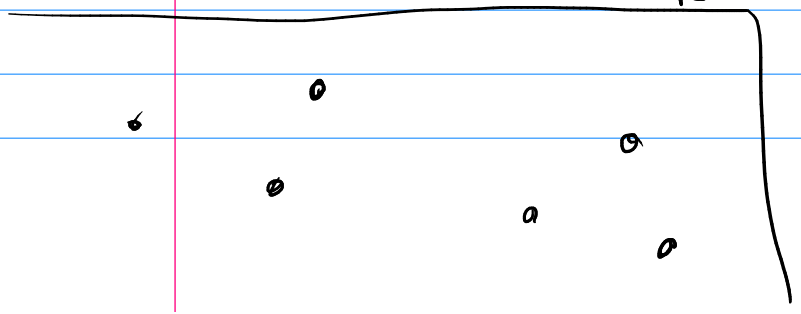
y_i

$\sum_i (y_i - \bar{y}) = 0$

$$\sum_i (*) = \sum_i \|x_i - \bar{x}_k\|^2 + N_k \|x_{i'} - \bar{x}_k\|^2$$

$$\sum_{i'} \sum_i (*) = N_k \sum_i \|x_i - \bar{x}_k\|^2 + N_k \sum_{i'} \|x_{i'} - \bar{x}_k\|^2$$

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



K-means: Lloyd's Algorithm

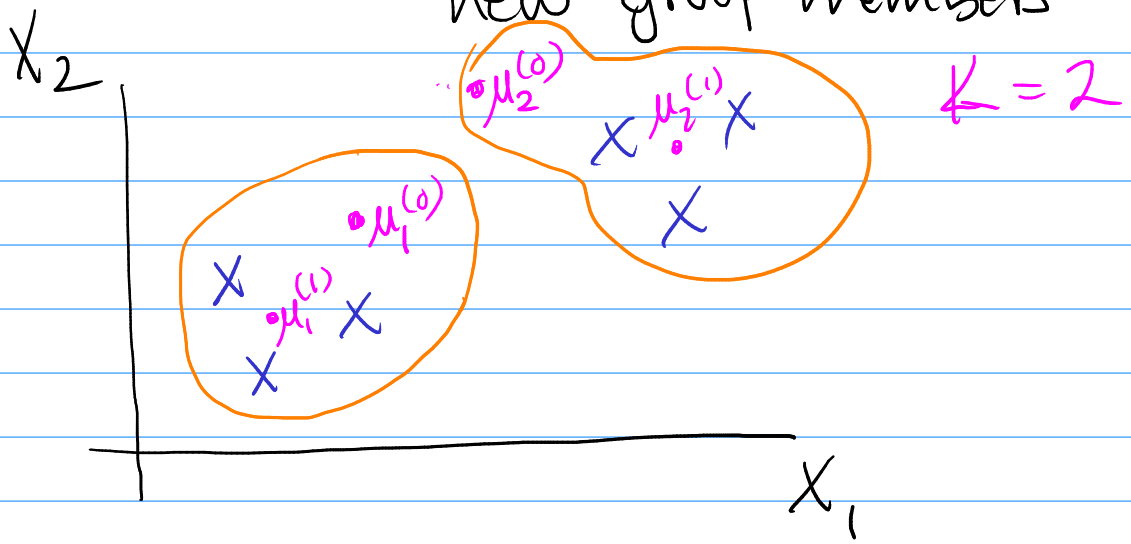
(0) **Initialize Step**: make random initializations of cluster means

$$\mu_1^{(0)}, \dots, \mu_K^{(0)}$$

For $t = 1, 2, 3, 4, \dots$ and at the t^{th} iteration

(1) **Assignment step** assign x_i to the group w/ closest mean $\forall i$

(2) **Update Step** re-compute the μs as the means of the new group members



Llogd's Algo

Assign: $G_k^{(t)} = \{ i \mid \|x_i - \mu_k^{(t-1)}\| \leq \|x_i - \mu_{k'}^{(t-1)}\| \ \forall k' \}$

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

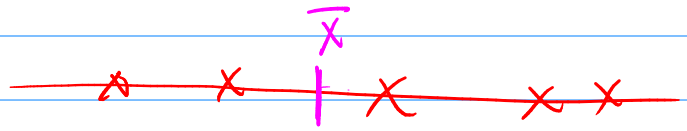
Why does this work?

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

Generalize problem

$$\hat{G}_1, \ldots, \hat{G}_K, \hat{m}_1, \ldots, \hat{m}_K = \underset{G_i, m_i}{\operatorname{argmin}} \underbrace{\sum_k N_k \sum_{i \in G_k} \|x_i - m_i\|^2}_{W}$$

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



Lloyds is basically coordinate descent

Step (1) Given ms find best Gs

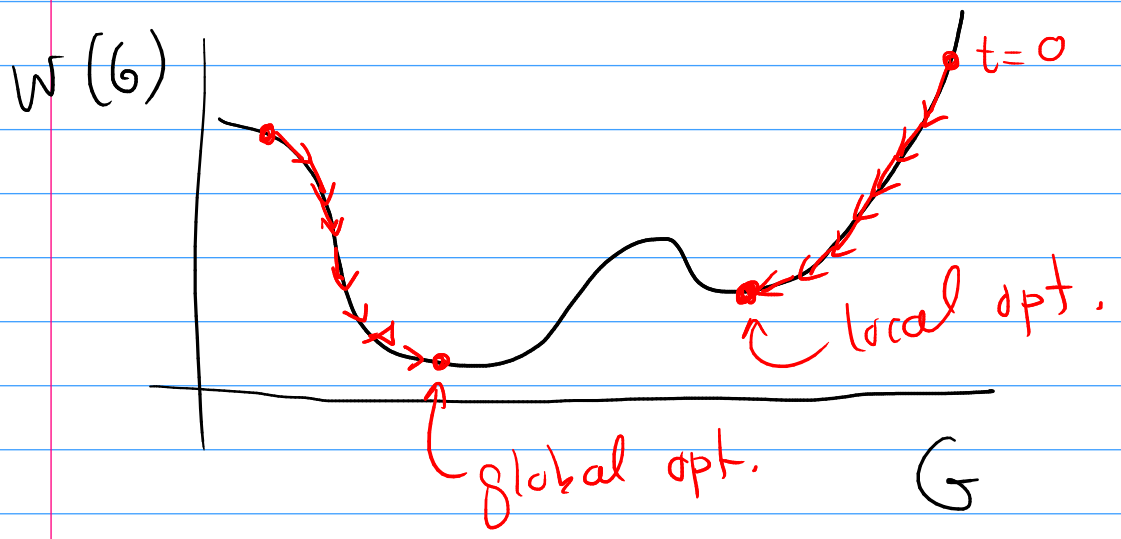
Assigning pts to nearest m reduces W

Step (2) Given Gs find best ms

Let m_k = mean of G_k $\Rightarrow$ reduce W

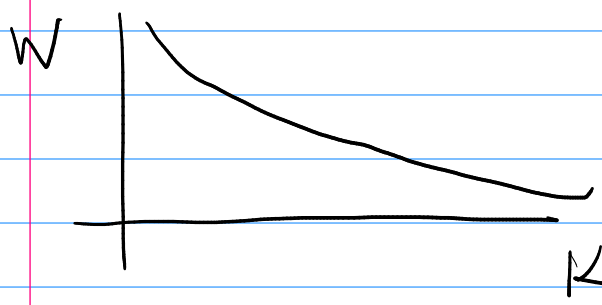
Idea: each step makes W better (smaller) & eventually we'll converge on some answer

problem: may get stuck in a local minimum



One potential soln: try several random inits

How do I choose K?



as $K \uparrow \Rightarrow W \downarrow$

(like KNN)

One way look for a "knee" in graph