

L9: Assignment-based Clustering

e.g. k -means / k -center
Lloyds / Gonzalez

Data Mining
Sep 22, 2026
Jeff M. Phillips

PATHWAYS COMPUTING CAREER EXPO

9/24

10AM – 2PM
Alumni House



**Sign Up
Now!**

U OF U DATA SCIENCE CLUB PRESENT

DATA DISCOVERY HUNT

**MONDAY
SEP 28**

Drop in ANYTIME between:
9:00AM - 4:00PM
@ MEB, Roundabout

Join us for The Data Discovery Hunt — solve clues, grab treats, and win prizes: hat, t-shirt, stickers, mugs and treats

How It Works

1. Check in at the main table and scan the QR code for your first clue.
2. Solve Data Science & CS-themed puzzles at stations throughout MEB.
3. Unlock the next clue with each puzzle you complete.
4. Finish at the main table to claim prizes and participation credit.

💡 You can stop anytime — you'll still get a participation treat!



Scan to RSVP



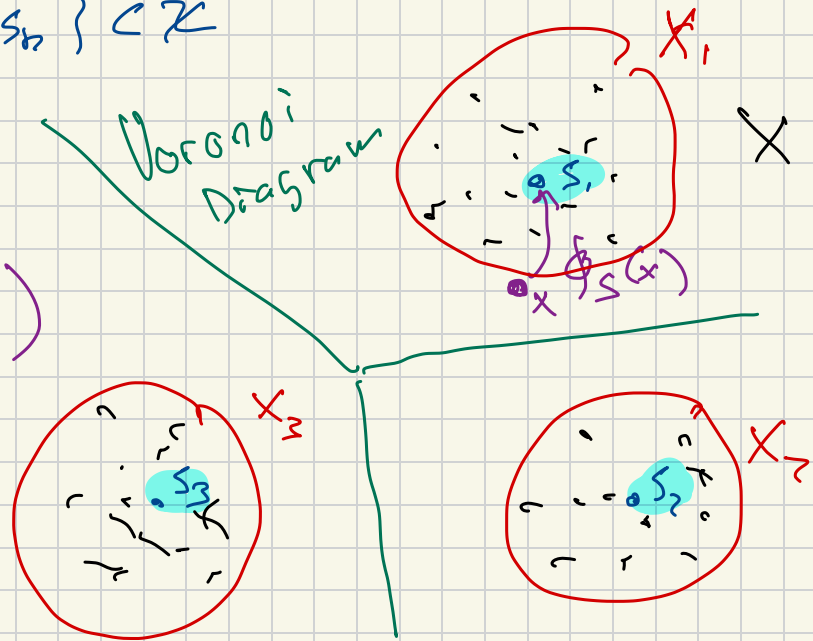
Join our discord

Input data $X \subset \mathcal{X} (\equiv \mathbb{R}^d)$ parameter k # clusters
 distance $D : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ (often $D(a,b) = \|a-b\|$)

Output: sets $\{X_1, X_2, \dots, X_k\} \subset X$ $s_j \in \mathcal{X}$
 defined sites $S = \{s_1, s_2, \dots, s_k\} \subset \mathcal{X}$

Mapping Function
 $s_j^*(x) = \phi_S(x) = \underset{s_j \in S}{\operatorname{argmin}} D(x, s_j)$

implicitly Define
 $X_j = \{x \in X \mid \phi_S^*(x) = s_j\}$



Variants

Assignment-based Clustering

Define

Cost of clustering, want to minimize.

• K-Means

$$\text{Cost}_2(X, S') = \frac{1}{|X|} \sum_{x \in X} D(x, \phi_{S'}(x))^2 = \frac{1}{|X|} \sum_{x \in X} \|x - \phi_{S'}(x)\|^2$$

Lloyd's Algo, $X \subset \mathbb{R}^d$ $D(x, s) = \|x - s\|$

• K-Median

$$\text{Cost}_1(X, S') = \frac{1}{|X|} \sum_{x \in X} D(x, \phi_{S'}(x))$$

better w/ outliers, not great Also

• K-Medoid

$$\text{Cost}_1(X, S') = \frac{1}{|X|} \sum_{x \in X} D(x, \phi_{S'}(x))$$

sites in data
 $S' \subset X$

• K-Center

$$\text{Cost}_\infty(X, S') = \max_{x \in X} D(x, \phi_{S'}(x))$$

select outliers, Gonzalez Algo

Gonzalez's Algo

for k -center

Input $X \subset \mathcal{X}$, distance D (metric), $k = \# \text{ sites}$

0. $s_1 \in X$ (arbitrarily?) $\mathcal{S}_1 = \{s_1\}$

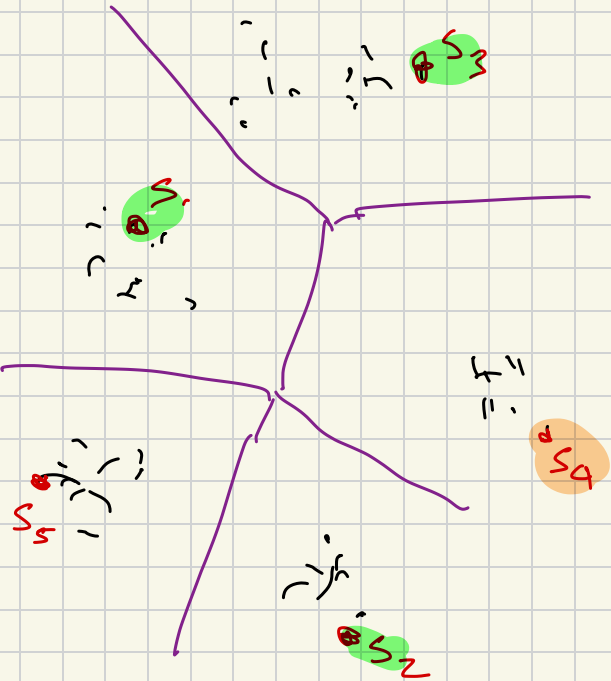
1. for $j = 2$ to k

Set $s_j = \underset{x \in X}{\operatorname{argmax}} D(x, \mathcal{S}_{j-1})$

2. Return $\mathcal{S} = \mathcal{S}_k$

sites found
so far

tends to pick
on boundary points



Lloyd's Algo

k-Means

Input $X \subset \mathbb{R}^d$

$$D(a, b) = \|a - b\|_2, \quad k$$

Lloyd's Algo

0. Choose k sites $S \subset X$ (how?)

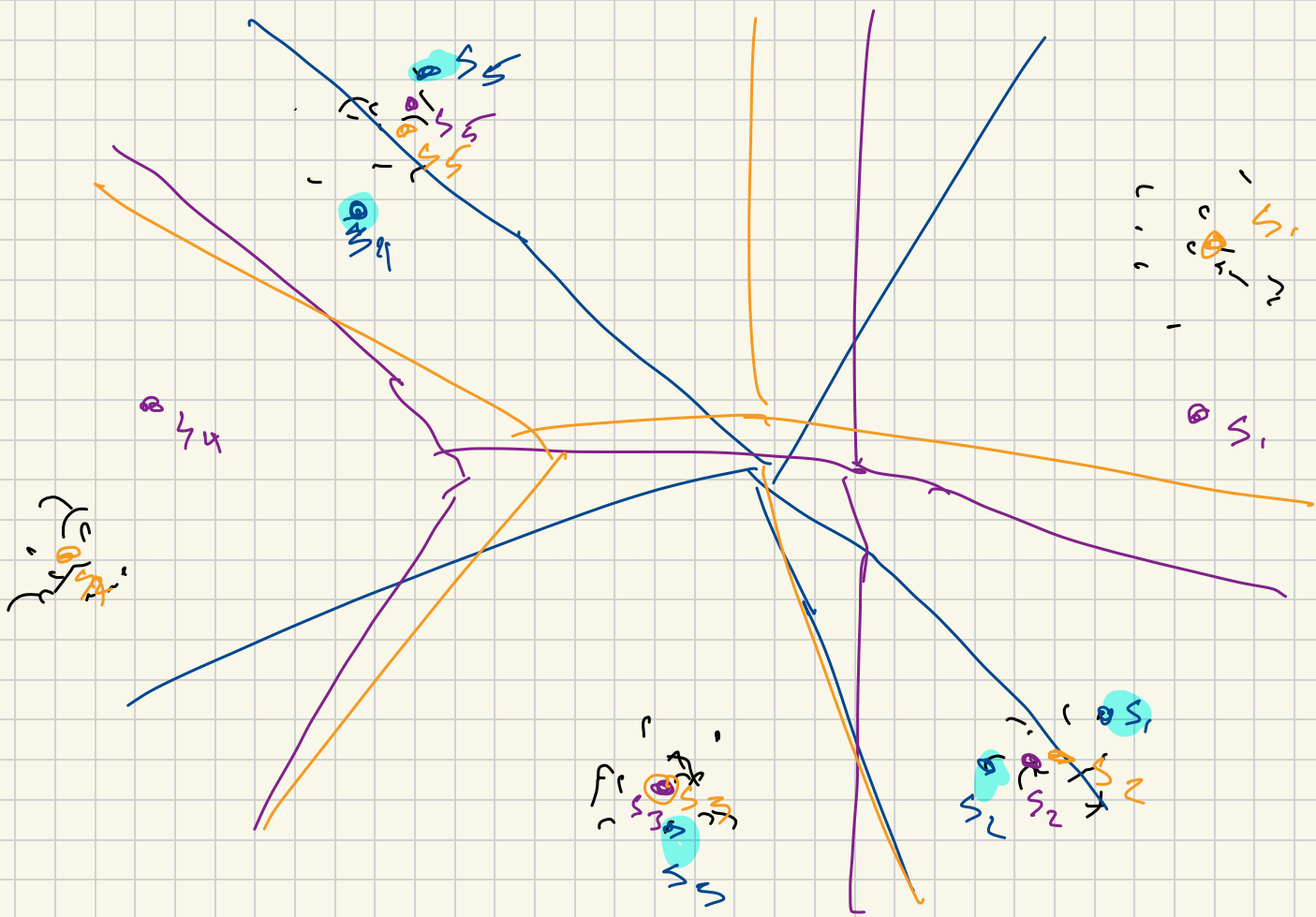
1. repeat

1.a: For all $x \in X$, set $\phi_S(x) = \arg \min_{s_j \in S} \|x - s_j\|$

1.b: For all $j \in [k]$, $s_j = \text{average} \{x \in X \mid \phi_S(x) = s_j\}$

2. until (S unchanged (or 20 steps))

Return S



Mean

$$S_j = \text{average} \{x \in X \mid \phi_S(x) = S_j\}$$

$$S_j = \text{average}(X_j)$$

$$X_j = \{x \in X \mid \phi_S(x) = S_j\}$$

Now

$$\begin{aligned} \text{Find } S_j^* &= \underset{S \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{x \in X_j} \|x - S\|^2 \\ &= \text{average}(X_j) \end{aligned}$$

Both

(1a) assignment

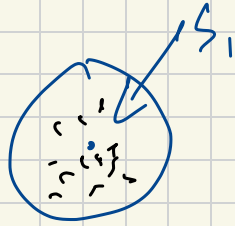
(1b) average

$\Rightarrow$ decrease $\text{cost}_r(X, S)$

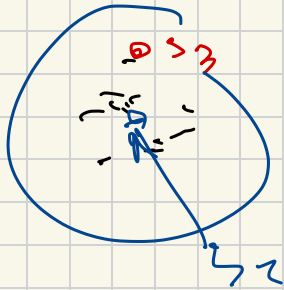
no infinite loop.

Lloyd's Algo

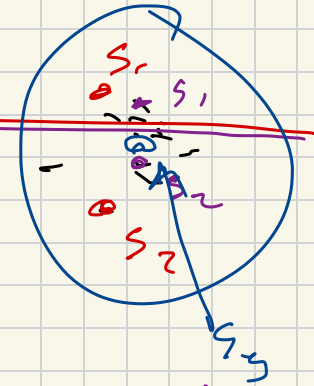
can get stuck in local optimum
 $X, k=3$



S_3



blue cost lower



purple sites
Lloyd's steps.

How to Initialize Lloyds?

1. Random $S \sim X$.

2. $S \leftarrow \text{Gonzalez}(X)$

3. k -Means $\neq f$

4. Run (say Gonzalez $\rightarrow$ Lloyds) w/ $2k$ clusters
Then compress down to k . $k \log k$

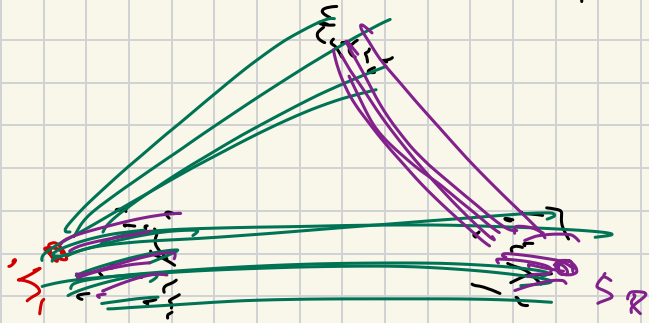
k-Means ++

①. Initialize $s_1 \in X$

1. for $j = 2$ to k

choose $s_j \sim X$ proportional

$$\text{to } \|x - \phi_{s_{j-1}}(x)\|^2 = w_x$$



n items $i=1 \dots n$ weights $w_1, \dots, w_n$

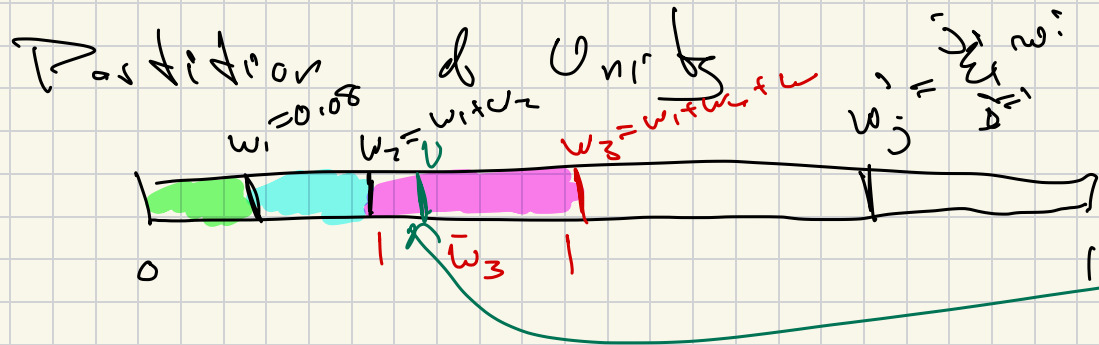
$$w_i > 0$$

Select item i proportion. / w_i

$$P_i = \frac{w_i}{\bar{w}} = \bar{w}_i \quad W = \sum_{i=1}^n w_i$$

$$\bar{w}_i \in [0, 1]$$

$$\sum_{i=1}^n \bar{w}_i = 1$$



uniform $u \in U_{unif}([0, 1])$

$$u \in [0, 1]$$