# L9: Assignment-based Clustering (e.g. k-means)

Sep 17, 2025

Data Mining

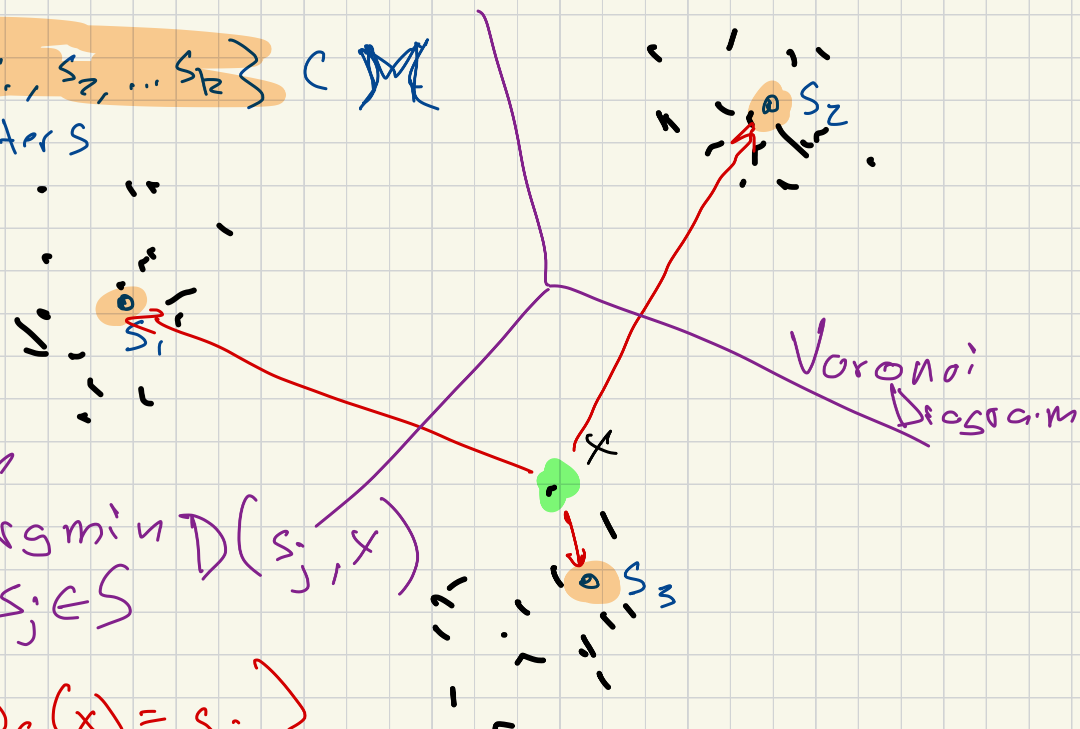
Jeff M. Phillips

**Input**: data $X \subset M$ $\qquad k = 3$

distance $D : M \times M \to \mathbb{R}_{\geq 0}$

**Output**: Set $S = \{s_1, s_2, \dots s_k\} \subset M$

sites / centers




Mapping Function

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

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

defined implicitly

# Variants Assignment-based Clustering

Define cost of clustering, want to minimize.

- **k-means** $\text{Cost}_2(X, S) = \frac{1}{|X|} \sum_{x \in X} D(x, \phi_S(x))^2$

  Lloyd's Algo, $X \subset \mathbb{X} = \mathbb{R}^d$, $D(x,s) = \|x-s\|_2$

- **k-median** $\text{Cost}_1(X, S) = \frac{1}{|X|} \sum_{x \in X} D(x, \phi_S(x))$

  better w/ outliers, not great Algo.

- **k-mediod** $\text{Cost}_1(X, S)$, but $S \subset X$.

- **k-center** $\text{Cost}_\infty(X, S) = \max_{x \in X} D(x, \phi_S(x))$

  outliers, Gonzalez Algo.

# Gonzalez Algo for k-center

Input: $X \in \mathcal{M}$, distance $D$ (metric), $k$ = # clusters

$S_j = \{s_1, s_2, \dots s_j\}$

$j \in [1, \dots k]$

## Gonzalez

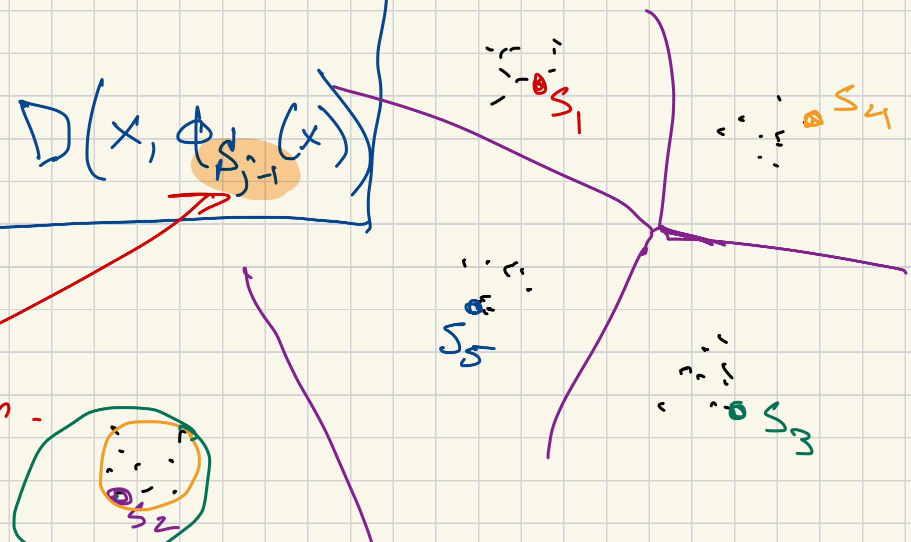
0. $s_1 \in X$ (arbitrarily?) $S_1 = \{s_1\}$

1. for $j = 2$ to $k$

   Set $s_j = \underset{x \in X}{\arg\max}\, D(x, \phi_{S_{j-1}}(x))$

2. Return $S = S_k$

Distance to set sites already chosen.



# Lloyd's Algorithm for k-means

Input $X \subset \mathbb{R}^d$, $D = \|\cdot - *\|_2$, $k$

## Lloyd's Algo

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

1. repeat

   1a. For all $x \in X$, set $\phi_S(x) = \operatorname{argmin}_{s_j \in S} \|x - s_j\|$

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

2. until ($S$ unchanged)

S1
S2
S2
S3
S31
S3
S4
S21

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

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

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

Find $s_j^* = \underset{s \in \mathbb{R}^d}{\arg\min} \sum_{x \in X_j} \|x - s_j\|^2$

$= \text{average}\{x \in X_j\}$

Both (1a) assignment ⟹ decrease $\text{Cost}_2(X, S)$
(1b) average

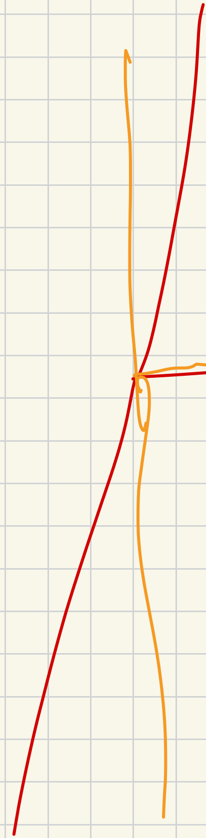
⟹ Lloyds Algo will terminate.

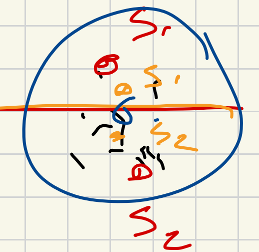
$$\mathrm{cost}_\tau(X, S) = \sum_{x \in X} \sum_{s_j \in S} \begin{pmatrix} \|x - s_j\|^2 & \text{if } \phi_S(x) = j \\ & \text{or } 0 \end{pmatrix}$$

$$= \sum_{s_j \in S} \sum_{x \in X_j} \|x - s_j\|^2$$

$k = 3$

Lloyds can get stuck in local opt.



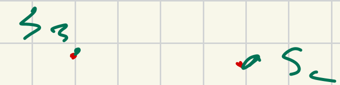

blue is better

orange terminates.

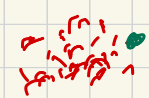
# How to initialize Lloyds?

1. Random restart, w $S \sim X$ at random

2. Gonzalez Algo. (deterministic)




3. k-means++



## k-means++

0. Initialize $s_1 \in X$

1. for $k = 2$ to $n$

   Choose $s_j \sim X$ proportional to

$$\|\phi_{S_{j-1}}(x) - x\|^2$$