L12: Assignment-based Clustering

k-means, k-center, k-median, k-mediod

Feb 19, 2025

Jeff M. Phillips

Input: 1. $X \subset \mathcal{X} = \mathbb{R}^d$ $\quad X = \{x_1, x_2, \dots x_n\}$

data point $x_i \in \mathbb{R}^d$

2. Distance $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$

metric

Lloyds requires $\mathcal{X} = \mathbb{R}^d$, $D = \|\cdot - \cdot\|_2$

Goal: $S = \{S_1, S_2, \dots S_k\}$ ← clusters

$S_j < X$ subsets

$S_i \wedge S_j = \phi$ (hard clustering)

$\bigcup_j S_j = X$

Assignment-based Clustering

k is part of input

Clusters $S_1, S_2, \dots S_k$

Centers $C = \{c_1, c_2, \dots c_k\} \subset \mathbb{R}^d$

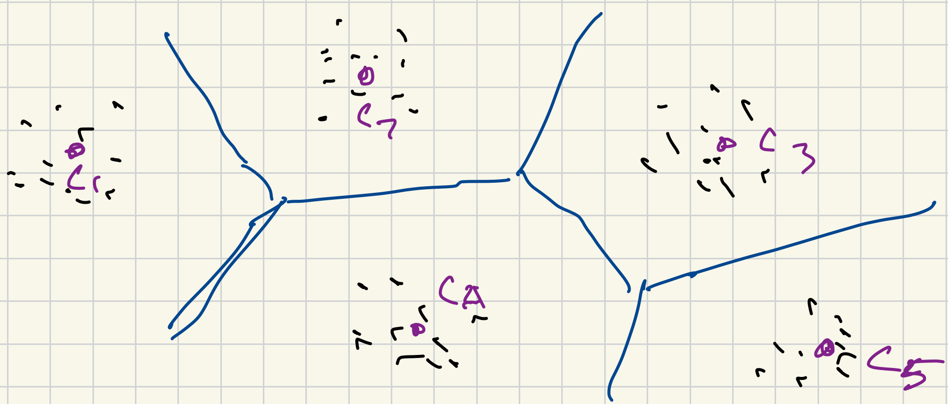
↑ representative of S_i

Nearest Neighbor:

$\phi_C : \mathbb{R}^d \to C$

$$\phi_C(x) = \arg\min_{c_j \in C} D(x, c_j)$$

X



Goal Find $C = \{c_1, c_2, \ldots c_k\}$

Formulations

k-means : minimize $\sum_{x \in X} D(x, \phi_C(x))^2$

Lloyd's D = Euclidean

k-center : minimize $\max_{x \in X} D(x, \phi_C(x))$

Gonzalez

k-median : minimize $\sum_{x \in X} D(x, \phi_C(x))$

k-medoid : minimize $\sum_{x \in X} D(x, \phi_C(x))$

$C \subset X$

Gonzalez Algo for k-center

Build centers C_j incrementally $C_1 \to C_2 \to \dots \to C_k$

$|C_j| = j$ $\quad C_1 = \{c_1\}$

find outliers!

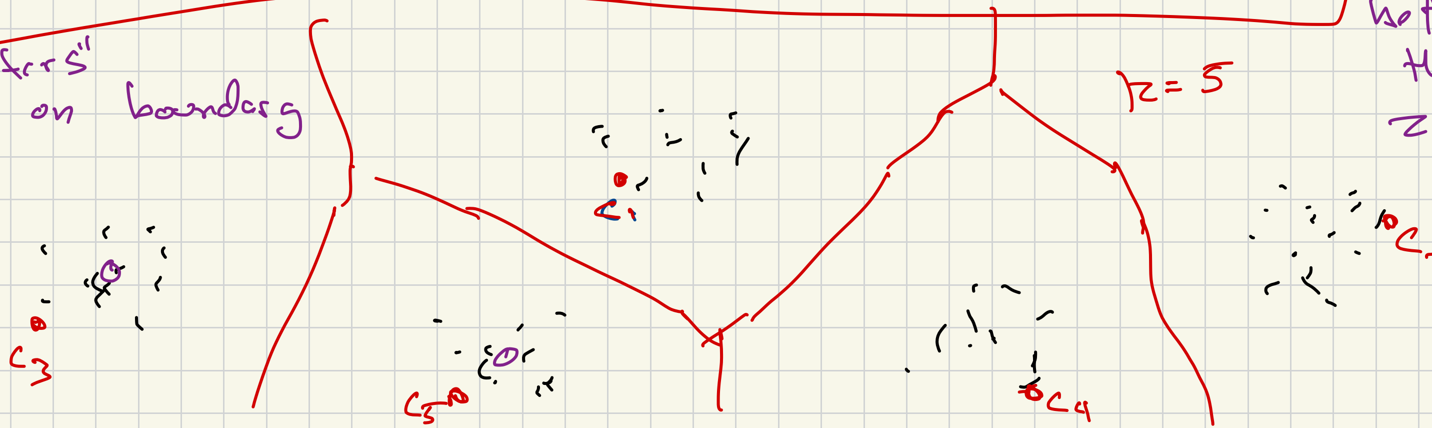
0. Choose c_1 arbitrarily $\to C_1 = \{c_1\}$
1. for $j = 2$ to k

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

Gives 2-approx. of optimal.

NP-hard to do better than 2-approx

"centers" on boundary



Lloyd's Algorithm for k-means

$D =$ Euclidean

0. $\boxed{\text{Choose } k \text{ pts} \to C}$ Do smartly / Random

1. repeat

   1a. For all $x \in X$, find $\boxed{\phi_C(x) \to c_j \Rightarrow x \to S_j}$ implicit

   1b. For all $j \in 1 \ldots k$, let $\boxed{c_j = \text{average}(S_j) = \frac{1}{|S_j|} \sum_{x \in S_j} x}$

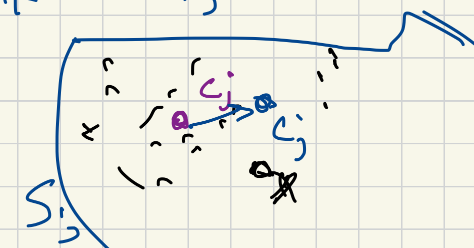
   $\Sigma \operatorname*{argmin}_{z \in \mathbb{R}^d} \sum_{x \in S_j} \|z - x\|^2$

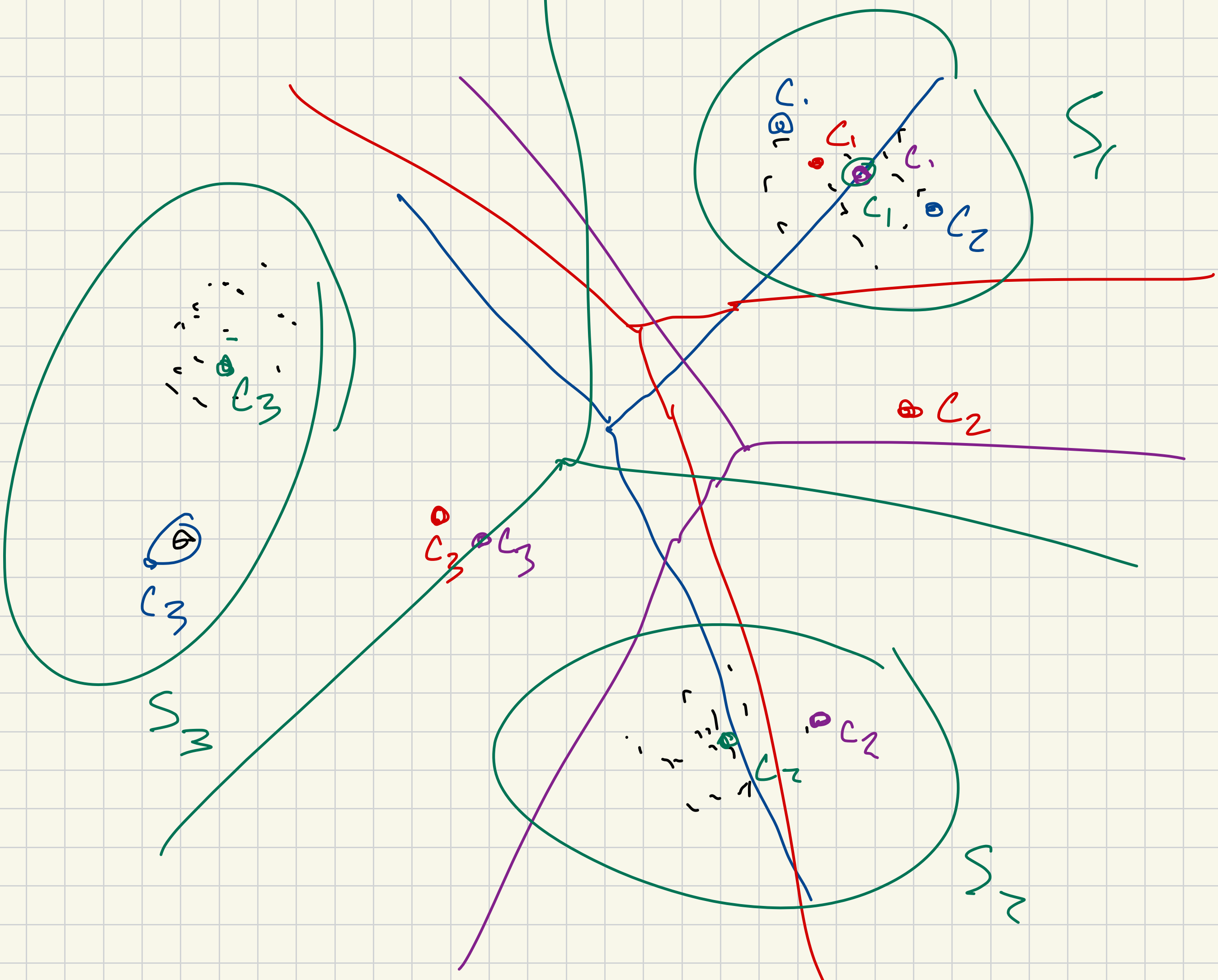
2. until (S unchanged

   or change is small)

in practice

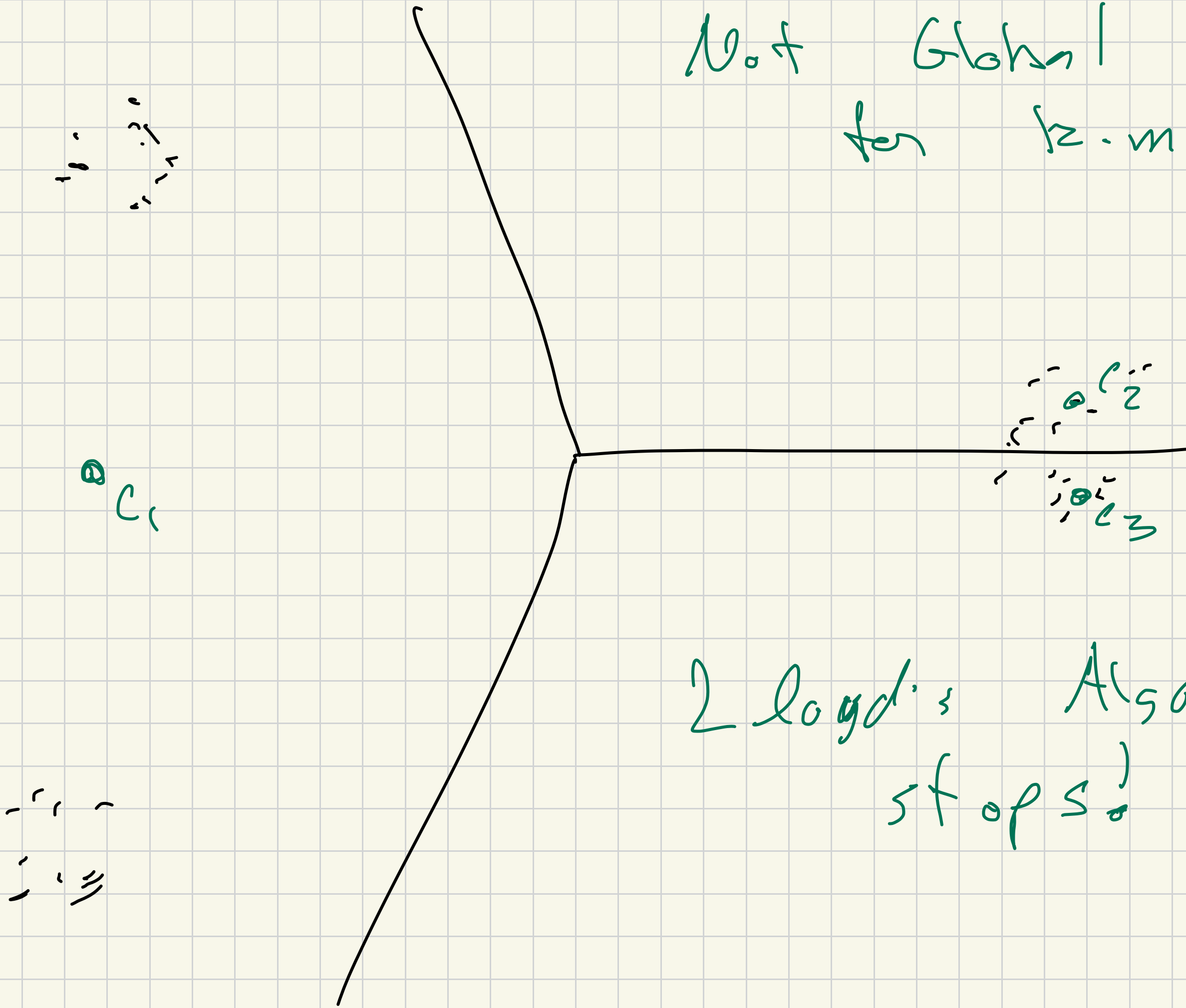
$k = 20$ iterations





C_1
C_1
C_1
C_1
C_2
S_1
C_2
C_3
C_2
C_3
C_3
S_3
C_2
C_2
S_2

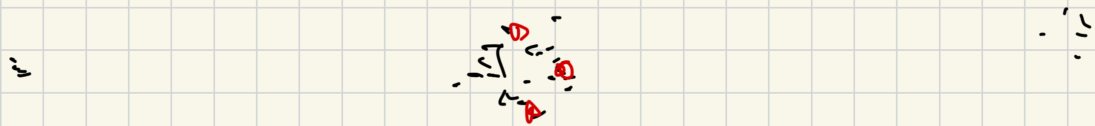
Not Global opt for k-means



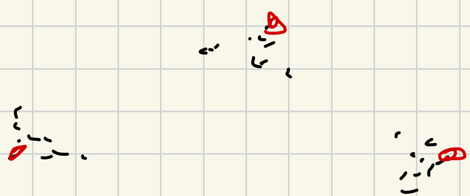
Lloyd's Also stops!

Initialize Lloyd's Algo for k-Means

1. Init Choose C at random.



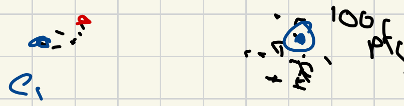
2. Init $C \leftarrow \text{Gonzalez}(X)$



always choose

3. k-means++

K-means++



0. Choose c_1 arbitrarily $c_i \in X$

1. for $j = 1$ to k

Choose c_j from X w/ probability proportional to

$x_i \in X$

$$v_i = D(x_i, \phi_C(x_i))^2$$

$V = \sum_{i=1}^{n} v_j$ select $x_i \sim X$ w.p. $\text{prob}(x_i) = \frac{v_i}{V}$

Implement $u \sim \text{Unif}(0,1)$ $V_i = \sum_{i=1}^{i} v_i$