Data Mining

L9 – Assignment-based Clustering

<u>Input</u>
- $X \subset \mathbb{R}^d$

  data point $x_i \in X$
- distance $d: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

  metric

<u>Goal</u>: $S = \{S_1, S_2, \ldots S_k\}$ ← clusters

$S_i \subset X$

$S_i \cap S_j = \phi$

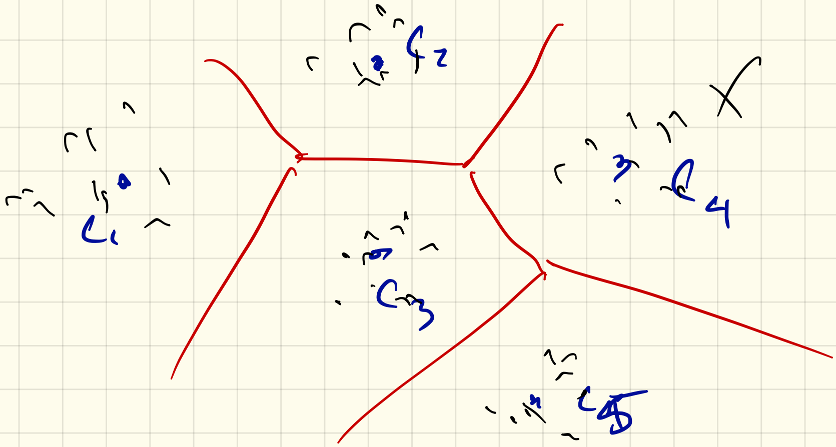
$\bigcup S_i = X$

Clusters $S_1, S_2, \ldots S_k$

center $C = \{c_1, c_2, \ldots c_k\}$ ← representative pts.

Nearest-Neighbor function $\phi_C : \mathbb{R}^d \to C$

$$\phi_C(x) = \underset{c_i \in C}{\arg\min}\, d(x, c_i)$$



Goal : $\boxed{\text{Find } C = \{c_1, c_2, \dots c_k\}}$ k given

Formulations

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

Lloyds $\quad d = \text{Euclidean}$

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-mediod : minimize $\sum_{x \in X} d(x, \phi_C(x))$

$C \subset X$

Gonzalez Alg. for k-center

Build C incrementally $\quad C_1 \to C_2 \to \dots \to C_k = C$

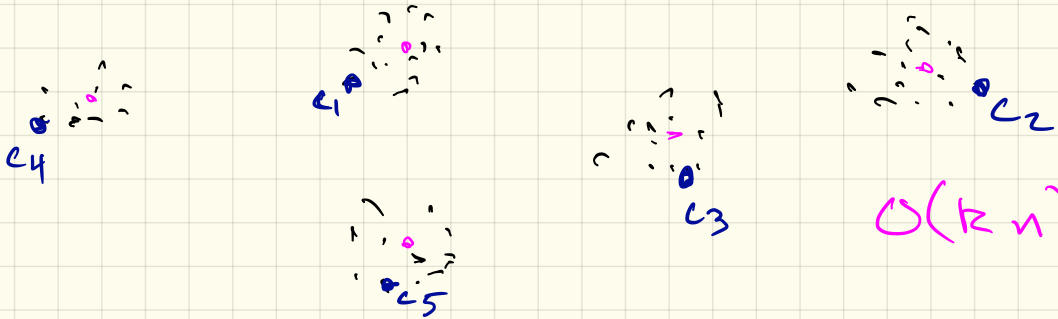
$|C_i| = i$

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

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

if d metric

↳ output 2-approx of optimal



$O(kn)$ time

Lloyd's Algo for k-means

$d =$ Euclidean

0. $\boxed{\text{Choose } k \text{ pts} \to C_i}$

1. repeat

$S_i = \{x \in X \mid \phi_C(x) = c_i\}$

usually 2-20 steps

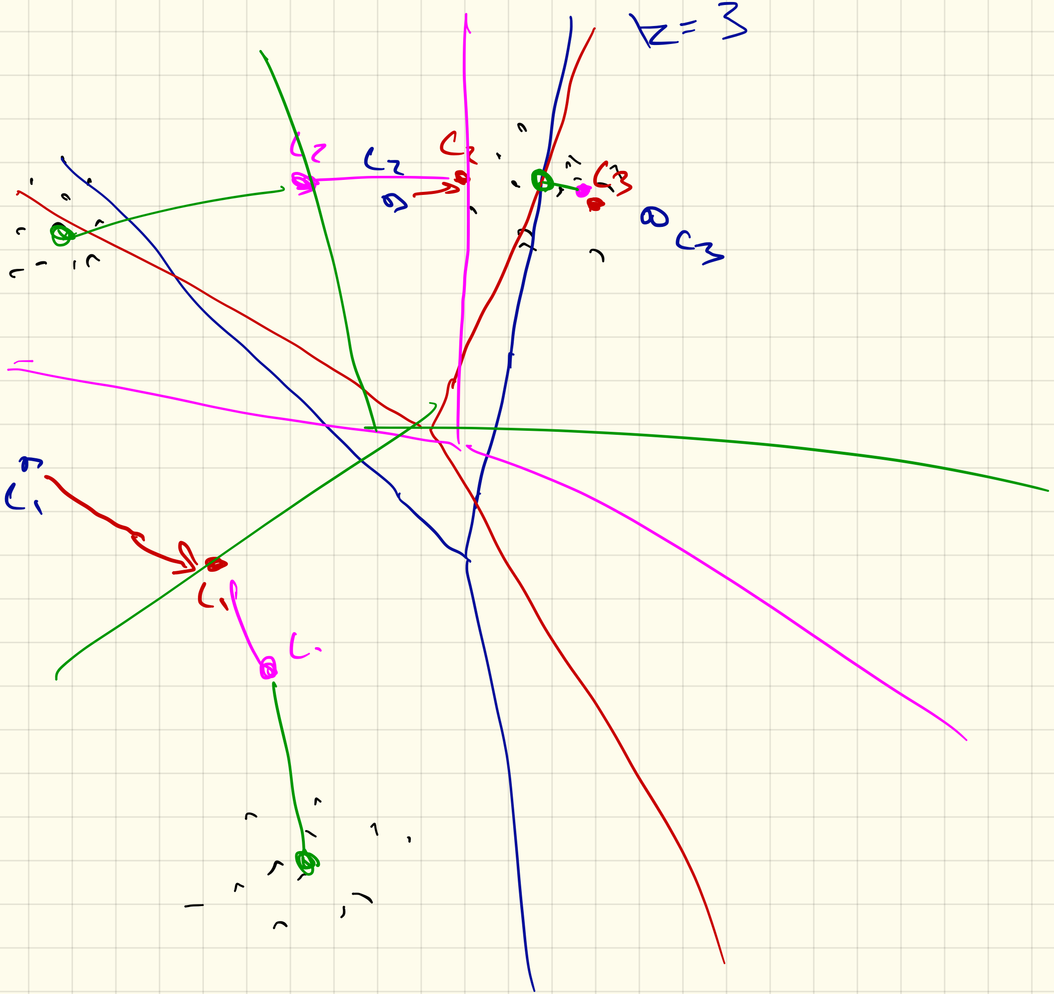
1a. For all $x \in X$, find $\phi_C(x) \to S_1, S_2, \dots S_k$

1b. For all $i \in 1..k$, let $c_i = \text{average}(S_i)$

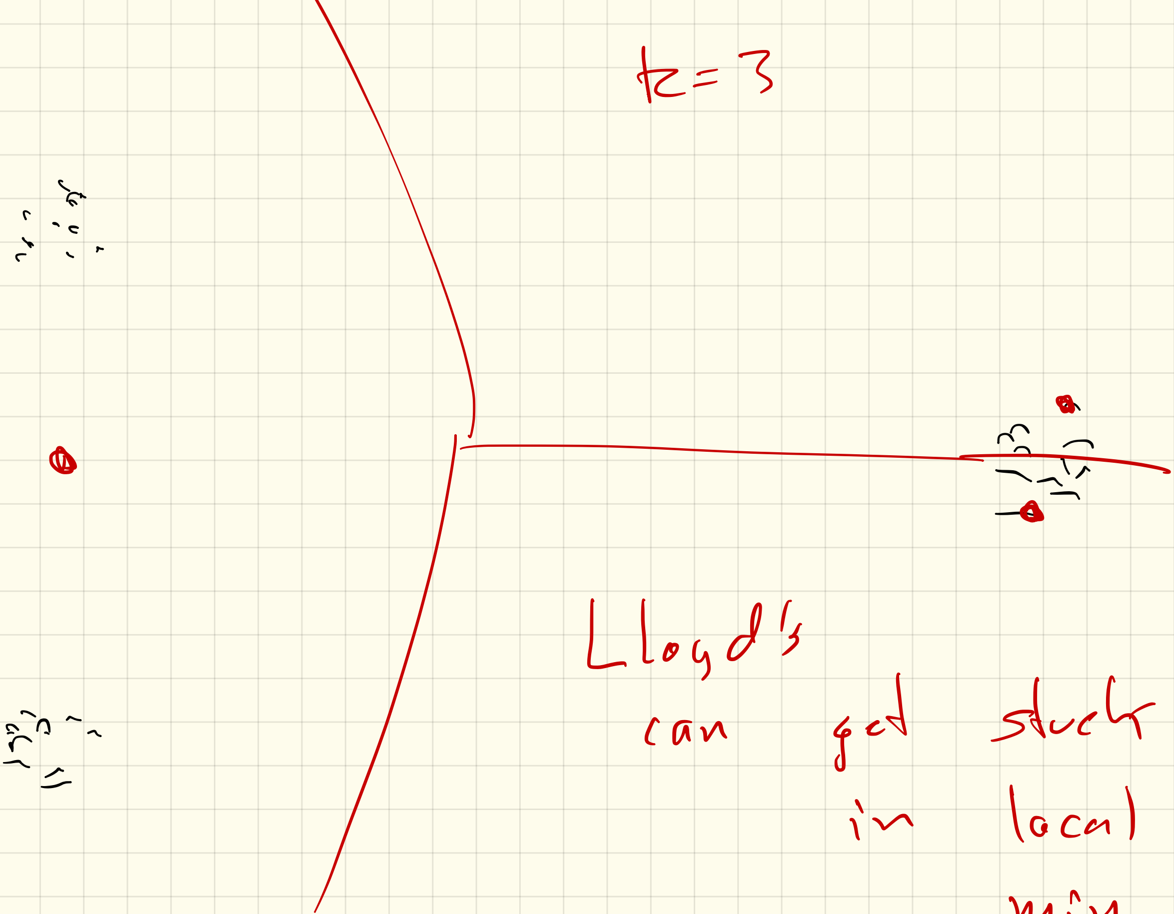
$$c_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$$

2. until (S unchanged or change is small)

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


k = 3
c2
c2
c2
c3
c3
c1
c1
ci

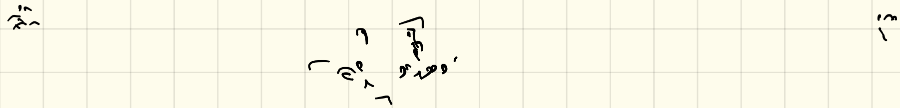
$k=3$



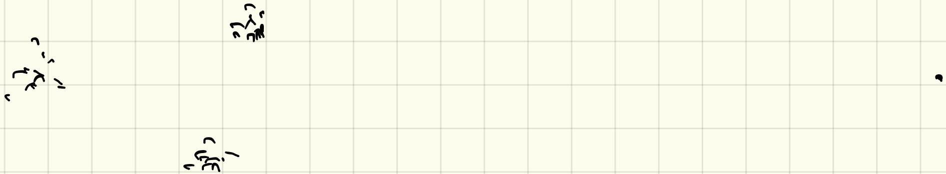
Lloyd's can get stuck in local min.

Choose Initial k Centers

1. Pick random subset $C \subset X$



2. Gonzalez Algo.



3. k-means ++

k-means ++

0. Choose c_1 arbitrarily $c_1 \in X$

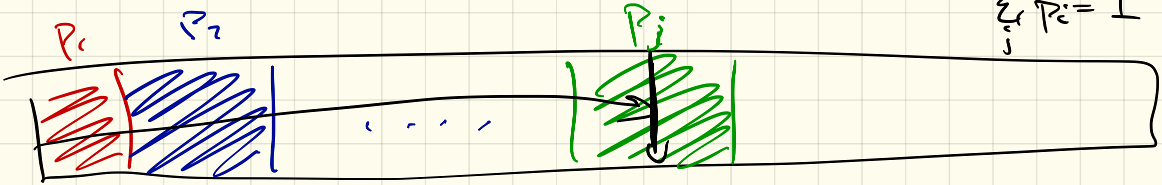
1. for $i = 2$ to k

   Choose c_i from X w/ prob proportional

$$v_j = d(x_j, \phi_{C_{i-1}}(x_j))^2$$

$$V = \sum_{j=1}^{n} v_j \qquad \text{prob}(x_j) = \frac{v_j}{V} = p_j$$

$$\sum_j p_j = 1$$



0 $\qquad$ $0.55 = u \in \text{Unif}(0,1)$ $\qquad$ 1