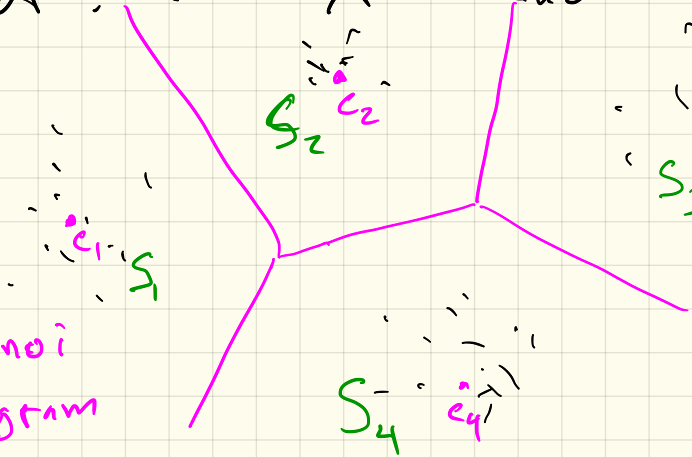
Assignment-based Clustering

Input: data X and dist $d: X \times X \to \mathbb{R}$



centers

$C = \{c_1, c_2, c_3, c_4\}$

k

Voronoi Diagram

$$\phi_C(x) = \arg\min_{c \in C} d(x, c)$$

$$S_j = \{x \in X \mid \phi_C(x) = c_j\}$$

$$X \to S_1, S_2, \ldots S_k$$

$$S_i \cap S_j = \emptyset$$

$$X = S_1 \cup S_2 \cup \ldots \cup S_k$$

$$\underline{\mathrm{Cost}_2(X, C)} = \sum_{x \in X} d(\phi_C(x), x)^2$$

$$\phi_C(x) = \arg\min_{c \in C} d(c, x)$$

$C \leftarrow$ minimize Cost_2

k-means clustering formulation

$$\underline{\mathrm{Cost}_\infty(X, C)} = \max_{x \in X} d(\phi_C(x), x)$$

k-center (Gonzalez)

$$\underline{\mathrm{Cost}_1(X, C)} = \sum_{x \in X} d(\phi_C(x), x)$$

k-median

<u>k-mediod</u>

minimize $\mathrm{Cost}_1(X, C)$

s.t. $C \subset X$ (subset)

Gonzalez Algo. k-Center

poorly w/ outliers

X, d metric

$\hat{C} \leftarrow C^* \leftarrow$ optimal $= \arg\min \text{Cost}_\infty(X, C)$

$C_1 = \{c_1\}$

0. Choose $c_1 \in X$ arbitrarily

$C_{j-1} = \{c_1, c_2, \ldots c_{j-1}\}$

1. for $j = 2$ to k

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

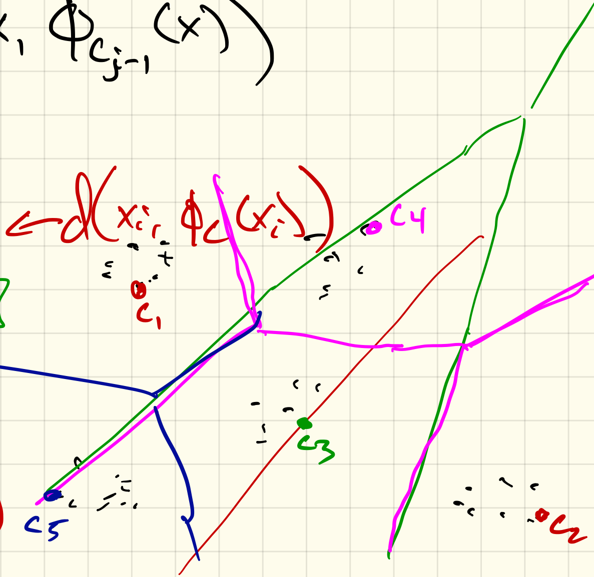
x_i

X: 1 2 3 4 … i … n

d 0 .3 .7 .8 $\boxed{1.2}$ .3

update 1 1 1 2 2 2 1

0 .3 .6 0



Provides 2-approximation

$\text{Cost}_\infty(X, \hat{C}) \leq 2 \cdot \text{cost}_\infty(X, C^*)$

k-means clustering

Lloyds Algo

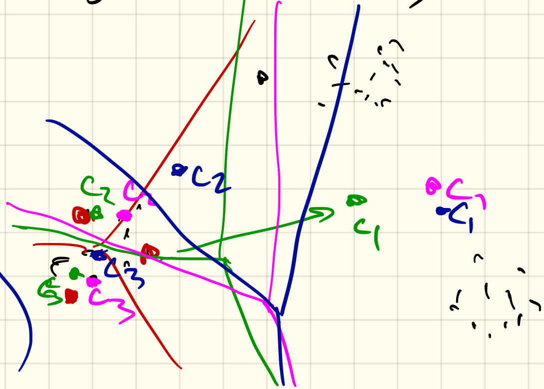
dist $d = \|\cdot - \cdot\|_2$ = Euclidean

0. Choosing k pts $C \subset X$ ← randomly / Gonzalez / kmeans++

1. repeat
   1a. for all $x \in X$ assign $\phi_C(x)$ $\quad S_j = \{x \in X \mid \phi_C(x) = c_j\}$
   1b. for all $j \in [k]$ let $c_j = \text{average}(S_j)$

2. until C is fixed

$$\text{Cost}_2(X, C) = \sum_{x \in X} d(x, c_x) = \sum_j \left(\sum_{x \in S_j} d(x, c_j) \right)$$



k-means++ (D^2-sampling)

↳ initialize Lloyds Algo

$C_j = \{c_1, c_2, \dots, c_j\}$

0. Choose $c_1 \in X$ arbitrarily $\quad C_1 = \{c_1\}$

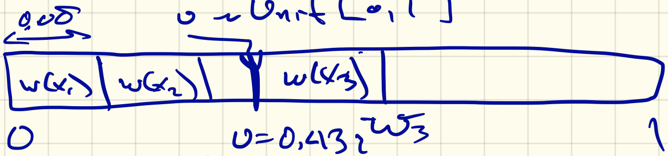
1. for $j = 2$ to k $\qquad$ M4D Sec 2.4

Choose c_j from X w/ prob proportional to

$$d(x, \phi_{C_{j-1}}(x))^2$$

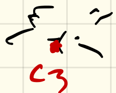
Partition of Unity

$u \sim \mathrm{Unif}[0,1]$



$$w(x_i) = \frac{d(x_i, \phi_{C_{j-1}}(x_i))^2}{\sum_{x'} d(x', \phi_{C_{j-1}}(x'))^2}$$

$$W_j = \sum_{i=1}^{j} w(x_i)$$



Choosing k

"elbow technique"