FoDA : Clustering

L22 : Voronoi Diagrams

What is Clustering?

Input Set of objects $X = \{x_1, x_2, \dots x_m\}$

Distance $D : X \times X \to \mathbb{R}^+$

(this class : $X \subset \mathbb{R}^d$, Euclidean $D(x_1, x_2) = \|x_1 - x_2\|$)

usually part of input

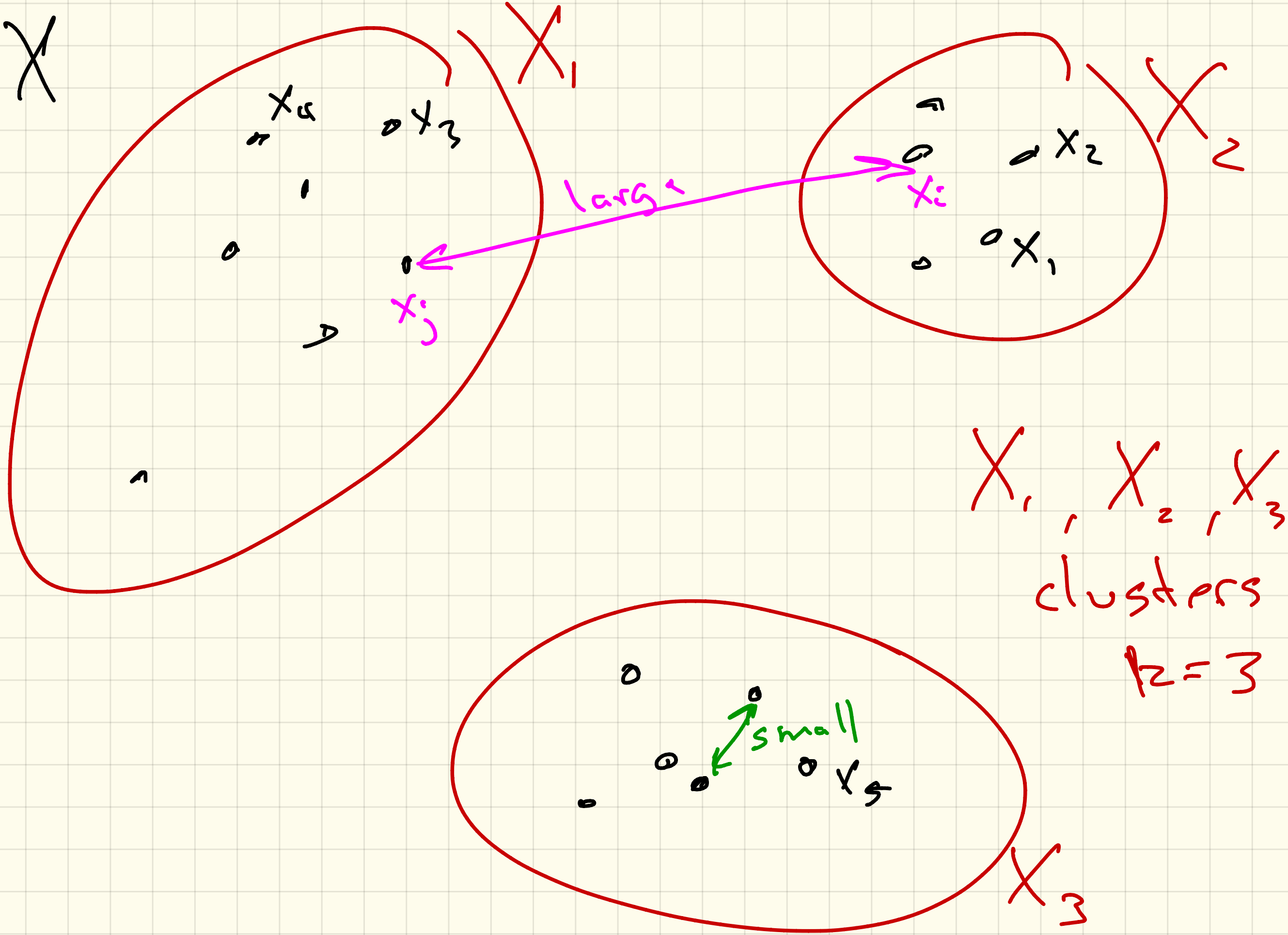
if ill-defined $\hookrightarrow$ trouble

Goal k subsets $\{X_1, X_2, \dots X_k\}$

$X_i \subset X$

$x_i, x_i' \in X \longrightarrow D(x_i, x_i')$ small

$x_i \in X$ $x_j \in X_j$ $i \neq j \longrightarrow D(x_i, x_j)$ large



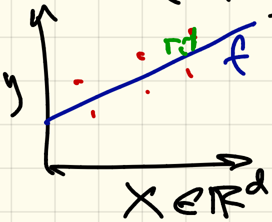
X_1, X_2, X_3 clusters

$k=3$

Regression

Input

$(X, y) \to f(x_i \in X) = \hat{y}_i$



measure $r_i = y_i - \hat{y}_i$

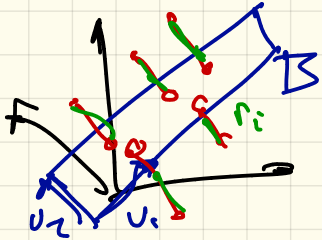
min $\sum_i r_i^2$

Dimensionality Reduction

PCA

$A = X \subset \mathbb{R}^d \to B \subset \mathbb{R}^d$

rank = k



min $\sum_i r_i^2$

$\|A - A_k\|_F^2 = \|A - \pi_B(A)\|_F^2$

$= \sum_{i=1}^{n} \|a_i - \pi_B(a_i)\|^2$

Assignment-based Clustering

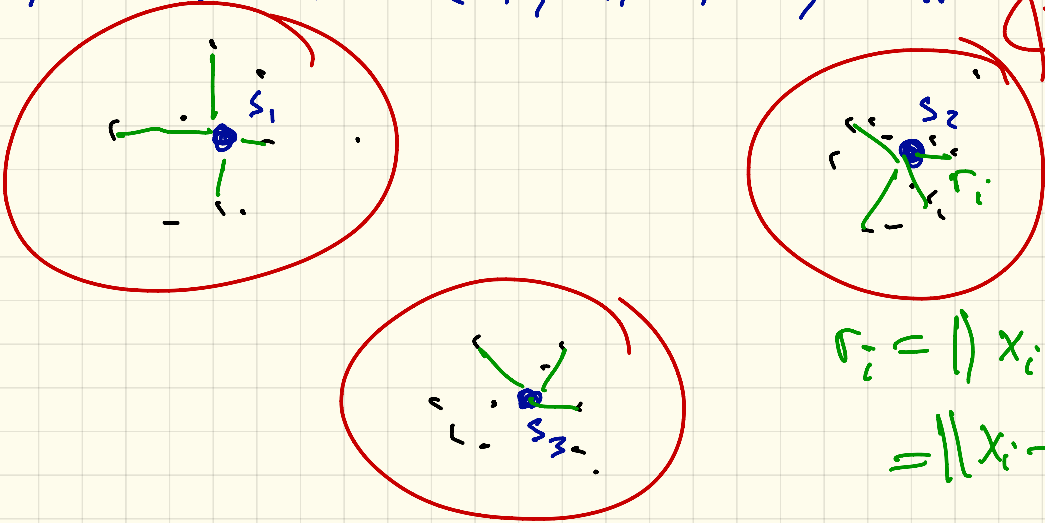
$X \subset \mathbb{R}^d$ $\quad D : \|\cdot - \cdot\|$ $\quad k$ clusters

$$D(x_i, x_j) = \|x_i - x_j\|$$

Model $\quad S = \{s_1, s_2, \dots, s_k\} \subset \mathbb{R}^d$

$$\phi_S(x) = \arg\min_{s_i \in S} \|x - s_i\|$$

maps to closest site



$$r_i = \|x_i - s_j\|$$

$$= \|x_i - \phi_S(x_i)\|$$

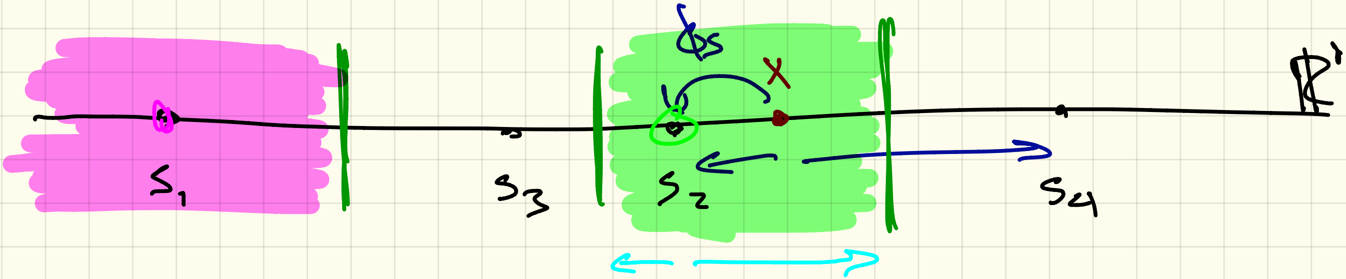
Post Office Problem

$$\phi_S(x) = \underset{s_i \in S}{\operatorname{argmin}} \|x - s_i\|$$

$$\phi_S : \mathbb{R}^d \to S$$

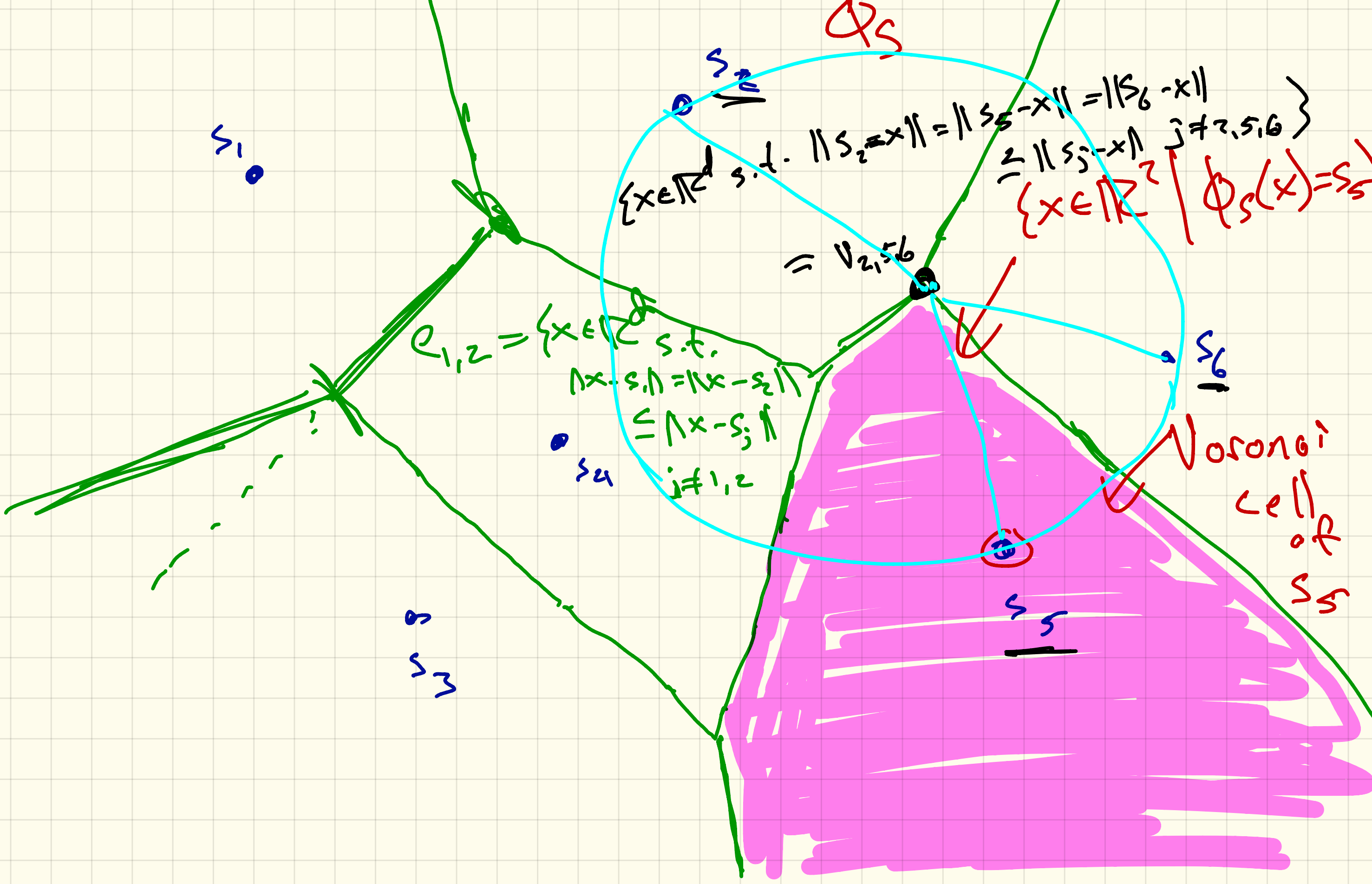
$|S| = k$

$d=1$ Solvable in $O(\log k)$ time



Voronoi Diagram

$d=2$



Voronoi Diagram in $\mathbb{R}^2$

→ "Complexity" is $O(k)$

vertices & # edges

→ Compute in $O(k \log k)$ time

→ Solve $\phi_S(x)$ in $O(\log k)$ time

Bad News not true for $d \geq 3$

complexity in $\mathbb{R}^3, \mathbb{R}^4$ $O(k^2)$

complexity in $\mathbb{R}^d$ $O(k^{\lceil d/2 \rceil})$ ← curse of dimensionality

$$\phi_S(x) = \underset{s_j \in S}{\arg\min} \| s_j - x \|$$

$S = \{s_1, \dots, s_k\}$

in high-d

$\Downarrow$

very high

complexity

0. $s = s_1$

   $m = \| x - s_1 \|$

1. for $i = 2$ to k

   if $m > \| x - s_i \|$

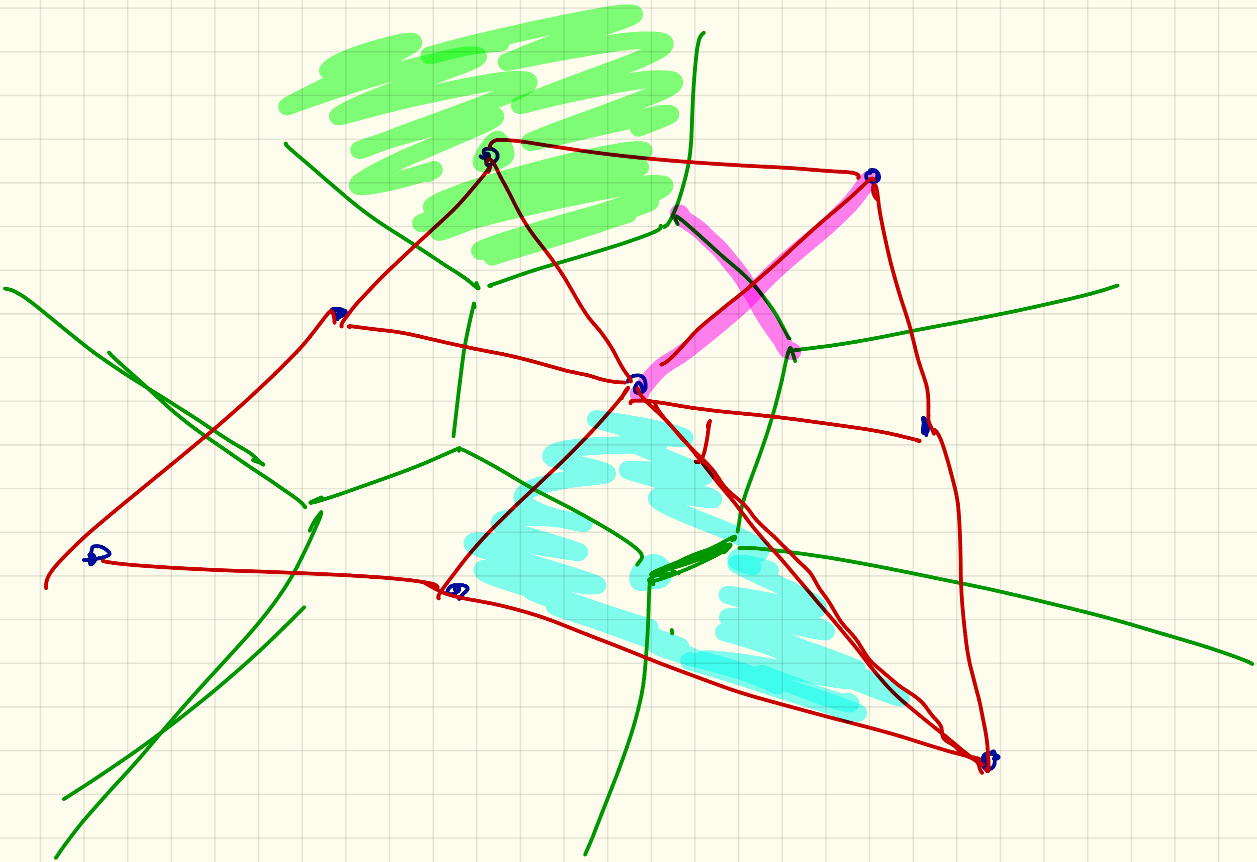
   $s = s_i$

   $m = \| x - s_i \|$

2. return s

$O(k)$ time

Delaunay Triangulation



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

Assignment-Based Clustering

Goal ~~S~~ Find $S = \{s_1, \dots s_k\} \subset \mathbb{R}^d$

$$\underset{S}{\text{minimize}} \ f\{\|x_i - \phi_S(x_i)\|\}$$

find $S = \{s_1 \dots s_k\}$ to try to

$$\underset{S}{\text{minimize}} \sum_i \|x_i - \phi_S(x_i)\|^2$$

$f = \sum_i r_i \leftarrow$ k-median $\leftarrow$ k-mediod ($S \subset X$)

$f = \sum_i r_i^2 \leftarrow$ k-means

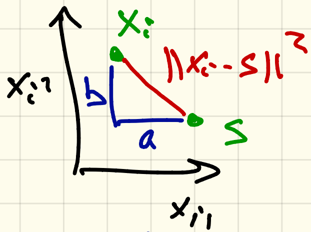
$f = \max r_i \leftarrow$ k-center

Dim Reduction (PCA)	k-means clustering
k basis functions $V_B = \{v_1, v_2 \dots v_k\}$ $\lVert v_i \rVert$ $\langle v_i, v_j \rangle = 0$	k sites $S = \{s_1, s_2 \dots s_k\}$
Assignment: projection $\pi_B(x) = \operatorname{argmin}_{b \in B} \lVert b - x \rVert$	Assignment: $\phi_S(x)$ NN $= \operatorname{argmin}_{s_j \in S} \lVert x - s_j \rVert$
Goal: $\sum_i r_i^2 = \sum_i \lVert x_i - \pi_B(x_i) \rVert^2$	Goal: $\sum_i r_i^2 = \sum_{i=1}^{n} \lVert x_i - \phi_S(x_i) \rVert^2$

Find

Set $X \in \{x_1, \dots x_n\}$ $\quad x_i \in \mathbb{R}^d$

Find s minimize $\sum_{i=1}^{n} \| s - x_i \|^2$



$$\| x_i - s \|^2 = \sum_{j=1}^{d} (x_{ij} - s_j)^2$$

$\|x_i - s\|^2 = a^2 + b^2$

Soln: $s = \frac{1}{n} \sum_{i=1}^{n} x_i$

$$s_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$