FoDA L22

Clustering:

Voronoi Diagrams → Assignment-based Clustering

Nov 10, 2022

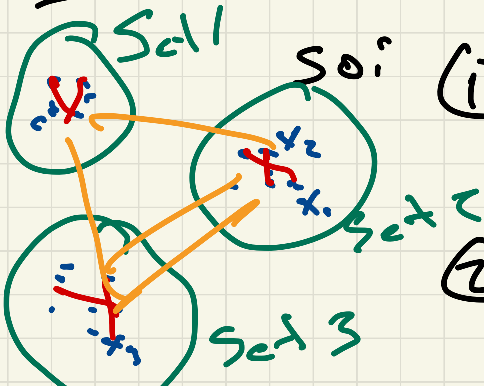
What is Clustering?

n objects (this class: set $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$)

distance function $d: X \times X \to \mathbb{R}_{\geq 0}$

(this class $d(x_i, x_j) = \|x_i - x_j\|$)

Goal: group objects into k sets



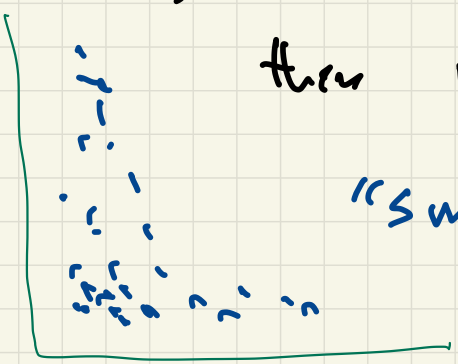
so: ① objects in the same set
↳ close.

② objects in different sets
↳ far.

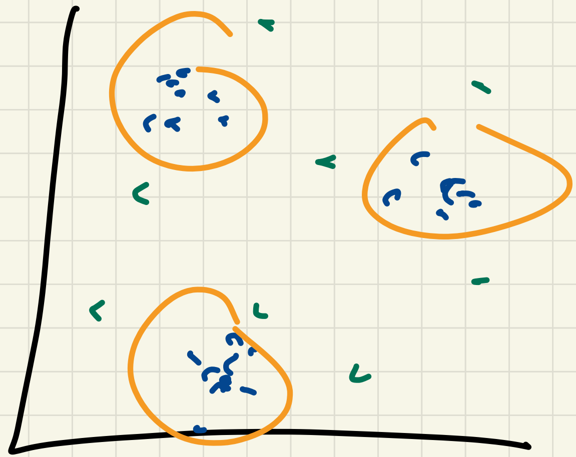
Clusterability

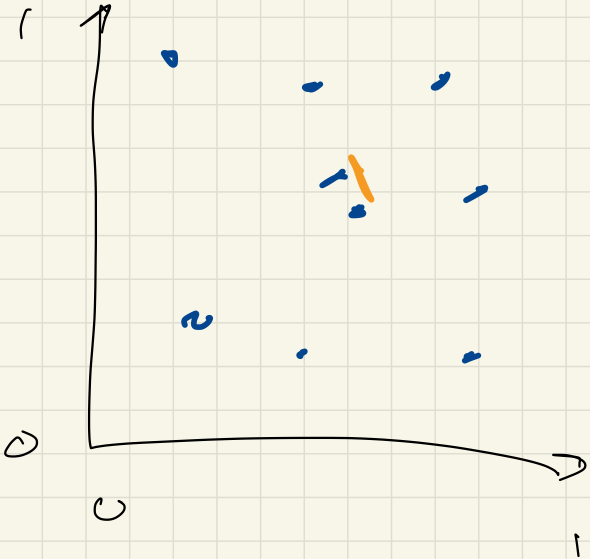
- When data is easily/naturally clusterable, most clustering algorithms work quickly and well.

- When data is not easily/naturally clusterable, then no algorithm will find good clusters.



"smear"




1
0
0
1

Clustering cost function (k-means)

data X, value k.

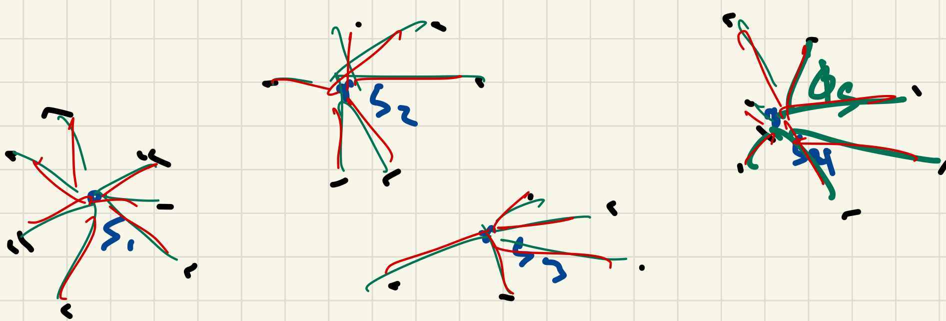
$$\text{cost}_2(X, S) = \sum_{i=1}^{n} \left(x_i - \cancel{\pi_S}(x_i)\right)^2$$

$\phi_S(x_i)$

find best S.

Project x_i onto closest $s_j \in S$

$S = \{s_1, \ldots, s_k\} \subset \mathbb{R}^d$ <- set k sites in $\mathbb{R}^d$



$4 = k$

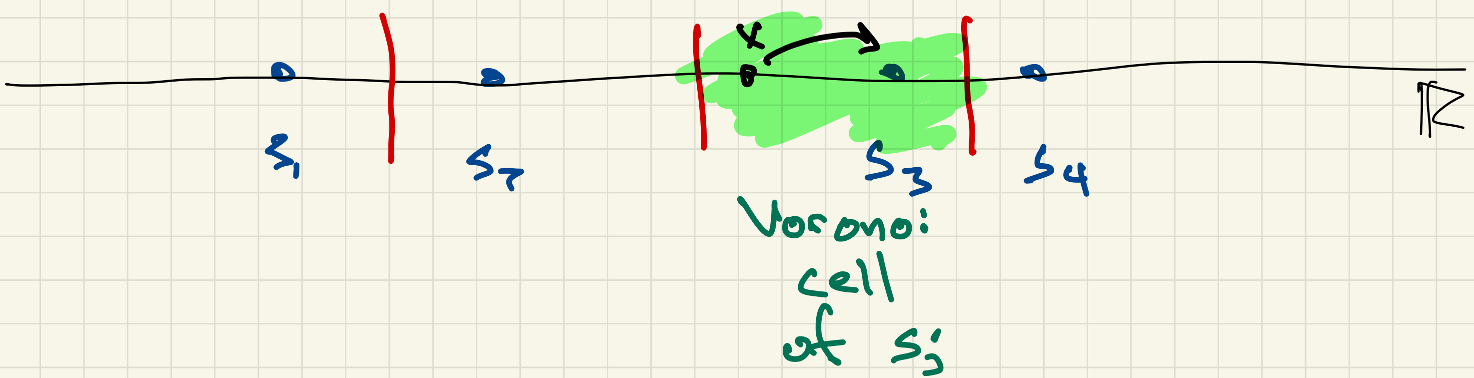
Voronoi Diagrams for set $S \subset \mathbb{R}^d$

$|S| = k$

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

what part of $\mathbb{R}^1$ maps to each s_i?

$d = 1$

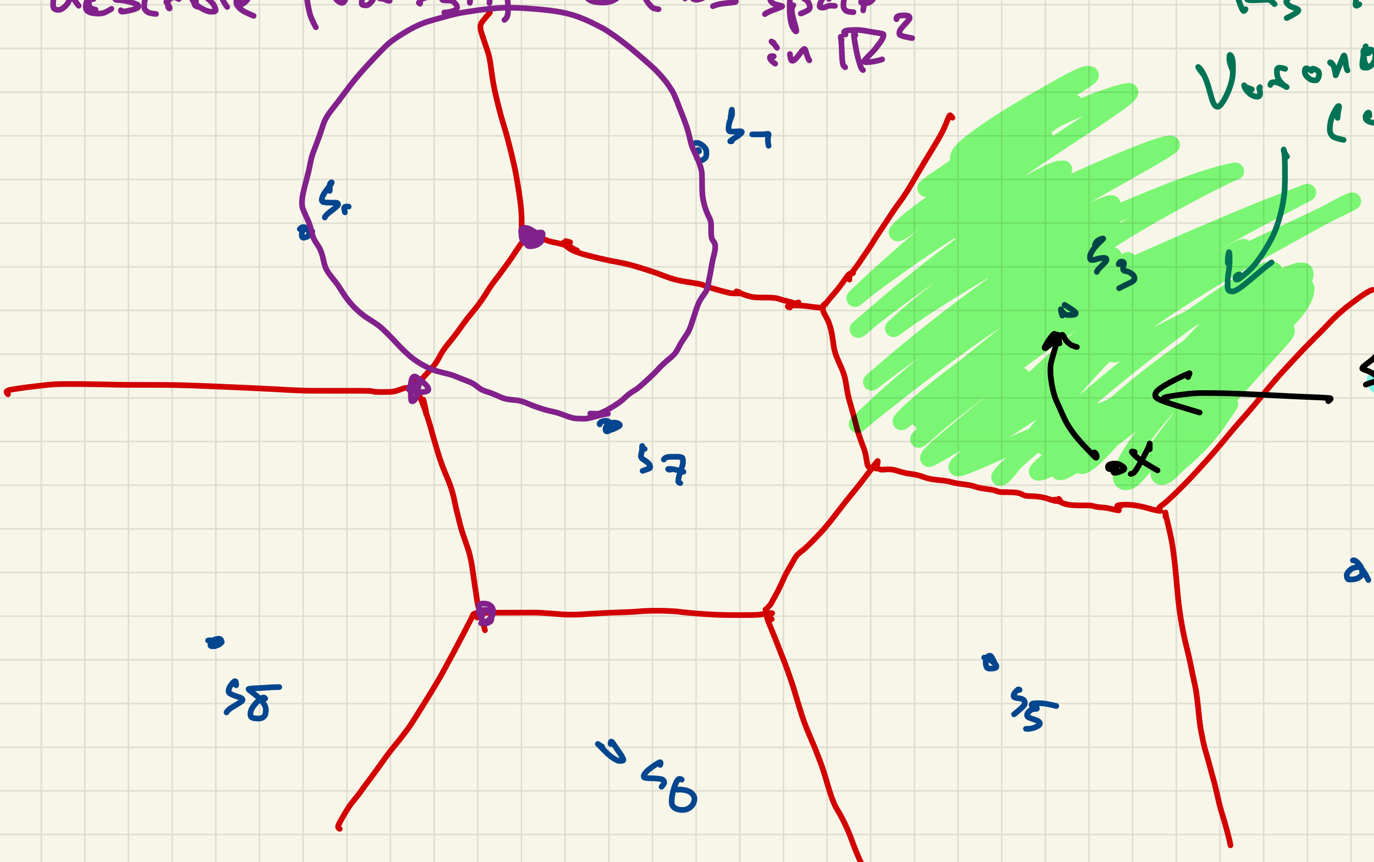


Voronoi Dgm in $\mathbb{R}^d$

describe $|\text{Vor Dgm}| = O(n)$ space in $\mathbb{R}^2$

$R_3 = \{x \in \mathbb{R}^d \mid \Phi_S(x) = s_3\}$

Voronoi cell



solve $\Phi_S(x)$ in $O(\lg n)$

Voronoi Diagrams in $\mathbb{R}^d$

size $\approx k^{\lfloor d/2 \rfloor}$

"Curse of Dimensionality"

Compute $\Phi_S(x) = \underset{s_j \in S}{\operatorname{argmin}} \|x - s_j\|$

$s^* = s_1$
for $j = 2$ to k
if $(\|x - s_j\| < \|x - s^*\|)$
$s^* := s_j$

$O(k)$

Assignment-based Clustering

$X \subset \mathbb{R}^d \qquad |S| = k$

k-means $\qquad S = \underset{|S|=k}{\arg\min} \sum_{j=1}^{n} \left(x_i - \phi_S(x_i) \right)^2$

k-center $\qquad S = \underset{|S|=k}{\arg\min} \max_{x_i \in X} \| x_i - \phi_S(x_i) \|$

k-median $\qquad S = \underset{|S|=k}{\arg\min} \sum_{i=1}^{n} \| x_i - \phi_S(x_i) \|$