L21: Clustering

Voronoi Diagrams → Assignment-based Clustering

Mar 30, 2026

EoDA

Jeff M. Phillips

What is Clustering

$X \subset \Omega$

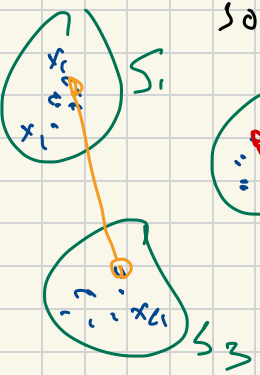
- n objects — a set $X = \{x_1, x_2, \dots x_n\} \subset \mathbb{R}^d$
- distance function $d : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$

trust distance as good modeling. $d(x_i, x_j) = \|x_i - x_j\|$

Goal: group objects into k sets — given as input

ex: $k = 3$

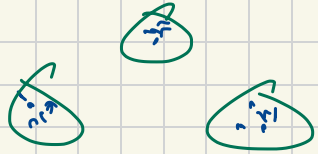
so:



(1) objects in same set $\Rightarrow$ $d(x_i, x_i')$ small.
x_i, x_i'

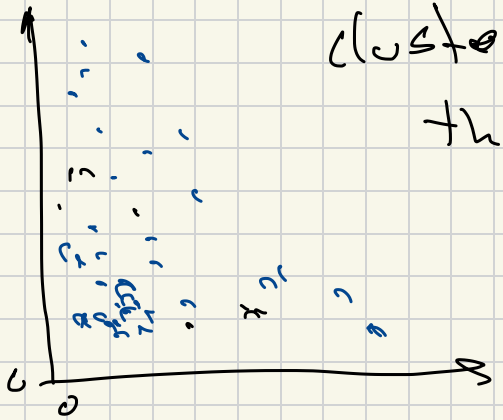
(2) objects in different sets $\Rightarrow$ $d(x_i, x_j)$ large.
x_i, x_j

Clusterability



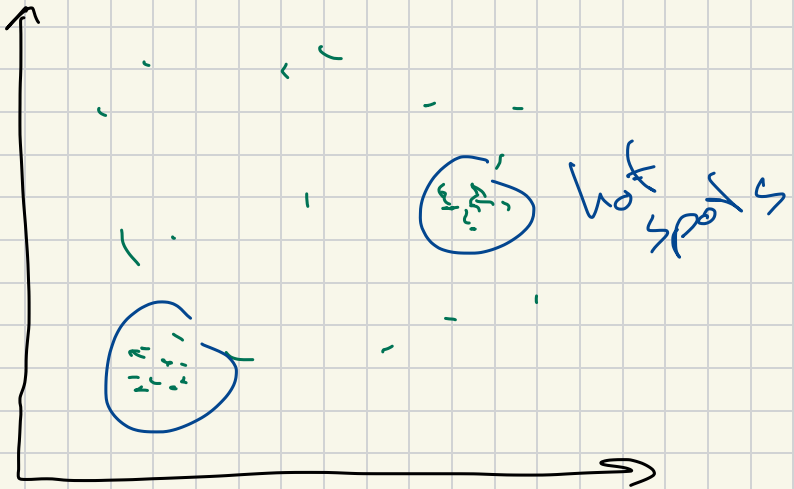
- When data is easily/naturally clusterable, most clustering algos. work well.

- When data is <u>not</u> easily/naturally clusterable, then no magical algorithm that will find a good clustering.



"smear"

"clusterings" in spatial data on a map



different

Decompose X in k sets each $x_i \in X$ in exactly 1 S_j

Clustering cost function (k-means)

Input data X, $d(x, x') = \|x - x'\|$, value $k > 1$

$$\text{cost}_2(X, S) = \sum_{x_i \in X} \|x_i - \phi_S(x_i)\|^2$$

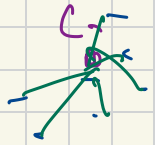
$S = \{S_1, S_2, \ldots S_j\}$

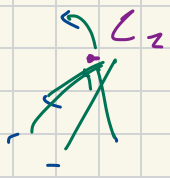
$S_j \subset X$

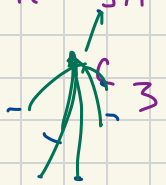
$\phi_S(x_i)$: similar to π_F (closest point on subspace F)

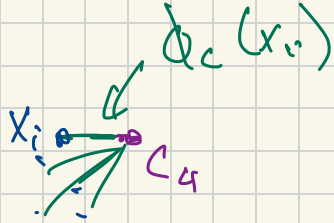
Sites $C = \{c_1, c_2, \ldots c_k\} \subset \mathbb{R}^d$ ← sites, centers

$\phi_C(x) = \underset{c_j \in C}{\text{argmin}} \|x - c_j\|$





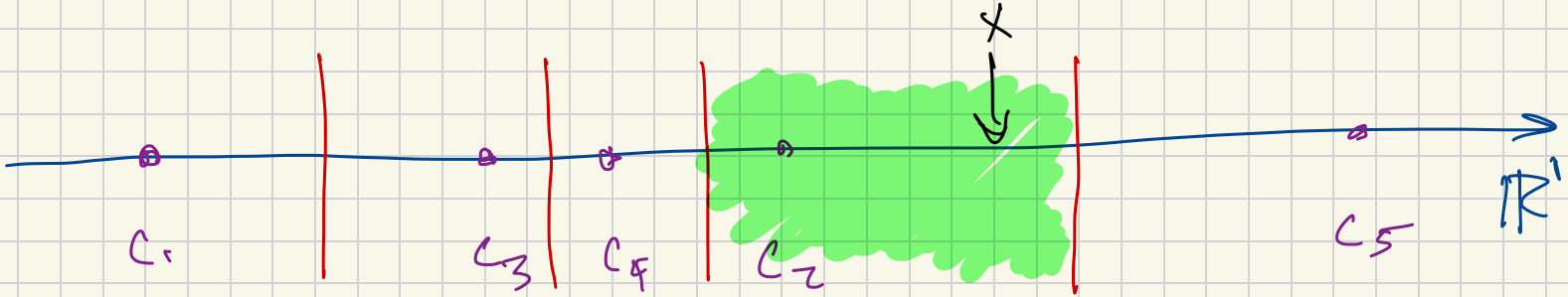




Voronoi Diagram

for set $C \subset \mathbb{R}^d$ $|C| = k$

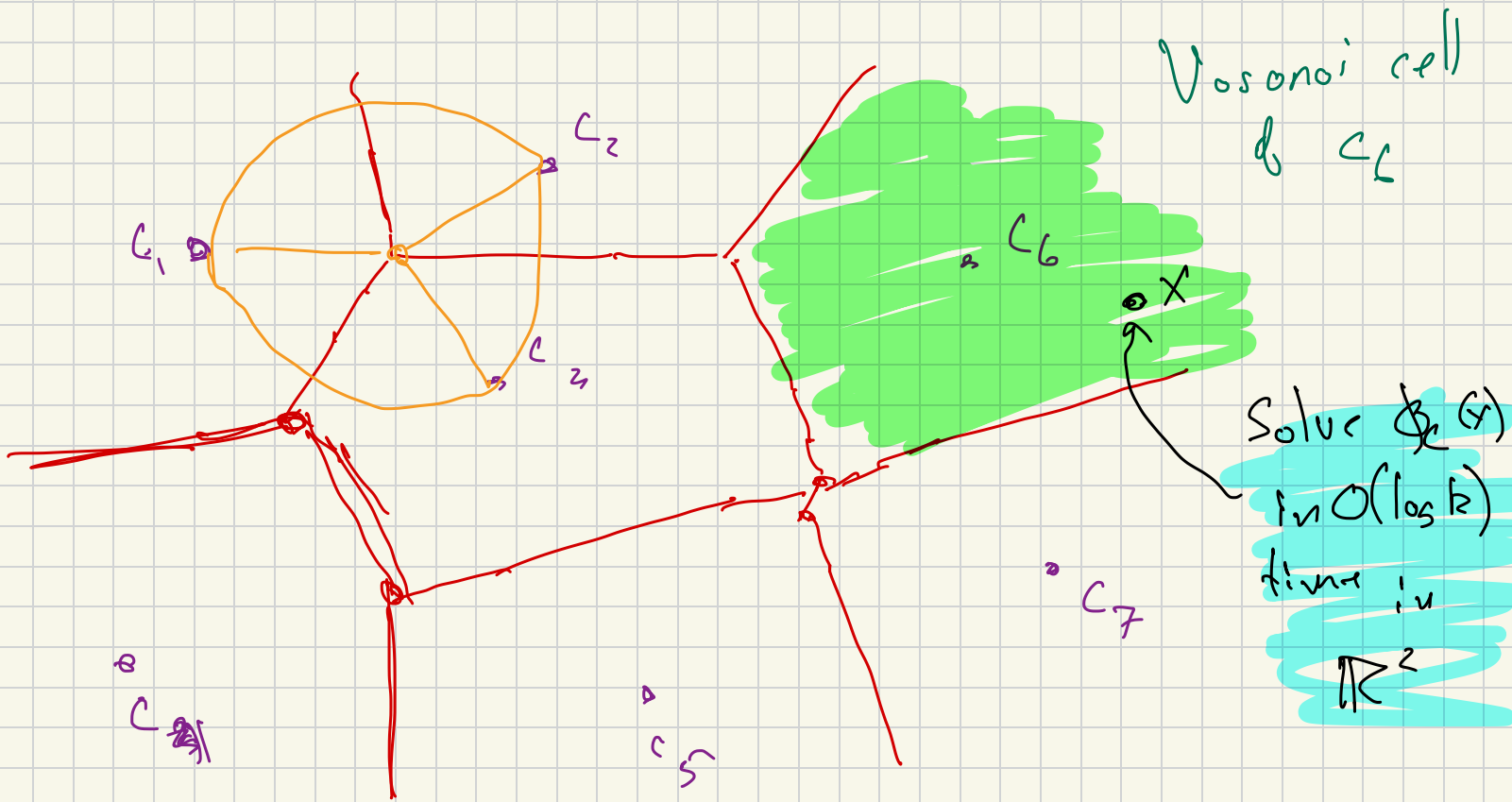
structure $\Phi_C(x) = \underset{c_j \in C}{\operatorname{argmin}} \|x - c_j\|$



$\{\Phi_C(x) = c_2\}$

Voronoi cell of c_2

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



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

size $\approx k^{\lceil d/2 \rceil}$

intractable

"Curse of Dimensionality"

Compute $\phi_C(x) = \underset{c_j \in C}{\operatorname{argmin}} \|x - c_j\|$

$c^* = c_1$ $j^* = 1$
for $j = 2$ to k
  if $(\|x - c_j\| < \|x - c_{j^*}\|)$
    $j^* = j$
return j^*

$O(k)$ time

Assignment-based clustering

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

- **k-means** — most common famous

$$C = \underset{|C|=k}{\operatorname{argmin}} \sum_{i=1}^{n} \| x_i - \phi_C(x_i) \|^2$$

→ soft clustering split decision

MoG

- **k-center** — Gonzalez Also

$$C = \underset{|C|=k}{\operatorname{argmin}} \max_{x_i \in X} \| x_i - \phi_C(x_i) \|$$

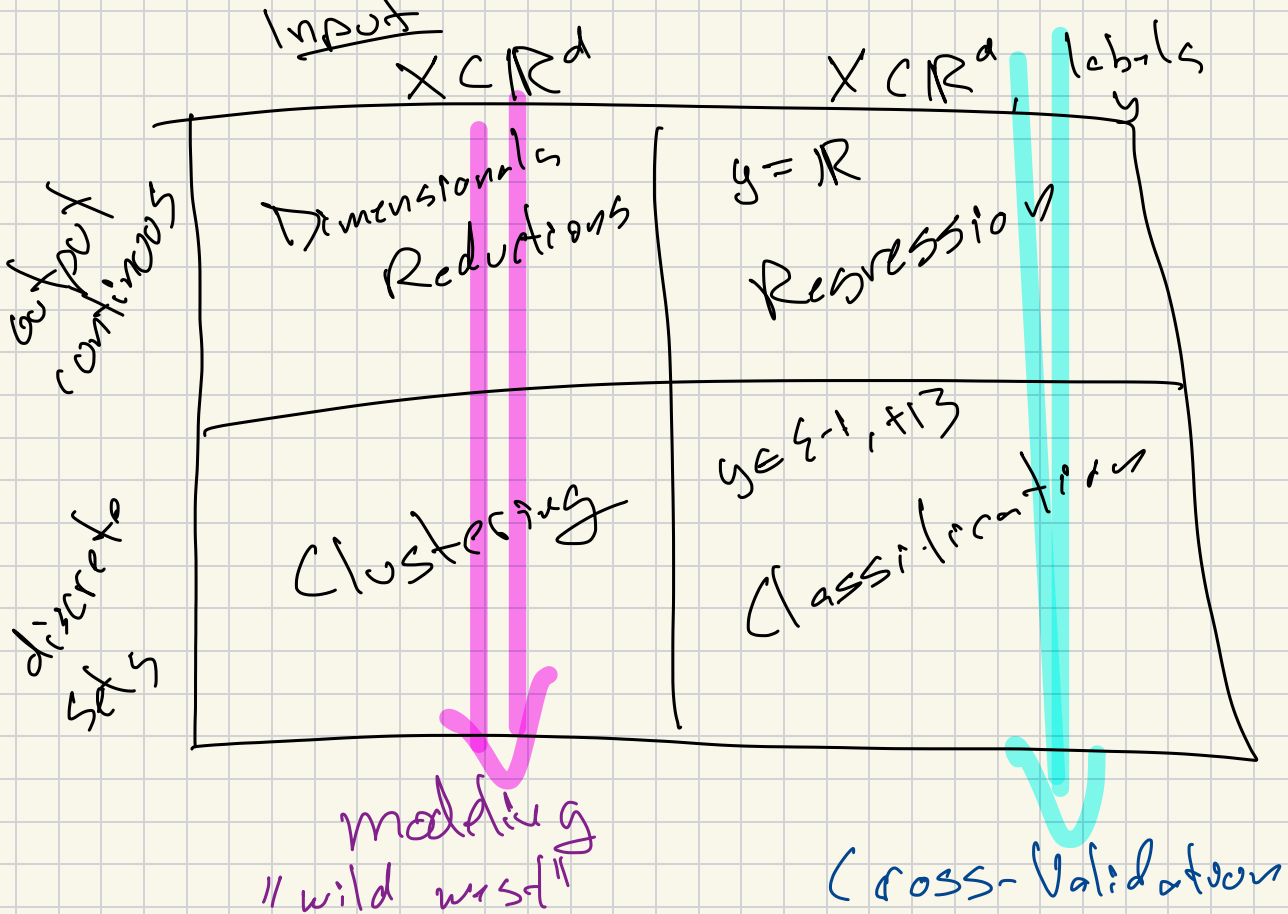
- **k-median**

$$C = \underset{|C|=k}{\operatorname{argmin}} \sum_{x_i \in X} \| x_i - \phi_C(x_i) \|$$

more resistant to outliers

- **k-medioid** — good alternative more interpretable

$$C = \underset{|C|=k}{\operatorname{argmin}} \sum_{x_i \in X} \| x_i - \phi_C(x_i) \| \qquad C \subset X$$


Input
$X \subset \mathbb{R}^d$
$X \subset \mathbb{R}^d$
labels
output continuous
discrete sets
Dimensionality Reductions
$y = \mathbb{R}$
Regression
$y \in \{-1, +1\}$
Clustering
Classifications
modeling
"wild west"
Cross-Validation