(2) → K-means Clustering
→ K-nearest neighbours

- Exploratory Data Analysis (EDA)

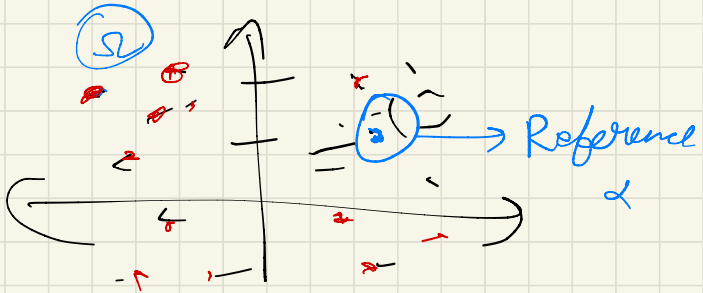
Virus $\xrightarrow{RNA}$ RT-PCR → Data

5000 samples

NYC, Pregnant women → covid testing

"Clustering"

Clusters → Groups



$(x_1, x_2, \ldots, x_n) \in \mathbb{R}^d \xrightarrow{\text{Partition}} P_1, P_2, P_3 \ldots, P_k$

$\hookrightarrow (S)$

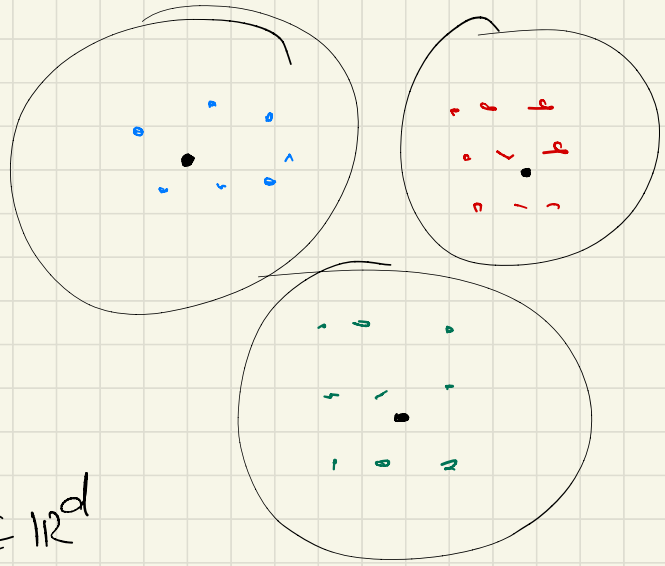
$P_1 \cup P_2 \cup \ldots \cup P_k = S$, $P_i \subseteq S \quad \forall i \in \{1, 2, \ldots k\}$

For any i, j s.t $i \neq j$, $P_i \cap P_j = \phi$

$\alpha, \beta, \gamma, \Omega,$ NA

Unsupervised.

• Pandas. "k" "MD"



k-means, k-medians, k-centers,
DBscan, k-medoids,

Objective fn: $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$

Goal: $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$

C_i = set of all points in dataset that closest to c_i

Mean-squared error

* $cost(X, C) = \sum_{i=1}^{n} (x_i - c(i))^2$, $c(i) = \min_{j=1}^{k} \|c_j - x_i\|$

- NP-hard to exactly minimize this cost.

$$\min_{C \in \mathbb{R}^d} \sum_{i=1}^{n} (x_i - C)^2$$

$$\hookrightarrow \sum_{i=1}^{n} (x_i - \mu + \mu - C)^2$$

n

μ

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

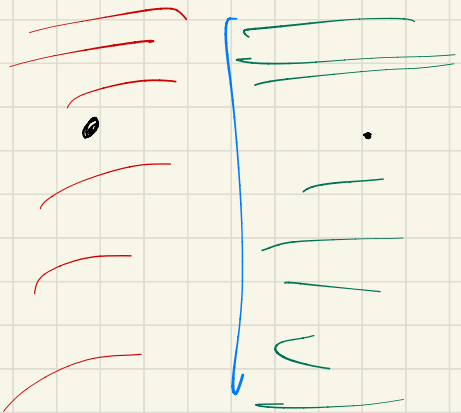
mean
centroid

$$= \sum_{i=1}^{n} (x_i - \mu)^2 + (\mu - C)^2 + 2(x_i - \mu)(\mu - C)$$

$$= \sum (x_i - \mu)^2 + \sum (\mu - C)^2 + 2(\mu - C) \sum_{i=1}^{n} (x_i - \mu) \quad \left(\sum_{i=1}^{n} (x_i - \mu) = 0\right)$$

- $\text{cost}(C) = \sum_{i=1}^{n} (x_i - \mu)^2 + (\mu - C)^2 n$

$$\text{cost}(c_1, c_2) = \sum_{i=1}^{n} (x_i - c(i))^2$$



- Lloyd's method / "k-means" algorithm

① Assign : For every data point, assign it to its closest center

② Re-center : "New centres" are the means of the clusters chosen in step ①.