① K-means clustering. ⟶ $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$

② K-nearest neighbors

③ neural network

↳ $X \to P_1, P_2, \ldots, P_k$

$P_1 \cup P_2 \ldots \cup P_k = X$

$P_i \cap P_j = \phi$

- Section (Initialization) ↳ Random

Iteration:

① Assignment

② Re-center

$$\left\{ \text{cost}(X, C) = \sum_{i=1}^{n} \| x_i - c(i) \|_2^2 \right\}$$

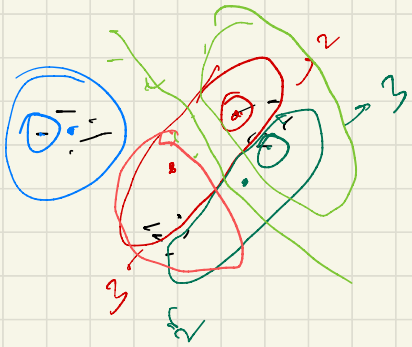
$|C| = k$

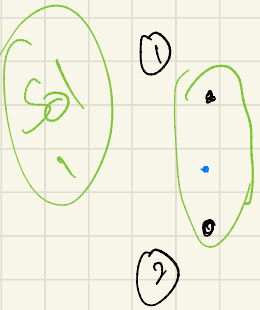
$c(i) \in C$

$$c(i) = \min_{c \in C} \| x - c \|_2$$

- Lloyd's algorithm

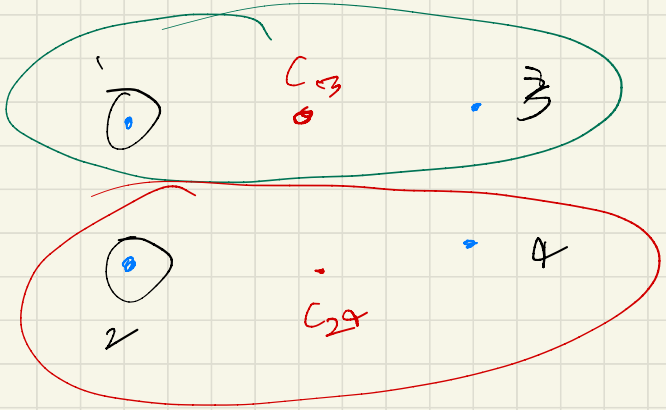
↳ K-means algorithm





$$\sum_{i=1}^{4} (x_i - C(i))^2$$

Random seeding $C_i = \{1, 2\}$



① Assignment

$1 \rightarrow 1$

$2 \rightarrow 2$

$3 \rightarrow 1$

$4 \rightarrow 2$

② Re-center

C_{13}

C_{24}

k-means ++ : (2008)

- Seeding
- k-means Algorithm

$C_1, C_2, \ldots C_k$, $C_i = \{c_1, c_2, \ldots, c_i\}$

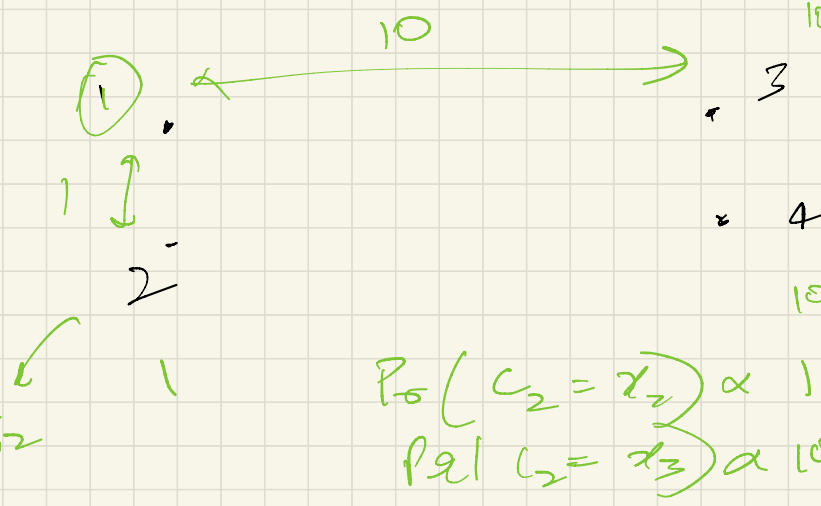
(1) First center is picked u.a.r

(2) For $i = 2$ to k :

$$Pr[C_i = x_j] \propto cost(x_j, C_{i-1})$$

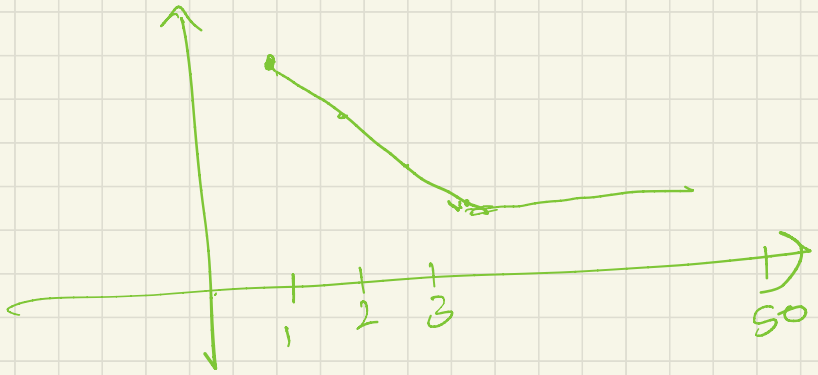
$$cost(x_j, C_{i-1}) = d^2(x_j, C_{i-1})$$

$O(nkd)$



How to choose k:

- Elbow method
  - k=1 to 50



k-nearest neighbours :

- Linear regression, neural networks, → parametric models.
  - ↳ $x \in \mathbb{R}^d$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$.

$$f(x) = w^T x + b$$

$$\min \sum_{i=1}^{n} |w^T x_i + b - y_i|^2$$

$$|Wx|^2$$

x_i, y_i $\quad$ Wx

- Non-parametric
  - Table look-up
  - k-nearest neighbours

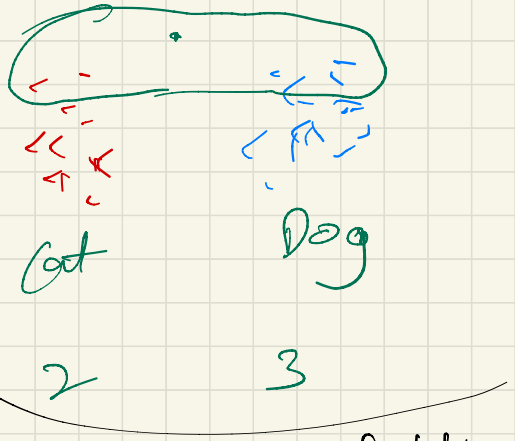
"Metric" → 1-nearest neighbour

k-nn:

① Find the 'k' nearest neighbours to the query point

② Output the mode of all the labels of the k-nn.

k → "odd"



Over-fitting

Under-fitting

(F)

Loss fn:

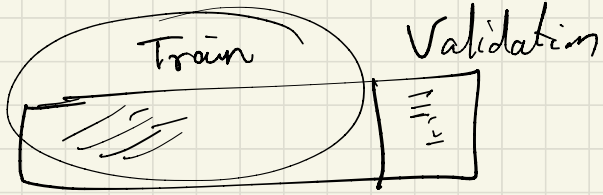
Train | Test.

0-1 error

How to pick "k":

Cross-validation

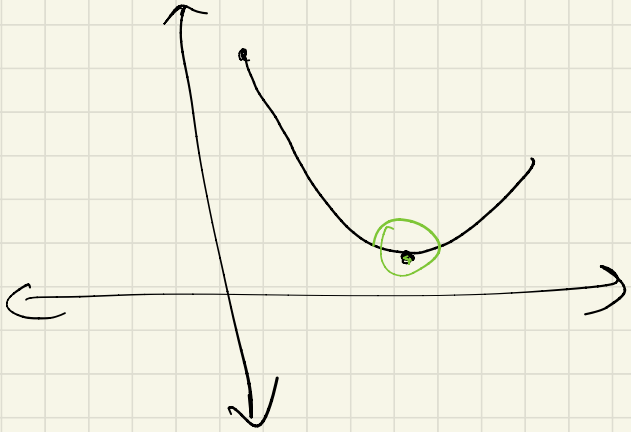
$k = 1, 3, 5, 7, 9, 11, \ldots, 99$

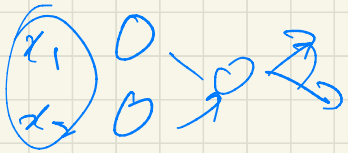


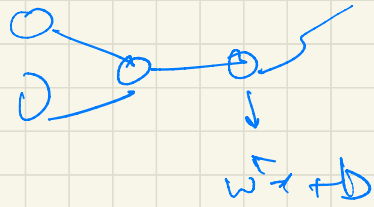
1000

$|V|$

$\sum_{k=1}$







3

x_1

x_2

w

1

2

3×100