K-nearest neighbors:

K-means ++ :

Clustering

Linear regression, neural network } → parametrized models → supervised learning

$y = w^T x + b$

$[x, 1]$

$w^T x$

• Non-parametric model:

Classification

Query:

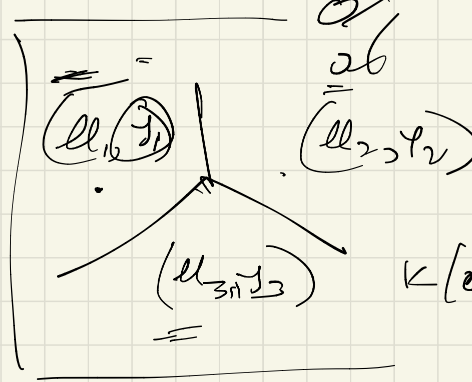
(1) Find the K-nearest neighbors to the data point

(2) Assign the mode of labels of the k-nn

0.0

0.2

0.6



$K(d+1)$

Time complexity: Turing machine

(1) Find k-nn.

(2) Assign mode

$\to O(k)$

[ ]

$\{1, 2, \dots, k\}$ $k+1$

$\{1, \dots, k+1\}$

$k \ll n$ $O(n \times [kd])$

n-datapoints

$x_1, x_2, \dots, x_n \in \mathbb{R}^d$

$\to$ k-numbers

$i = 1$ to k

$x \leftarrow x + x_i$

[, ,] + C

Query point $\to x \in \mathbb{R}^d$

$\sqrt{\sum_{i=1}^{d} (x_{ij} - x_{ij})^2}$

"C"

One-hot encoding

$[0, 0, 1, 0, 0]$

binary vector of size C

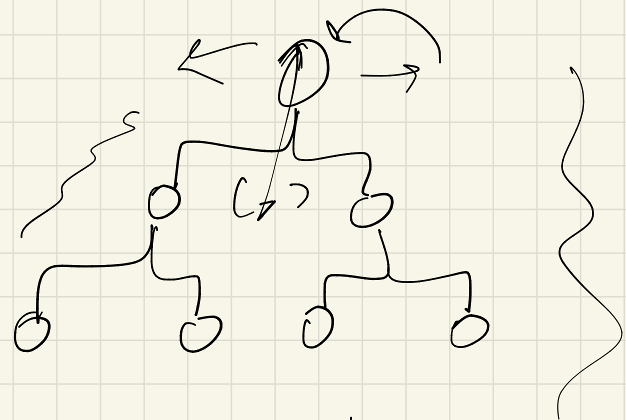
$\{1, 2, \dots, C\}$

Cat, Dog, Ship, Flower, Bicycle

Image 1k

Speed-up:

(1) k-d trees



$n \longrightarrow$ stored in k-d tree

log k

(2) Locality sensitive hashing

Time complexity

Space complexity $\rightarrow$ RAM

hash-code {

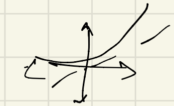
$\hookrightarrow$ function(File) $\longrightarrow$ very small string

$\hookrightarrow$. 128 bytes .

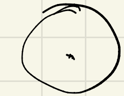
$Pr[f(x_1) = f(x_2)]$ astronomically low if $x_1 \neq x_2$

LSH-based data structure

$1 + x \le e^{x} \quad \forall x \in \mathbb{R}$



$F(x) \longrightarrow \mathbb{R}^{c}$



"Constant query time" } Probabilistic and it can fail.

99.999% → success probability

$1920 \times 1080 \to 10^{6}$

• concentration bounds / inequalities.

"Curse of dimensionality"

πr^{2}

$r(1-\varepsilon)$

$\hookrightarrow \pi r^{2}(1-\varepsilon)^{2}$

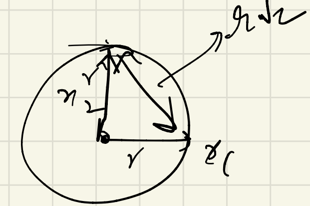
Most of the volume in high dimensions is at the boundary

$$\frac{Vol(A_{1-\varepsilon})}{Vol(A)} = (1-\varepsilon)^{d} \le e^{-\varepsilon d}$$

As $d \to \infty$

ratio $\to 0$

- Unit sphere in d-dim, x_i, x_j



$\langle x_i, x_j \rangle$ is very close to 0.

w.h.P

↳ with high Prob.

Law of large numbers

Central limit thm.

$x_i \to +1 \quad 1/2$

$x_i \to -1 \quad 1/2$

$x_1, x_2, \ldots, x_p$ i.i.d

$$X = \sqrt{p} \cdot \frac{x_1 + x_2 + \ldots + x_p}{p} \xrightarrow{d} N(0, 1)$$

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2]$$

$$\boxed{Var[\alpha X] = \alpha^2 Var X}$$

$$E[X] = E\left[\frac{x_1 + \ldots + x_p}{p}\right] = \frac{1}{p} \times p \times E[x_1] = 0$$

$$X = \frac{x_1 + \dots + x_p}{p},$$

$$x_i \begin{cases} \to +1 & \frac{1}{2} \\ \to -1 & \frac{1}{2} \end{cases}$$

$Var(x_1) = E[(x-\mu)^2] = E[x^2]$

$$Var[X] = Var\left[\frac{x_1 + \dots + x_p}{p}\right] = \frac{1}{p^2} \times p \times Var[x_1]$$

$$= \frac{Var(x_1)}{p} = \frac{1}{p}$$

Dist of $\sqrt{p}(X - \mu) \xrightarrow{d} N(0,1) \longrightarrow$ Central limit thm.

$$X = \frac{x_1 + \dots + x_p}{p}$$

converges in distribution

$$\frac{1}{\sqrt{x_1^2 + x_2^2 \dots + x_d^2}} (x_1, x_2, \dots, x_d) \longrightarrow \text{unit ball}$$

$x_i \in N(0,1)$