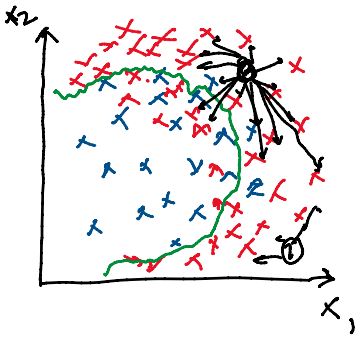
KNN

11 February 2022 20:56

Binary

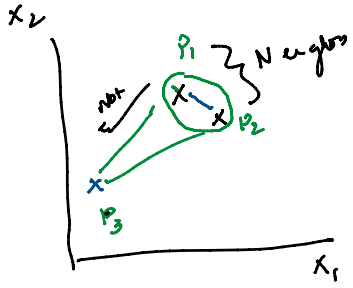
Classification

K- Nearest Neighbours



K- number ?

Nearest Neighbour



distance

(m) dimensional

Neighbours (metric)

(1) distance (Eucledia, Manhattan)

(2) Similarity (Consine)

Euclidian

$$d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + (p_3-q_3)^2 + \cdots} = \sqrt{\sum_{i=1}^{n} (p_i-q_i)^2}$$

$$(x,y) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

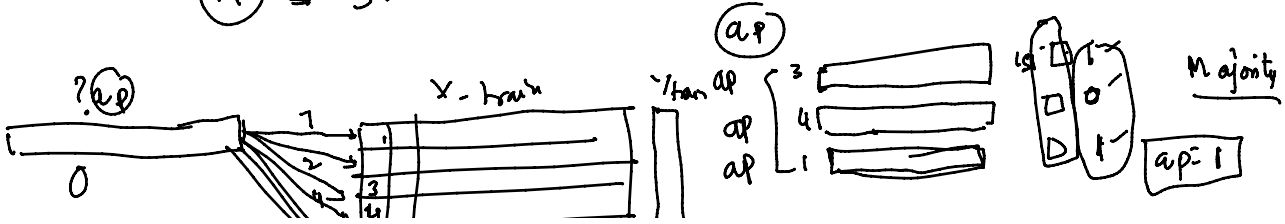
$$\text{Manhattan} = \sum_{i=1}^{n} |p_i-q_i|$$

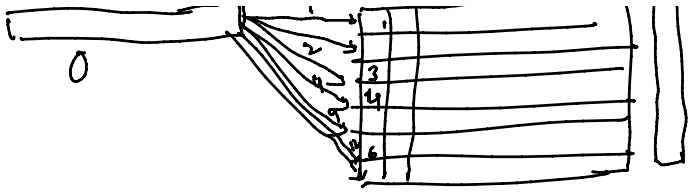
$$\text{Minkowski} = \left(\sum_{i=1}^{n} \left(|p_i-q_i| \right)^p \right)^{\frac{1}{p}} \qquad p=1$$

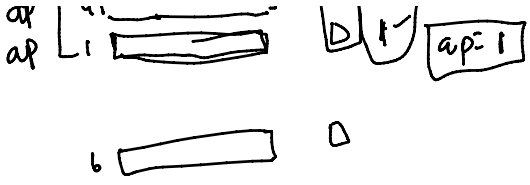
$p=1 \rightarrow$ Manhattan ✓

$p=2 \rightarrow$ Euclidian

(K) = 3 ✓







$\hat{Y} = 12$

Why ?

Training data. 065 = 1000 ✓ ← Var = 10 Try L.R data

065 = 1000 Var = 500 ✗ Try L.R data

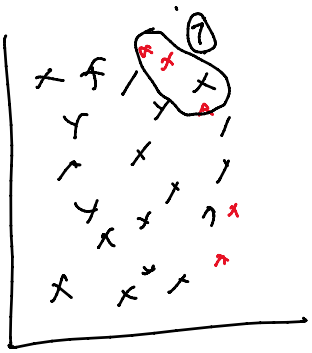
"dimensionality curse"

Gene

Biotech P 500 N 500 [1000] [1000 variables]

$n = 1000$, $m = 1000$ ✗ ML

Imbalanced data



distance based metrics :- *Normalize data ✓

	Sal	Yrs of exp
①	25000	3
②	30000	7

$$d = \sqrt{(30000 - 25000)^2 + (7-3)^2}$$

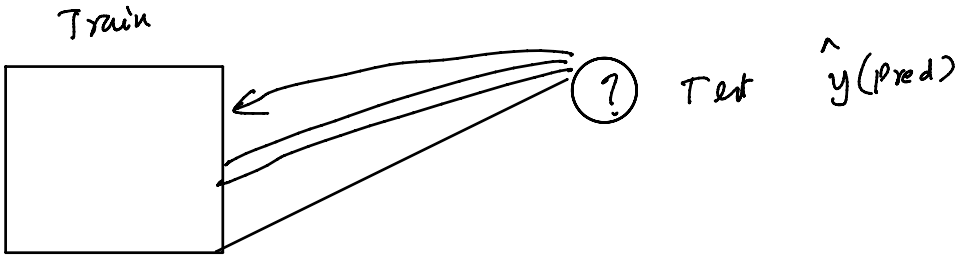
Numeric data (One hot encoding)

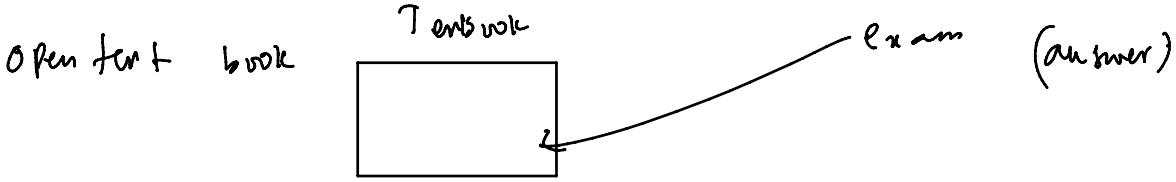
① Normalize data

② Distance (Euclidian, Manhattan) } ✓

③ K ?

Disadvantages :- lazy learning



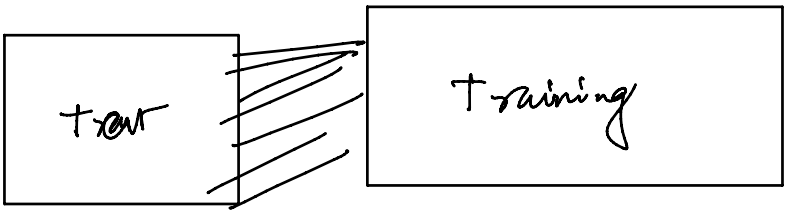


Complexity

Train O(1)

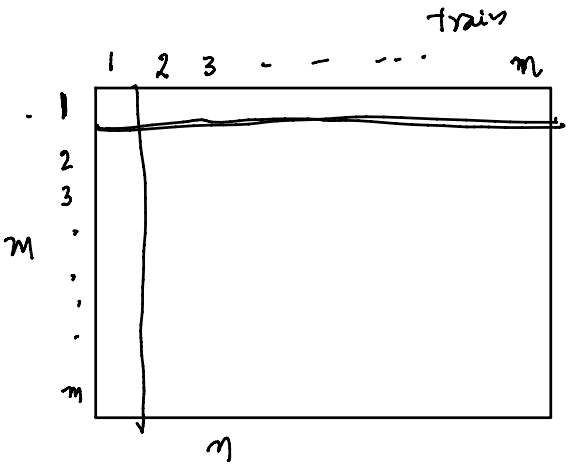
Calculation O(n)

Test(m) → O(n x m)

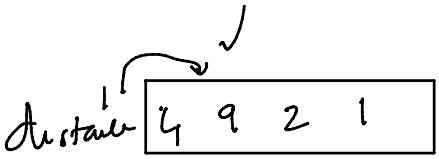


m obs

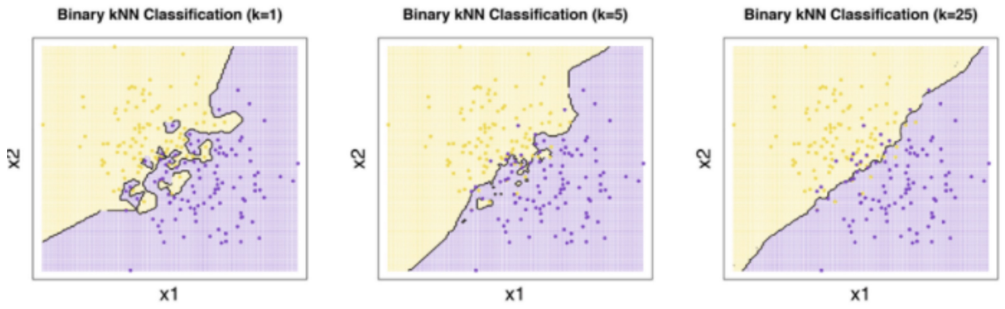
n obs

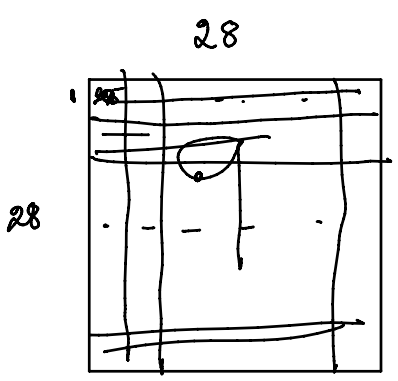


(distance)



(n x m)





Y

9

1x1 1x2 1x3 - - - 1x 28 2x1 2x2 - - - 2x28 - - - 28x1 - 28x2 - - 28x28

784

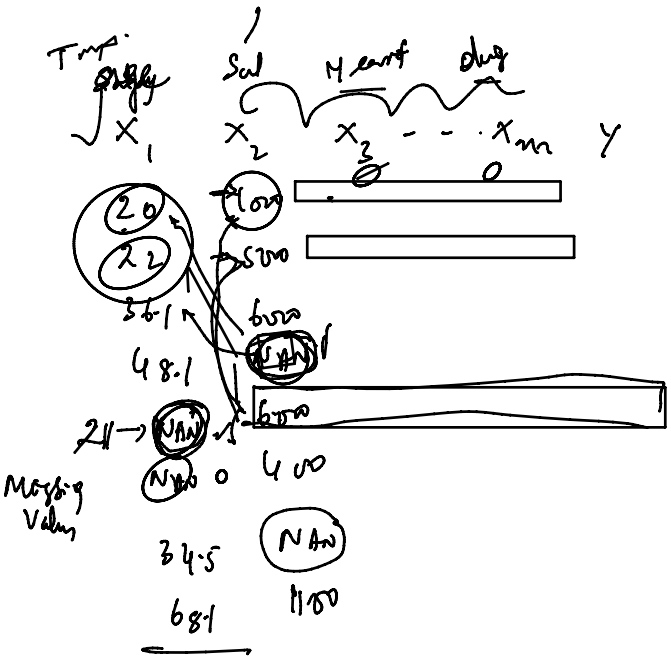
Multi class Problem

Yn ~ Sn - a?

Multi Class Problem

$Y \sim \{0 - 9\}$

$Y \sim 0, 1$



Imputation? {KNN Imputation}

Handling missing values

L.R $\theta_0 + \theta_1 (x_1)$ mistake

Replacing Missing - (Imputation)

Imputation → Mean Imputation

Median Imputation

Zero Imputation

Placeholder = (99)