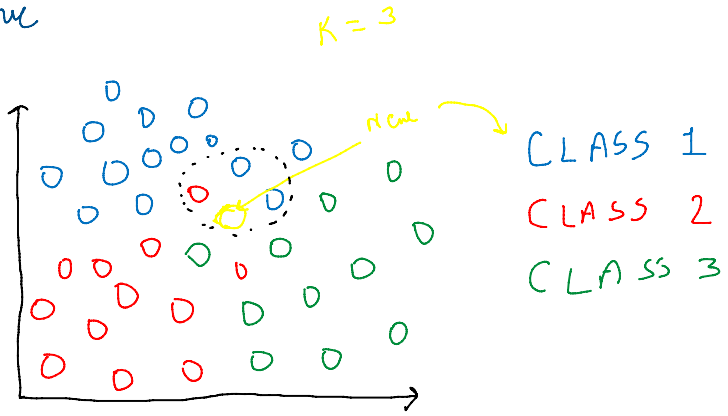
K Nearest Neighbors

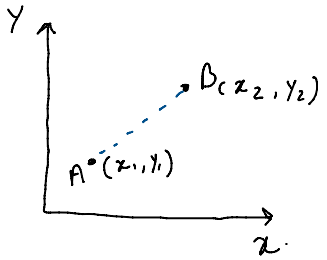
Sunday, June 12, 2022 9:36 AM

* Supervised learning Algorithm
* Non Parametric



> KNN algo is based on the principle that similar data points are close to each other.

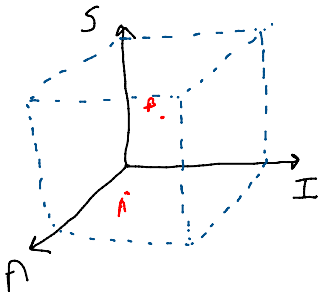
* Distance between two data points : (2D Plane)



$$d(A,B) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$ (Euclidean dist.)

Income Speed Age

$$d(A,B) = \sqrt{(S_A-S_B)^2 + (I_A-I_B)^2 + (A_A-A_B)^2}$$



I S A

	$\bar{I}$	$\bar{S}$	$\bar{A}$
A	I_A	S_A	A_A
B	I_B	S_B	A_B

* DISTANCE METRICS :-

1) MINKOWSKI DISTANCE :

$$d(A,B) = \left(\sum_{i}^{n} |A_i - B_i|^p \right)^{1/p}$$

Ex $A(A_1, A_2, A_3)$ $B(B_1, B_2, B_3)$

$p = 2$

$$d(A,B) = \left[(A_1 - B_1)^2 + (A_2 - B_2)^2 + (A_3 - B_3)^2 \right]^{1/2}$$

2) EUCLIDEAN DISTANCE :

$$d(A,B) = \sqrt{\Sigma (A_i - B_i)^2}$$

> Minkowski with $p = 2$

(iii) MANHATTEN DISTANCE :

$$d(A,B) = \Sigma |A_i - B_i|$$

> Minkowski with $p = 1$

* KNN ALGORITHM :-

> Select a value of K
> Calculate the distance of unknown data points from all other data points.
> Select K data points which are closest to the unknown data point.
> Predict the class of unknown data point from the most popular response from K nearest neighbors.

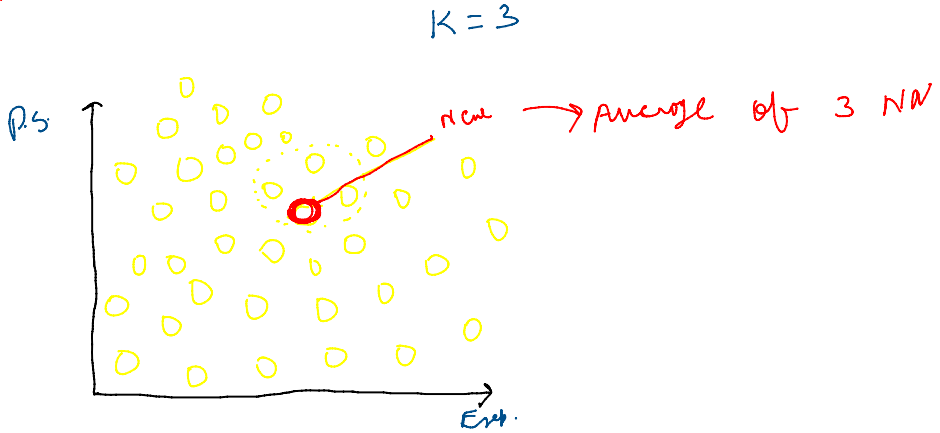
* EFFECT OF VALUE OF K :-

> CASE 1 : K is very small
  > Overfitting
  > Complex and inappropriate model

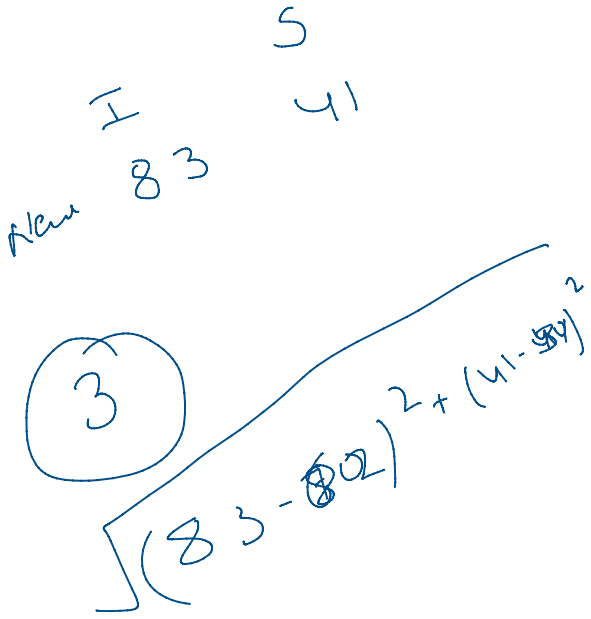
> CASE 2 : K is very large
  > Underfitting
  > Overly Generalized model

* Test the model for multiple values of K and choose the one which gives best result.

* KNN REGRESSION :-



> Select a value of K
> Calculate the distance of unknown data points from all other data points.
> Select K data points which are closest to the unknown data point.
> Predict the output of unknown data point by taking average of output of KNN.



I	S	D
70	40	0
80	40	0
60	30	1
82	41	1
50	25	0
60	45	0
85	45	1