Unit-III (Part-2).

Ensemble learning:-

* Ensemble learning Combine individual model together to improve the stability & Predictive Power of model.

* It Combines multiple machine learning model in to one Predictive model

* The Combined Strength of the models offsets, individual model, Variances & biases.

machine-1 — which is giving 50% of required o/p

machine-2 — " 50% " in "

(+)

Will make Some improvement. — 70%. We can get by together.

Two approaches are Bagging & Boosting.

Bagging:- [Parallely we do].

* Bagging is a homogeneous weak learners from each other independently in parallel & Combines them for determining the model Average.

* It is used to decrease the variance in the prediction model.

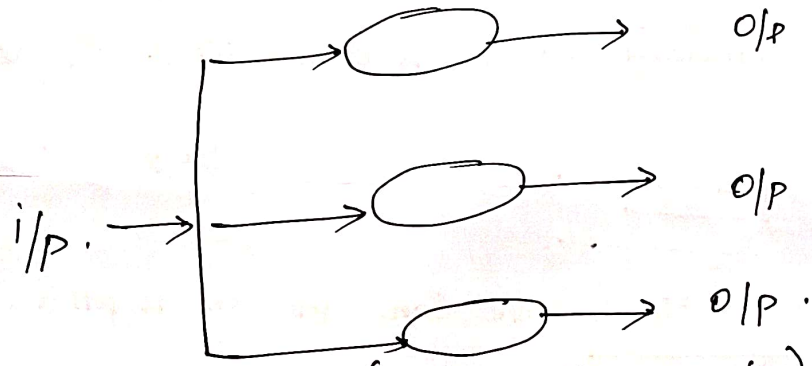
* Bagging is a parallel method that fits different

Considered learner, independently for each, making it possible to train them simultaneously.

→ Bagging generates additional data for training from Data set.

→ This is achieved by Random Sampling with Replacement from the Original dataset.

→ Sampling with Replacement may repeat some observation in each new training data set.



(Parallel i/p get o/p).

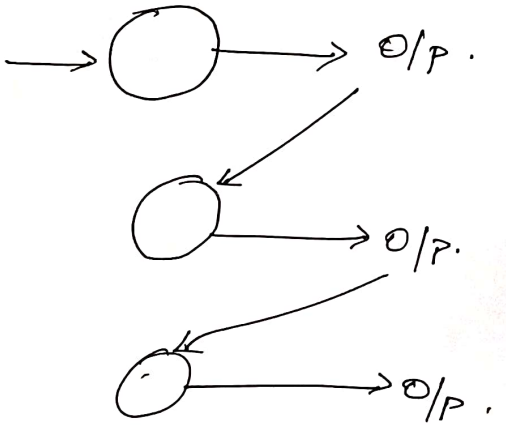
— x —

Boosting :- [Sequential].

* Boosting is a Sequential ensemble method that iteratively adjust the weight of observation as per the last classification

* If an observation is incorrectly classified, it increases the weight of that observation.

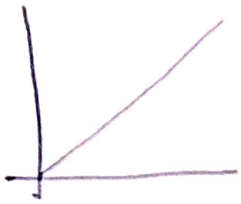
* The team Boosting in a layman language refers to algorithms that convert a weak learner to a Stronger One. It decrease the bias Error and builds Strong Predictive models.



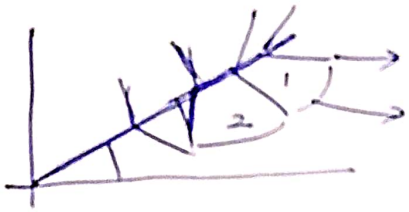
Boosting Sequential.

— x —.

Ex. Initially, We have $f(y) = 2x+5$ (Linear)



If we have $f(y) = 2x^2+3x^2$ (Quadratic eqn.) and divides into Pocket's.



// 1st fig in Linear, 2nd fig not linear, It is Quadratic, But it is trying to make Linear, why means, Based on the Previous (Least) / memorize it is trying to implement Linear with Pockets. //

— x —

KNN Algorithm:- [K-nearest neighbour Algorithm].

Ex:- for Lazy learning (It wont learn from training data, it just memorize the training data.

Given Data Query ⇒ x = (Maths = 6, CS = 8)

and K = 3 (Constant which is gives nearest neighbours)

Classification - Pass / fail.

Data related to 5 Students Consider.

	Maths	CS	Result.	
1	4	3	F	
2	6	7	(P)	⇒ Existing data.
3	7	8	(P)	
4	5	5	F	
5	8	8	(P)	

How do you Prob means Using Euclidean distance (d)

Euclidean Distance (d)

$$d = \sqrt{|x_{o_1} - x_{A_1}|^2 + |x_{o_2} - x_{A_2}|^2}$$

O → Observed Value

A → Actual Value

1 - Math

2 - CS.

i) Calculate d_1

$$d_1 = \sqrt{(6-4)^2 + (8-3)^2} = \sqrt{2^2+5^2} = \sqrt{29} = 5.38$$

Stu 2) ✓ $d_2 = \sqrt{(6-6)^2 + (8-7)^2} = \sqrt{0+1} = \sqrt{1} = 1$

✓ $d_3 = \sqrt{(6-7)^2 + (8-8)^2} = \sqrt{1+0} = 1$

$d_4 = \sqrt{(6-5)^2 + (8-5)^2} = \sqrt{1+9} = \sqrt{10} = 3.16$

✓ $d_5 = \sqrt{(6-8)^2 + (8-8)^2} = \sqrt{4+0} = \sqrt{4} = 2$

Next Among these 5 values have to choose 3 values Because value of K = 3. Should Choose neighbour values.

Choosing 1, 1, 2 (neighbhour values).

In Existing data Tabulation, Choose 2, 3, 5 (stu) III-1 (40)

So we will get 3P, OF. ⇒ (3P)/OF

Obviously pass is dominating due to 3pers. Which ever is majority have to choose that.

Ex (2P)/1F , 1P/(2F)

—X—

Regression :-

Statistical Tool used to understand and Quantify the Relation between 2/more variables.

// understand and express some Relation between 2-Variables then the tool is called Regression//.

Linear Regression :-

$$y = \beta_0 + \beta_1 x + E.$$

→ y depends on x, y is dependent variable, x is independent variable.

→ If any Changes in x reflect same in Other's (i.e) y.

y = Dependent Variable

x = Independent Variable

β_0 = Constant / Intercept

β_1 = x - Slope / Co-efficient

E = Error

*It will work good in Linearly Seperable Data Only.

(+)(+) (+) | (-) (-) (-) → with Single Straight line will divide

Locally Weighted Regression:-

- To Overcome the Problem of non-linearly Seperable Data.
- LWR Alg assigns weights to data to Overcome the prob
- Computationally more expensive (Each & every data Assigning weights).

(*) Assings weights with the help of Kernel Smoothing.

If it close to the predicting value it will Assign the weight.

Kernel Smoothing:-

$$D = ae^{-\frac{||x - x_0||}{2c^2}}$$

(Dot product)

C is Constant

x = Each training i/p

X_0 = value we are predicting

* x - If 10, 10 data set, for every data allocating weight on it.

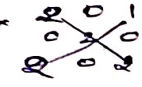
* x_0 Predicting

* for Each x value we need to allocate weight matrix(w)

How weight matrix (w) calculating?

If x = 10, 11 - weight matrix assiging.

10 - for x, Remaining 1 for Predict (x_0).

* Weight matrix Called Diagonal matrix - Ex  values.

$$B = (x^{T}Wx)^{-1}x^{-1}Wy$$

(T - Trampose)

Unsupervised Learning:-

Data { Label / Unlabelled.

* In unsupervised Learning using unlabelled data. (Human teaches to computer to behave in this fashion).
* In Supervised learning i/p data is labelled. (vice versa).

[The System knows for this i/p, This o/p will come].

Color i/p → o/p (Blue/Green].

* In unsupervised learning, Unlabelled, there is No Prior knowledge of the data i/p. & No knowledge about o/p.
* What will do in Unsupervised ?

The o/p depends on Algorithm, Alg means Not set of Explicitly pgm Inst. Alg means the model we are designing for making i/p in to o/p.

Here The o/p depends on Alg designing.

* Here there is No training data (i.e) Not giving any Training to the Computer.

Infact the "Computer teaches you" for this i/p This is the o/p. Based on intresting pattens. the Computer design the i/p. and o/p will Produce

(Ex) Search Engines. [How google works. Whenever giving Query in Search Engine, what is to be

* Which class is to be produced.

So I donot have any info about Query, and what is to be produre also don't know. Because, Whenever Query is given. Different No: of Webpages / Links to be displayed.

* But in Supervised learning, Whenever I am giving the Question, Compulsory it will move to particular class.

Whenever giving Car it moves to Blue

Color " " green / Red.

i/p Exactly moves to one particular class.

* But in Unsupervised learning, " " . . "

" , What answer we don't know.

* Diff. No. of Links displayed on the Screen. Among this user will select one choice.

Search Engine

↓

many web / link display

* Ex Add pgm in Search Engine → o/p displays → Display Same pgm (Website) Related Addition.

* Supervised learning (2 TYPES) Consist " Classification & Regression "

Based on data, Real value / Categorized data.

* III'y in unsupervised learning

1) Clustering: Group of data. (Ex) market Analysis.

We have many Number of Customers, the customers are grouped Based on purchasing behaviour. (Because) The customers what purchasing we don't know. Snacks, Electronics, Group wise spritted.

2) Association Rule Mining:

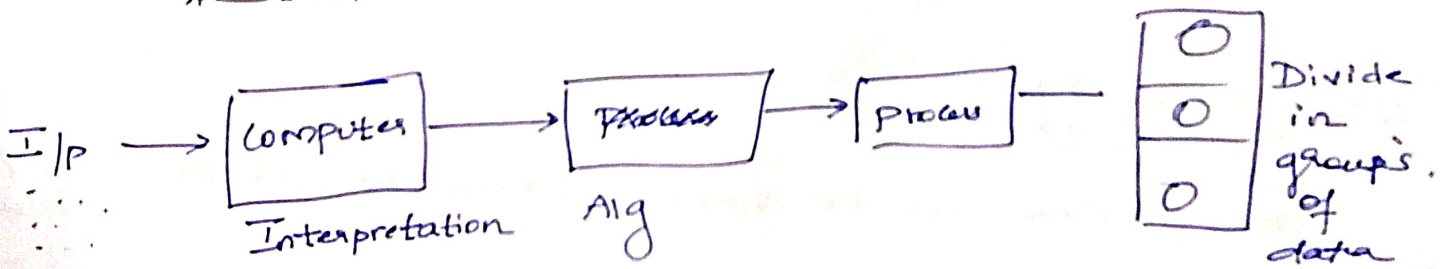
* It discover some Rule, such that

if a person Buy 'x' item, He tends to Buy 'y' item also.

* The computer trained by itself. By Checking 10 Customes, Buy "Laptop" tends to by "Keyboard".

* Washing machine
Related together
* Covers
* Stand

} Behaviour of Previous Customes, Computer Trained.



This is the Procau of unsupervised

——x——.

K- Means Algorithm :-

Partitioning Methods :-

n data items / objects $\xrightarrow{\text{Divided in to}}$ K Partition / Part's / Group's / Cluster.

* Each partition will represent one cluster
* $K \leq n$, The size of K should be less than / equal to n

n - Total data items

K - Each individual.

2 Rules :-

1) Each partition - atleast one object
2) Each object should belong to only 1 partition.

Ex. K-means Alg :-

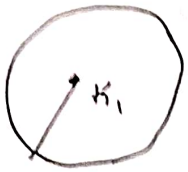
Data divided in to cluster based on distance and values.

	Height (x)	Weight (y)
1)	185	72
2)	170	56
3)	168	60
4	179	68
5)	182	72

	Height (x)	Weight (y)
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	180	76

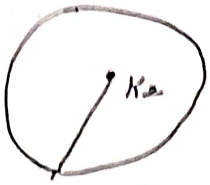
* Our main moto, divide data into Clusters.

* Each Cluster have Centroid.



(185, 72)

This order may take any value from data set.



(170, 56)

* Now Remaining values should be divide Based on Euclidian distance (ED)

$$ED = \sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$

x_0 → height, x_c → Centroid values, y_0 → weight, y_c → Centroid values

for (3)

(168, 60) → $K_1 = \sqrt{(168-185)^2 + (60-72)^2} = 20.8$

$K_2 = \sqrt{(168-170)^2 + (60-56)^2} = 4.48$

Which is smaller among these 2 values.

So ∴ $K_2 < K_1$ $K_2 \in K_2$.

* Assign the data point K_2.

* So Third row Belongs Cluster K_2.

* Now Calculate new Centroid for K_2.

$$\left(\frac{170+168}{2}, \frac{56+60}{2}\right) = (169, 58)$$

(170, 56) & (168, 60)

*

$K_2 = (169, 58)$ Updated.

---

for (4) (179, 68)

$$K_1 = \sqrt{(179-185)^2+(68-72)^2} = 6.32$$

$$K_2 = \sqrt{(179-169)^2+(68-58)^2} = 14.14$$

$K_1 < K_2 \therefore$ (4) εK_1

↓

Belongs.

New Centroid for K_1

$$= \left(\frac{179+185}{2}, \frac{68+72}{2}\right)$$

Old (185, 72) New data point (179, 68)

$K_1 = (182, 70)$

Like wise (5), (6), (7) ..... etc.

Rows.

$K_1 = \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ (Belong K_1)

$K_2 = \{2, 3\}$ (Belong K_2).

*Now the data divided in to 2 different cluster

— x —