M T W T F S S

Page No.:

Date:

- There are 3 types of Learning they are :-
  Supervised Learning
  Unsupervised Learning
  Reinforcement Learning

1. SUPERVISED LEARNING

- The Learning Agent (LA) learns from the enviornment as well as his/her previous output. The supervisor and LA both links to the enviornment.
- The supervisor is having the ideal solution.
- The LA receives the feedback from the supervisor as well as enviornment.
- ⊕ - It is called as an adder it provides the differentce between Actual and predicted.

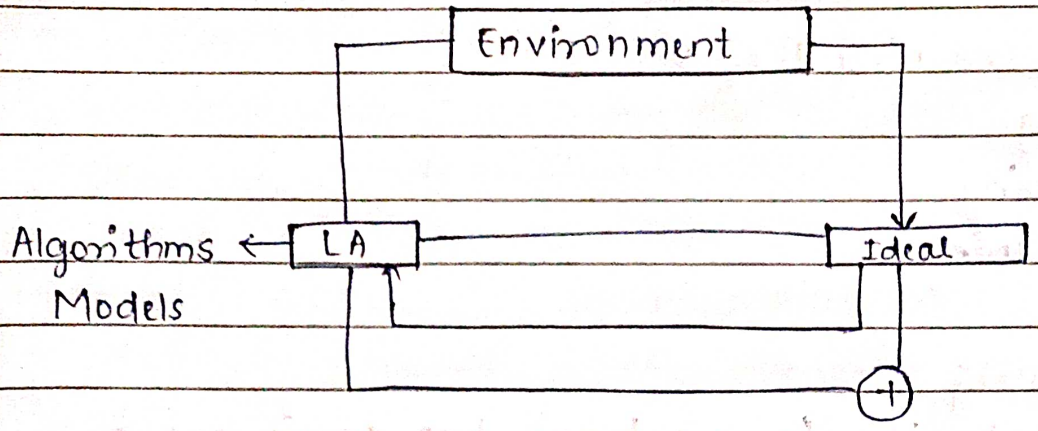
In this way the machine trains / learns and predict the outputs.

- There are different terminology available
  feature → attribute
  predictor of variable or independent variable.
- Target variable - Dependent variable / Reponse variable.

SUPERVISED LEARNING

Regression	Classification
- Linear Regression	- Support Vector Machine
- MLR	
- Lasso Regression	- Decision Tree
- Ridge Regression	- KNN - K Nearest Neighbour.
- polynomial Regression	- LR

Date:



Adder = Actual - Predicted

2. UNSUPERVISED LEARNING

- In unsupervised Learning the model has to learn the unlabelled data.
- The LA learns the input from the enviornment. It has the actual output where he tries to learn from own.
- LA learns by himself not from any supervisor everytime the LA tries to improve the accuracy by comparing his previous responses.
- It learns its own pattern (→ Predicted) again and again and it generate the o/p without the corresponding output (labelled data).

Examples of unsupervised learning are clustering, Association rules etc.

- Why we use unsupervised Learning?
  The model extract the hidden features from the dataset. It provides the insightful information about a data set.
  For example - Hidden pattern

Algorithms — K-mean, K-medoid, DBSCAN, HC → Clustering
Aprior, FP-Growth, CLAT → Association

Labelled data - Model is trained

x - study time
y - Grade.

x	y
5	D
6	C
7	B

M T W T F S S
Page No.:
Date:

Dataset

Association

T_1 : Bread, cheese, butter
T_2 : chicken pakoda, Beer
T_3 : Book, Pencil

Bread → cheese
(Bread, Butter) → Cheese

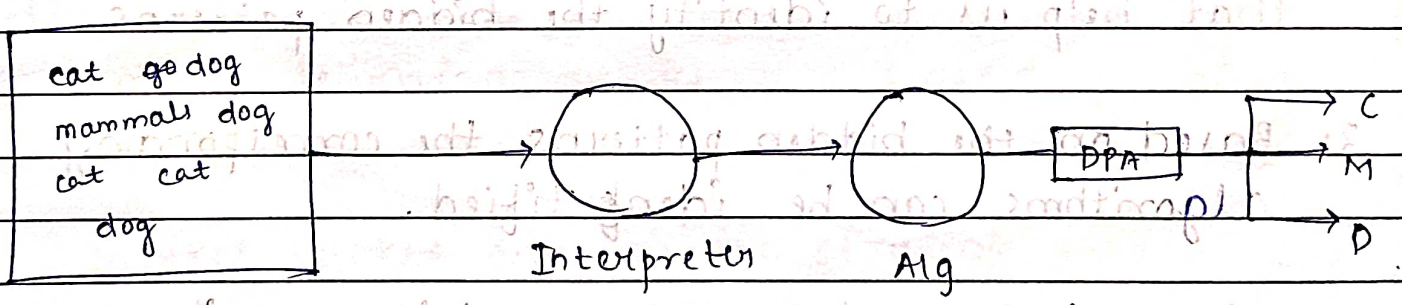
MBA - Market Basket Analysis.

That hidden patterns provide to the model. The model learn and predict it.

- Models of unsupervised learning

Let us consider a model which contains huge number of models of and our task is to catagorise the different objects.

1. Raw data
2. Interpretation - (finding hidden data from the data set)
3. Algorithm
4. DPA (Data pre-processing ~~Algorithm~~ Analysis)
5. Result.



Raw data

Features { C → 1 2 3 4 ....
M → 11 22 33 44 ....
H → 111 222 333 444 .... }

Outline of DPA (For cleaning the data)

(i) Removing Noise
(ii) Handling Missing (unknown value, unassigned value)
(iii) Identify Outlier.

1. Initial we will collect a data from different sources and prepare the raw data.
The rawdata passed into the model

2. In this step, the model generates the features which are available in the raw dataset. It finds the hidden patterns and establish a strong association rules.

Ex-

Features { C → 1, 1, 2, 3, 4
M → 11, 22, 33, 44
H → 111. 222 333 444

Transactional

R_1 : Bread → cheese
R_2 : Pen → Book

The R_1 and R_2 are called as strong association rules that help us to identity the hidden patterns.

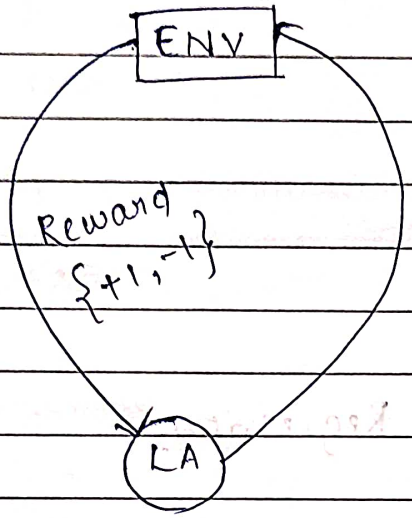
3: Based on the hidden patterns the corresponding algorithms can be identified.

4. DPA helps us to remove the Noise and identity the outlier from the dataset. It will clean the dataset. This process is called as data cleaning.

5. Finally, the algorithm divides the data objects into groups according to the similarities and find the difference b/w them.

Date:

3. REINFORCEMENT LEARNING
(Reward based Learning)



(i) +ve
(ii) -ve

In Reinforcement learning the learning agent learns from the environment and provide a series of actions each actions are called as the decissive statements. It will be provide the Rewards each reward is further classified in two types *positive and Negative. The +ve reward is treated as +1 and the -ve reward is treated as -1. The model learns all the +ve instances and build a model. It enforce to learn hence it is called as Reinforcement learning or Reward based learning.

MBA - Market basket analysis is a strategic data mining technique used by retailers to enhance sales by gaining a deeper understanding of customer purchasing patterns.

SUPERVISED LEARNING

It provides the different algorithms to predict the future value and these are

- Linear Regression
- Multiple
- polynomial
- Ridge
- Lasso
- Logistic Regression
- Support Vector Machine Regression
- Decision Tree Regression

In all cases the machine trains with the training data and validate with the testing data.

1. LINEAR REGRESSION

The term linear Regression means how one variable is linearly dependent into another variable. The Objective of LR is to make predictions like cells prediction, Salary prediction, Age prediction, price of product prediction.

LR algorithm shows a linear relationship b/w a dependent variable(Y) and one or more independent variable and Hence it is called as LRs it means finds how the value variable according to the value of independent variable

Let us take an example —

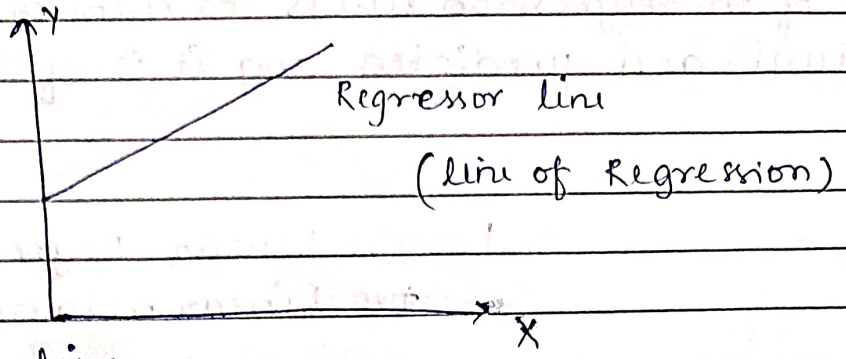
x ← Number of hours	y → grades
5	D
6	C
7	B
8	A

M T W T F S S

Page No.:

Date:

Let us plot the graph and find the regressor line



The Regressor line

$y = mx + c$

OR

$y = b_0 + b_1 x_1$

The objective of the LR is to find the best fit line from equation where x is independent y =

$b = 0$ intercept of the line

b_1 = coefficients of the line.

Types of linear regression

(i) Linear

If a singal independent variable is nearly equal to the value of a numerical depended variable

(ii) Multiple Linear Regression

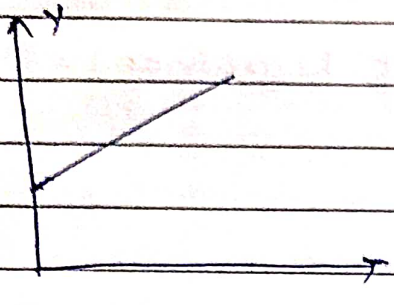
The equation is $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$

when more than 1 independent variable is used to predict the value of y then it is called MLR.

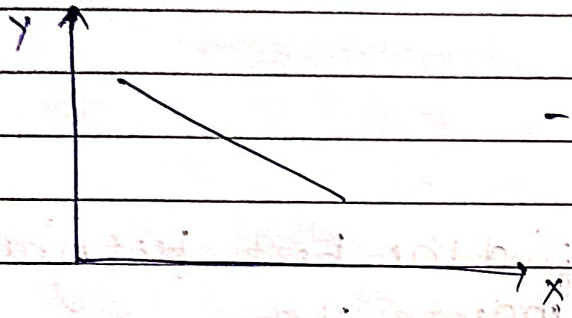
x_1	x_2	y

Linear Regression lines

The Objective of the regression line is to minimize the error between actual and predicted and it is of two types



Positive Linear Regression
+ve Linear relation.



-ve linear relation

- COST FUNCTION

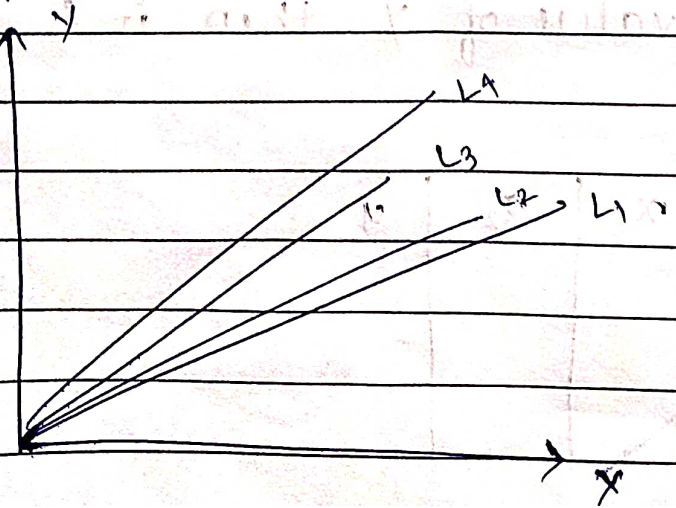
To find the best fit line we required to find the difference blw the predicted value and the actual value the error is called as loss.

Error can be defined as Actual - Predicted.

Loss → $E = A - P$

The different values for weight or coefficient of lines b_0 and b_1 gives the diff and the cost function is used to estimate the values of the coefficient for the best fit line.

In this case there are 4 lines that can be obtained and can be represented in 2 dimensional coordinate.



Out of these 4 line which one is the best line?

Those lines which obtain the less error that is the best line that means it is the less loss between actual and predicted.

	I/N		Act	Pred	D
L_1 :	5	E	90	85	5%
L_2 :	6	E	90	86	4%
L_3 :	7	E	90	88	2%
L_4 :	8	E	90	89	1%

Loss

Avg = 3%

$$\text{Mean Square Error} = \frac{1}{N}\sum_{i=1}^{N} y_i - (b_0 + b_1 x_i)^2$$

LINEAR REGRESSION.

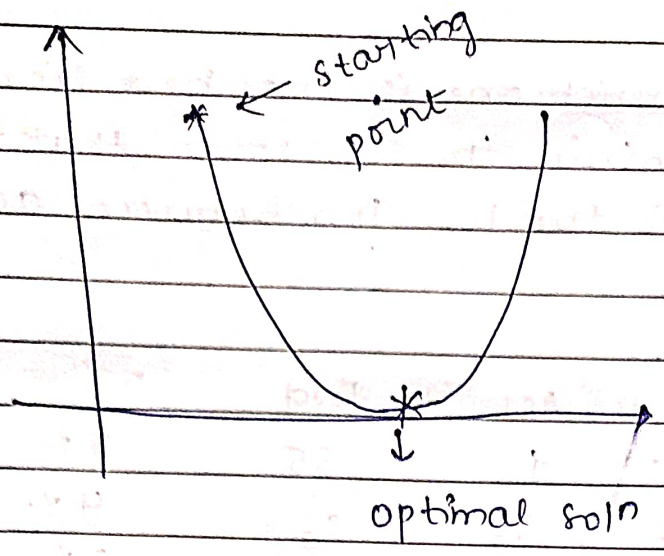
Through Linear Regression we identity the regression line by using the least Square Method.

Let us consider the 4 data points if we plot then the 4 lines can be obtained by using the linear regression equation

but our objective is to find the best regression line or the optimal line

The achieve the optimization, gradiention algorithm provides the idel soln to minimize MSE (Mean square Error). Gradientation is used to minimize the MSE by calculating the gradient of the cost function it will update the co-efficient of the line by reducing the cost function.

A gradientation can be represented as:

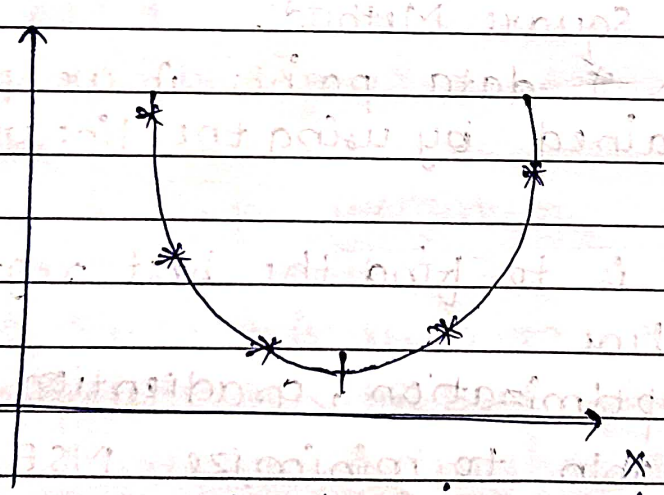


There are 2 different scenario can be possible

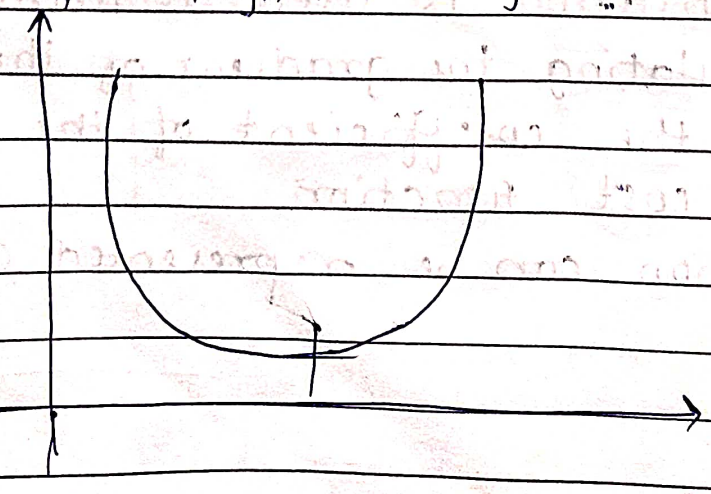
(i) If you choose to take longer steps then u may get sooner but there is a changce that you can overshoot the bottom of the pit.

(ii) If you take lesser steps then u will reach the optimal soln. But it will take much time each step can be idely represented as

$\alpha = 0.0001$ which is called as learning rate.



High learning rate.



Each time the value of x is updated and checks with the optimal solution and this way gradientation works.

Let us consider a dataset which consist of 2 features x and y and find the followings

(i) Find the b_0

(ii) Find b_1

(iii) Generate the Regressor line.

(iv) Find the Error term

(v) Find R^2.

x	y
1	2
2	4
3	5
4	4
5	5

Apply the linear regression algorithm for the following

The Formula

$$y = b_0 + b_1 x \longrightarrow (1)$$

Calculate the b_1 from the formula.

$$b_1 = \frac{(x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2}$$

$\bar{x} = 3$

$\bar{y} = 4.$

~~Calculat b_0 from the formula.~~

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	$\hat{y}$
1	2	-2	-2	4	4	4	2.8
2	4	-1	0	1	0	-0	3.4
3	5	0	1	0	1	0	4
4	4	1	0	1	0	0	4.6
5	5	2	1	4	1	2	5.2

$b_1 = \frac{6}{10}$

$b_1 = 0.6$

$\hat{y}-y$	$(\hat{y}-y)^2$
0.8	0.64
-0.6	0.36
-1	1
0.6	0.36
0.2	0.04

$\Sigma(\hat{y}-y)^2 = 2.4$

Calculate b_0 from equation ①

$$y = b_0 + b_1 x$$

$$y = b_0 + 0.6(3)$$

$$y = b_0 + 1.8$$

$$4 - 1.8 = b_0$$

$$2.2 = b_0$$

$$Y = b_0 + b_1 x$$

$$\boxed{Y = 2.2 + 0.6x}$$

Regressor line
Best fitted line
optimal line for the prediction.

Calculate $\hat{y}$ by substituting the values of x

$$\hat{y} = 2.2 + 0.6(1)$$

$$Y = 2.8$$

Substract y from $\hat{y}$.

$$MSE = \sqrt{\frac{\Sigma(\hat{y}-y)^2}{N-2}}$$

$$MSE = 0.8944$$

MSE - Mean square error and it can obtain by using this formula

$$MSE = \sqrt{\frac{\Sigma(\hat{y}-y)^2}{N-2}}$$

Date:

This MSE is also called as standard error of estimate (SEE)

The total loss of the model is 0.89 which is less than 1 i.e the model behaves well for training data as well as unseen data.

R^2 - coefficient of determination.

If $x = 10$ what would be y

$$y = 2.2 + 0.6 \times 10$$

$$y = 2.2 + 60.$$

$$y = 62.2$$

* Calculate total sum of squares (SST).

The formula for SST is ~~Sum over~~ $\Sigma(y_i - \bar{y})^2$

$$\Sigma(y_i - \bar{y})^2 = 6$$

* Calculate the coefficient of determination

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$R^2 = 0.6.$$

$$R = \{-1, 0, 1\}$$

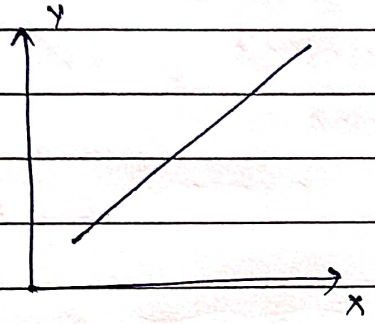
Page No.:
Date:

* Multiple Linear Regression

The term Multiple Linear Regression finds the regressor line b/w the dependent variable and independent variable.

It is a statistical method used to model the relation-ship b/w multiple independent variable and a single dependent variable.

LR



$$Y = b_0 + b_1 x$$

OR

$$Y = b_0 + b_1 x + \epsilon$$

$$Y = f(x)$$

MLR

Y	x_1	x_2

$$Y = f(x_1, x_2)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon$$

$$Y = b_0 + b_1 x_1 + b_2 x_2 \quad \text{OR}$$

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \epsilon$$

In Multiple linear regression we need to find $b_0, b_1,$ as well as b_2 from the equation $Y = b_0 + b_1 x_1 + b_2 x_2$

The formula for $b_0 = \bar{Y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 Y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

Page No.:
Date:

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

Q. Find the regressor line of the following tabular data.

QD (Y)	Price (x_1)	Income (x_2)	$y[(y-\bar{y})]$	$x_1[x_1-\bar{x})]$	$x_2[(x_2-\bar{x}_2)]$
100	5	1000	20	-1	200
75	7	600	-5	1	-200
80	6	1200	0	0	400
70	6	500	-10	0	-300
50	8	300	-30	2	-500
65	7	400	-15	1	-400
90	5	1300	10	-1	500
100	4	1100	20	-2	300
110	3	1300	30	-3	500
60	9	300	-20	3	-500

y^2	x_1^2	yx_1	yx_2	x_1x_2	x_2^2
400	1	-20	4000	-200	4000
25	1	-5	1000	-200	4000
0	0	0	0	0	160000
100	0	0	3000	0	90000
900	4	-60	15000	-1000	250000
225	1	-15	6000	-400	160000
100	1	-10	5000	-500	250000
400	4	-40	6000	-600	90000
900	9	-90	15000	-1500	250000
400	9	-60	10000	-1500	250000
		-300	65000	-5900	1580000

$\bar{y} = \frac{80\cancel{0}}{1\cancel{0}} = 80$

$\bar{x}_1 = \frac{6\cancel{0}}{1\cancel{0}} = 6$

$\bar{x}_2 = \frac{800\cancel{0}}{1\cancel{0}} = 800$

$$b_1 = \frac{(\Sigma x_2)^2(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)^2(\Sigma x_2^2)^2 - (\Sigma x_1 x_2)^2} = \frac{(1580000)^2(-300) - (-5900)(6500)}{(30)(1580000) - (-5900)^2}$$

$b_1 = -7.1882$

$$b_2 = \frac{(30)(65000) - (-5900)(6550 - 300)}{(30)(1580000) - (-5900)^2}$$

$$= \frac{1950000 - 1770000}{47400000 - 34810000}$$

$$= \frac{180000}{12590000}$$

$$b_2 = 0.01429$$

~~$y = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2$~~

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$80 - (-7.188)(6) - (0.014)(800)$$

$$b_0 = 111.928$$

$$y = b_0 + b_1 x_1 + b_2 x_2$$

$$Y = 111.928 + (-7.188)(60) + (0.014)(800)$$

$$y = 80$$

Date:

The Regression coefficient for x on y is written as b_{xy}.

y on $x = b_{yx}$

The coefficient of correlation can be represented as r and it can be written as

$$r = \sqrt{b_{xy} * b_{yx}}$$

where b_{xy} = coefficient of x on y.

b_{yx} = coefficient of y on x.

The value of r lies between {-1 to +1}

if $r = +1$ then it is highly correlated

$r = -1$ there is no correlation among x and y.

$r = 0$ there is partial correlation among x and y

Let us consider two regression line $x + 6y = 6$. ① and $3x + 2y = 10$. ②

find the correlation coefficient (r).

treat eqn ① as y on x.

$$x + 6y = 6.$$

$$6y = 6 - x$$

$$y = \frac{6 - x}{6}$$

in the $y = 6\left(1 - \frac{x}{6}\right)$

$$b_{xy} = -2/3.$$

the coefficient of correlation can be determined as

$$r = \sqrt{b_{xy} * b_{yx}}$$

$$r = \sqrt{\frac{-2}{3} * \frac{-1}{6}}$$

$$r = 0.33$$

r is

From the two given equations there is a partial correlation among two variables.

of two lines

Q. The two equations of the ~~line~~ regressions or the obtained in a correlation analysis are the following

$$2x = 8 - 3y \longrightarrow (1)$$

$$2y = 5 - x \longrightarrow (2)$$

Obtain the value of the correlation coefficient

$$x = \frac{8 - 3y}{2}$$

$$x = 4 - \frac{3}{2}y$$

$$x = 4 - \frac{3}{2}\left(\frac{5-x}{2}\right)$$

$$x = 4 - \frac{15 + 3x}{4}$$

$$4x = 16 - 15 + 3x$$

$$x = 1$$

$$2y = 5 - x$$

$$y = \frac{5 - x}{2}$$

$$b_{xy} = \frac{-3}{2}$$

$$b_{yx} = \frac{-1}{2}$$

Page No.:
Date:

$$r = \sqrt{\frac{3}{2} \times \frac{1}{2}}$$

$$\boxed{r = 0.86}$$

* A student obtained the following two regression equations Do you agree with him.

$6x = 15y + 21$

$x = \frac{15}{6}y + \frac{21}{6}$

$x = \frac{15}{6}y + \frac{7}{2}$.

$b_{xy} = \frac{15}{6}$

$21x + 14y = 56.$

$\frac{21x}{56} + \frac{14y}{56} = 0.$

$\frac{14y}{56} = \frac{21x}{56}$

$\frac{1}{4}y = \frac{21x}{56}$

$y = 84x/56$

$y = 3/2\,x$

$r = \sqrt{\frac{15}{6} \times \frac{3}{2}} = 1.93$

- If the obtained line is correct
- any regression coefficient is -ve and other are +ve
- if both the regression coefficients are >1
- It is not possible.

* The height of the fathers and sons are given below. find the height of the son when the height of the father is 70 inches.

Father	Son
71	69
68	64
66	65
67	63
70	65
71	62
70	65
73	64
72	66

65	59
66	62

$\bar{x} = 759/11 = 69$

$\bar{y} = 704/11 = 64$

Date:

The above tabular data indicates to find the regression equation and we have to fit the regression equation for answering the derived value.

The linear regression x on y can be written as

$$Y - \bar{Y} = r \frac{\sigma_{xy}}{\sigma_{yx}} (X - \bar{X})$$

In the above equation the deviations are taken from actual mean

$r = \frac{\sigma x}{\sigma y}$ can be written as $\frac{\Sigma xy}{\Sigma x^2}$

↓

deviation of actual mean.

$\frac{\Sigma xy}{\Sigma x^2}$

$\bar{X} = 69$

$\bar{Y} = 64$

The calculation requires

X	$(x-\bar{x})^x$	x^2	Y	$(y-\bar{y})^y$	y^2	xy
71	2	4	69	5	25	10
68	-1	1	64	0	0	0
66	-3	9	65	1	1	-3
67	-2	4	63	-1	1	2
70	1	1	65	1	1	1
71	2	4	62	-2	4	-4
70	1	1	65	1	1	1
73	4	16	64	0	0	0
72	3	9	66	2	4	6
65	-4	16	59	-5	25	20
66	-3	9	62	-2	4	6
					66	39

$$Y - \bar{Y} = \frac{\Sigma xy}{\Sigma x^2}(X - \bar{X})$$

$$Y - 64 = \frac{39}{74}(X - 69)$$

$$Y - 64 = 0.52(X - 69)$$

$$Y = 0.52X - 35.88 + 64$$

$$Y = 0.52x + 28.12$$

$$Y = 0.52 \times 70 + 28. = 64.4$$

Hence the likely height son when the height of the father is 70 inches shall be 64.53 inches

Q. Given the following data $\bar{x} = 36$ $\bar{y} = 85$ $\sigma_x = 11$ $\sigma_y = 8$. $r = 0.06$.

find the two regression equations

$$Y - 85 = 0.06\left(\frac{11}{8}\right)(X - 36)$$

$$Y - 85 = \frac{0.66}{8}(X - 36)$$

$$Y - 85 = 0.082(X - 36)$$

$$Y - 85 = 0.08x - 2.88.$$

$$Y = 0.08x - 2.88 + 85$$

✓ $$Y = 0.08x + 82.12$$

$$Y - 85 + 2.88 = 0.08x.$$

$$Y - 82.12 = 0.08x$$

$$\frac{Y - 82.12}{0.08} = x.$$

$$\frac{y}{0.08} - 1026.5 = x.$$

Date:

$$b_{x \text{ on } y} = x - \bar{x} = r \frac{\sigma y}{\sigma x} (y - \bar{y})$$

$$x - 36 = 0.06 \left(\frac{8}{11}\right)(y - 85)$$

$$x - 36 = 0.04(y - 85) \quad \checkmark x = 0.04y + 32.6$$

$$x = 0.04y - 3.4 + 36$$

Q. Find the two regression equation for the following two series what is most likely value of x when $y = 20$ and most likely value of y when $x = 20$

		x		y		
x	y	$(x - \bar{x})$	x^2	$(y - \bar{y})$	y^2	xy
35	23	5	25	-2	4	-10
25	27	-5	25	2	4	-10
29	26	-1	1	1	1	-1
31	21	1	1	-4	16	-4
27	24	-3	9	-1	1	3
24	20	-6	36	-5	25	30
33	29	3	9	-4	16	-12
36	30	6	36	5	25	30
240	200	0	142		92	26
8	8					50

$\bar{x} = 30 \quad \bar{y} = 25$

y on x.

$$y - \bar{y} = r \frac{\sigma y}{\sigma x} (x - \bar{x})$$

x on y.

$$(X-\bar{X}) = r\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

As there is no information about coefficient of correlation then we assume deviation takes from assumed mean Hence

$$r\frac{\sigma x}{\sigma y} = \text{Sum of } \frac{\Sigma xy}{\Sigma x^2}$$

$$X - 30 = 0.35(Y-25)$$

$$X = 0.35y - 8.75 + 30$$

$$\hat{x} = 0.35y + 21.25 = 28.25$$

~~Substitute~~

Similarly the regression equation y on x,

$$y - 25 = 0.54(x - 30)$$

$$y = 0.54x - 16.2 + 25$$

$$y = 0.54x + 8.8$$

$$y = 19.6$$

Q. You are given below the following information about advertisement expanditure and shells. the correlation coefficient 0.8.

Calculate the two

find the likely shell when advertisement expandi-ture 25.

What would be the advertisement budget if the company wants attained shells target of 150 cr.

x on $y = b_{xy}$.

y on $x = b_{yx}$.

M T W T F S S

Page No.:

Date:

The regression equation x on y

	Adv. Ep.	Sales y
Mean	20	120
SD	5	25.

$\bar{x} = 20 \quad \bar{y} = 120 \quad \sigma_x = 5 \quad \sigma_y = 25$

$x - \bar{x} = 0.8 (y - \bar{y})$

$x - 20 = 0.8 (y - 120)$

$x = 0.8y - 96 + 20$

$x = 0.8y - 76$

$x - 20 = 0.8 \left(\frac{5}{25}\right)_5 (y - 120)$

$x - 20 = 0.16y - 19.2$

$x = 0.16y - 19.2 + 20 \quad 0.8$

$x = 0.16y - 19.2 + 20 \cdot 0.8 = 0.16(25) + 0.8 = 4.8$

$y - \bar{y} = 0.16 (x - 20)$ $\qquad y - \bar{y} = 0.8 \left(\frac{25}{5}\right)^5 (x - \bar{x})$

$y = 0.16x - 3.2 + 120$

$y = 0.16x + 116.8.$

$y - 120 = 4x - 80$

$y = 4x - 80 + 120$

$y = 4x + 40.$

$y = 640$

Date:

$\bar{X}$ = is the mean of x series

$\bar{Y}$ = is the mean of y series

$r \frac{\sigma x}{\sigma y}$ is the regression coefficient of x on y.

$r \frac{\sigma y}{\sigma x}$ is the regression coefficient of y on x.

Q. prove $r = \sqrt{b_{xy} * b_{yx}}$.

We know that

$$b_{xy} = r \frac{\sigma x}{\sigma y} \longrightarrow (1)$$

$$b_{yx} = r \frac{\sigma y}{\sigma x} \longrightarrow (2)$$

$$b_{xy} * b_{yx} = r \frac{\sigma x}{\sigma y} \cdot r \frac{\sigma y}{\sigma x}$$

$$r = \sqrt{b_{xy} * b_{yx}}.$$

Assumptions for the coefficient of correlation

(i) Both the regression coefficient will have the same sign either there will be a +ve or -ve it is never possible that one coefficient is +ve another one is -ve.

(ii) The coefficient of correlation are should not greater than 1. must be less than 1.

(iii) If $b_{xy} = 1.2$

$b_{yx} = 1.4$

$r = 1.29$.

M T W T F S S

Page No.:

Date:

(iii) The coefficient of correlation will have the same sign as that of regression coefficient.
If regression coefficient have -ve sign r be will also be -ve and if regression coefficient +ve r is also +ve.

II

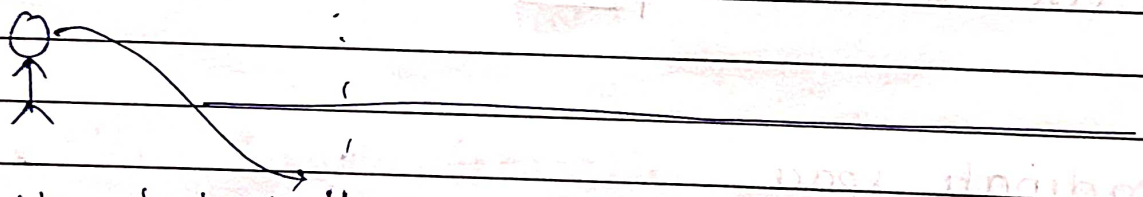
Page No.:
Date:

Logistic Regression

(Predict the future data point)

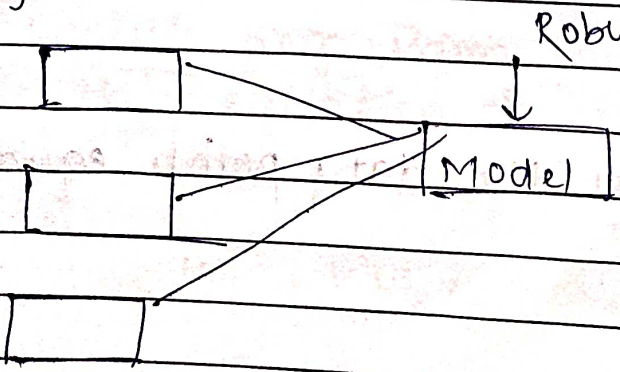
The main objective of the LogisticRegression is to find a line which is perfectly separable between +ve and -ve classes. It is a binary class classification problem where the target variable (Y) is a binary nature

	INPUT FEATURE		OUTPUT FEATURE	
	Test 1	Test 2	Y	↓
	1	2	1	+1
	3	4	0	-1
	5	6	1	
	⋮	⋮	0	
	⋮	⋮	⋮	
Testing	8	⋮	⋮	
	9	⋮	⋮	
	10	18	0	
	⋮	⋮	⋮	



Test 1 and Test are called as the i/p feature and y is the target feature.

Let us take the 10 data points out of this 80 % data will be training and 20 % data will be used for testing



Once the model is trained and it is ready to accept the unseened data.

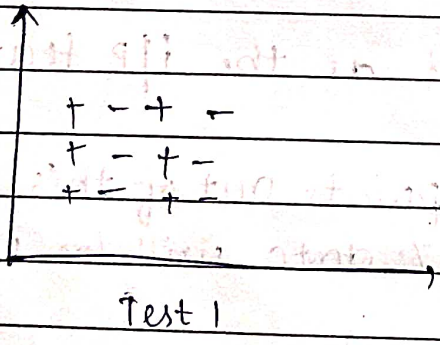
	test 1	test 2	Y
Newdata	11	8	?

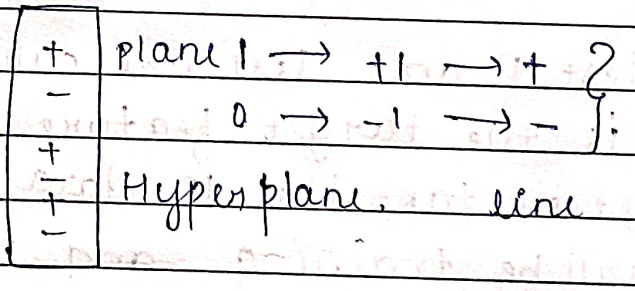
The model predicts either 1 or 0. If it 1 then the person is having a symptom otherwise not.
If the model says the robust then the difference b/w training and testing is minimized.
that means a model generalises.

Training Accuracy	Testing Accuracy	Differnce
92 %	91 %	1 %

If we have the two I/P features then we can visulise with the line if we have 3 I/P features we can visulise with the plane otherwise hyperplane.

Why 2D coordinate space
Test 1 and Test 2 consist of +ve and -ve

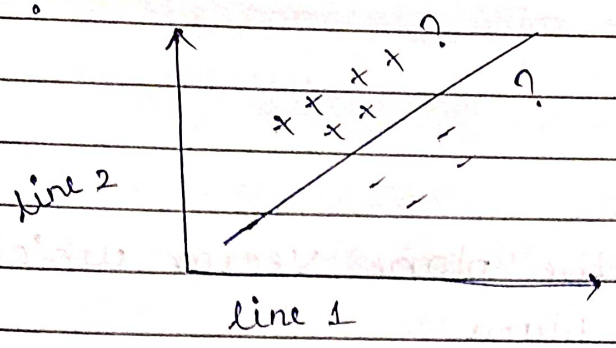




* Linearly separable
The concept says that all the data points equally distributed.

Date:

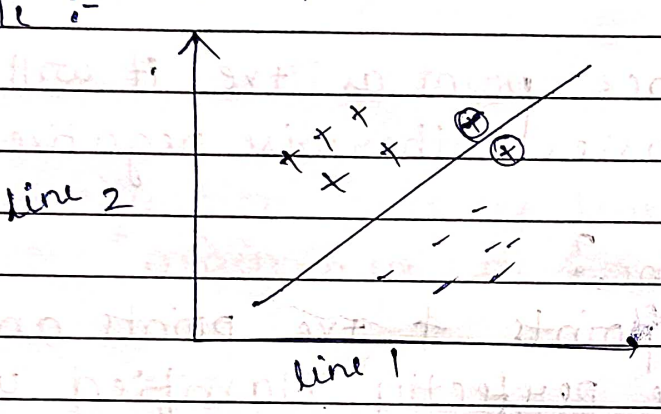
example :-



* Almost linearly Separable

In this case all most all the data points are classified and some points available both sides hence it is called as Almost linearly separable.

example :-



* Non- linearly Separable.

You take any line in any direction you cannot separate the +ve and -ve classes in the dataset that's why we calle as non-linear lines.

Logistic regression won't work with this dataset.

Q. How this line helpful for predicting ~~feature~~ future data points ?

The Logistic regression found as the best line it will use this equation of the line for predicting future data points. Now apply equation of the line

$$w_1 x_1 + w_2 x_2 + w_0 = 0 \longrightarrow (1)$$

Date:

$w : [w_1, w_2]$

$x : [x_1, x_2]$

where w = Normal to the plane vector which is 1er to the plane

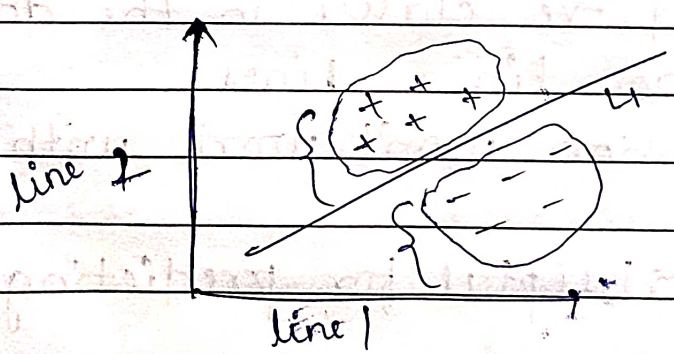
Assume, w is the 1er line if the point is available on the side then it is a +ve distance otherwise -ve distance

based on this concept our logistic regression our future predicted data point.

If you got the distance point as +ve it will predict the +ve classes otherwise negative classes.

Q. Why it is the best line.?

Out of this 8 data points 4 +ve points and 4 -ve points are perfectly classified with each other out ~~of~~ of these 8 lines L_1 is the best line ~~as~~ because maximum data points are correctly classified.



Page No.:
Date:

If we put all the concepts into the equations to identify the best line we use

$$z_i = y_i d_i \longrightarrow (1)$$

where y_i is the actual value
d_i is the predicted value.
z_i is the Resultant.

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

$w : [w_1, w_2]$ $\quad x : [x_1, x_2]$

$$\boxed{w^T x_i + w_0 = 0} \longrightarrow (2)$$

w_0 = Origin

* How to calculate d_i ?

$$d_i = \frac{w^T x_i + w_0}{\|w\|} \longrightarrow (3)$$

From equation (3) if the line passes through the origin then d_i can be written as

$$d_i = \frac{w^T x_i}{\|w\|} \longrightarrow (4)$$

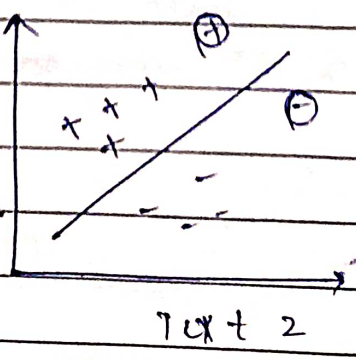
w^T - Transpose of w

From equation (4)
w is the unit vector.

$$d_i = w^T x_i \longrightarrow (5)$$

equation (4) and (5) we are taking as our assumptions and $y_i d_i$ can have the followings

$z_i = y_i d_i$



future data points

1. If y_i is +ve
   d_i is +ve
   then the resultant is $z_i = y_i d_i$ is +ve.

2. If y_i is +ve
   d_i is -ve
   then the resultant is $z_i = y_i d_i$ is -ve.

3. If y_i is -ve
   d_i is +ve
   then the resultant is $z_i = y_i d_i$ is -ve.

4. If y_i is -ve
   d_i is -ve
   then the resultant is $z_i = y_i d_i$ is +ve.

$$\text{argmax} \sum_{i=1}^{N} y_i d_i$$

$w = w_1, w_2$

Let us consider to calculate $y_i d_i$ all the datapoints and sum them

$$\sum_{i=1}^{N} y_i d_i$$

$y_1 d_1 +$
$y_2 d_2 +$
$y_3 d_3 +$
$y_4 d_4 +$
:
:
$y_n d_n$

Date:

The general term of the equation is

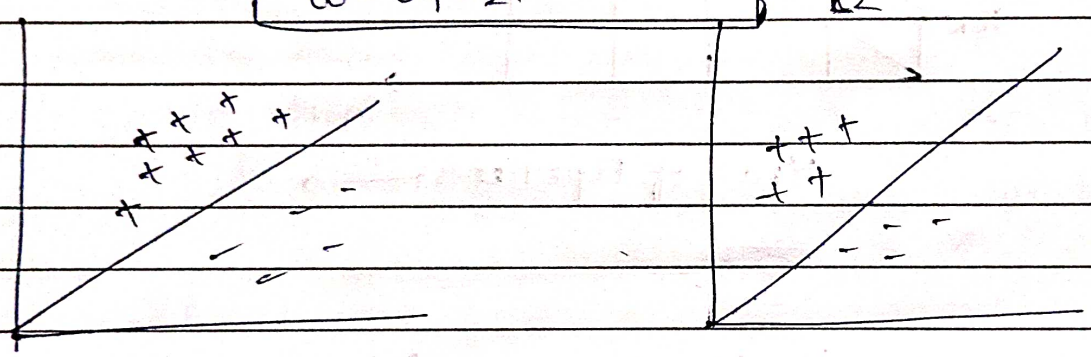
$$\sum_{i=1}^{N} y_i d_i \longrightarrow (6)$$

Let us take two lines L_1 and L_2 calculate the summation of both L_1 and L_2 lines for which line summation value got maximum.

Let L_1 = 8 +ve and L_2 = 5 +ve out of these two lines L_1 is having correctly classified and the logistic regression equation can be written as for best line is

$$\text{argmax} \sum_{i=}^{N} y_i d_i \longrightarrow (7)$$

$$w = w_1 w_2.$$



The above is called as optimization function the meaning of this equation is for what values of w the entire summation function will be maximum that w's need to be calculated.

We will solve this equation and identify w_1 and w_2 in equation 7 and it can be written as $y_i d_i$

Substitute $d_i = w^T x_i$

Hence eq① can be written as

$$z_i = y_i(w^T x_i) \longrightarrow (8)$$

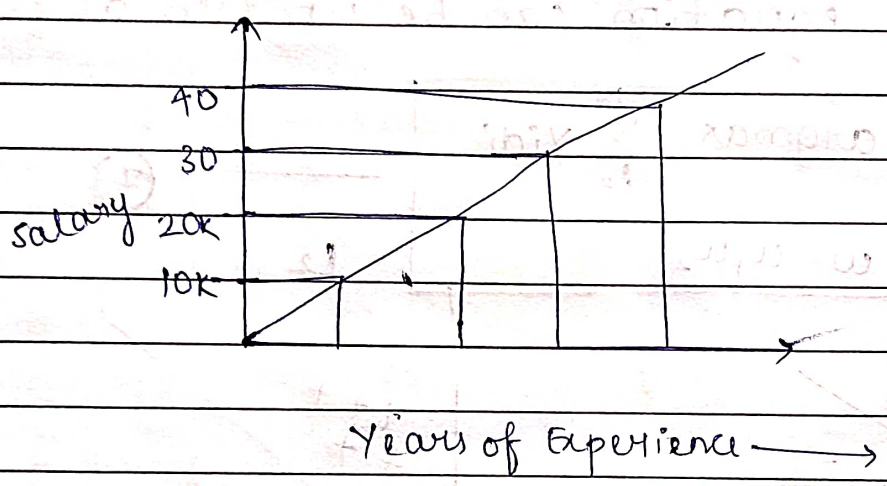
Substitue $w_1 x_1$

Hence $z_i = y_i(w_1 x_1 + w_2 x_2) \longrightarrow (9)$

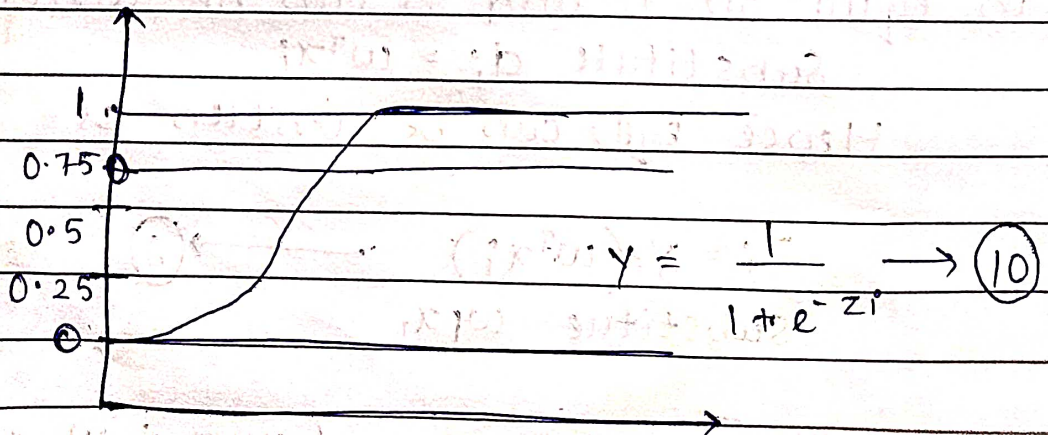
In this equation (9) Y_1, X_1, X_2 are given and w_1 and w_2 are required to calculate.

Let us take two different examples

① An Organization want to determine the employees based on the year of experience in this example we will apply the linear regression and we can predict the salary.



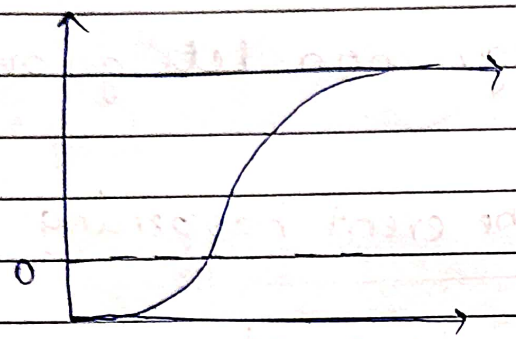
② In this example, if the Organization want to know wheater employee would get employee based on their performance. In this case the linear regression won't help us. We will clip the line from 0 and 1 and convert it into a sigmoid curve.



$$y = \frac{1}{1+e^{-zi}} \quad \rightarrow (10)$$

* Difference b/w Logistic Regression and Linear Regression.

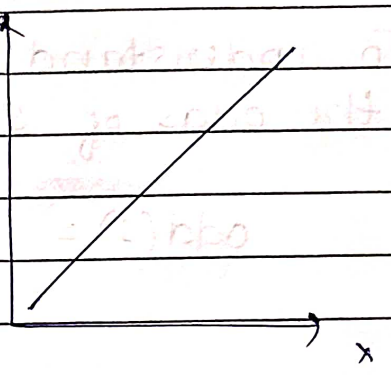
Classification



Sigmoid function
(S-shaped)

$$y = \frac{1}{1+e^{-zi}}$$

Prediction



$y = ax+b$

$y = b_0 + b_1 x$

(i) This approach is useful when the nature of the dataset says binary or multi class classification approach

* It is used to predict the future events by taking dependent and independent features.

(ii) In Logistic regression the features are categorical nature.

* the response of linear regression is a continuous nature.

(iii) In linear regression the best fit line obtain in the straight line.

(iii) In Logistic regression the line is obtained as Sigmoid shape.

Page No.:
Date:

* In linear regression the equation is

$$y = b + ax$$

(iv) In LR the equation is

$$y = \frac{1}{1 + e^{-zi}}$$

* How does the logistic regression algorithm works.?

To understand the Logistic regression lets go over the odds of success.

$$\text{odd(s)} = \frac{\text{Probability of the event happening}}{\text{Probability of the event not happening.}}$$

$$\text{odd(Q)} = \frac{P}{1-P} \quad \text{——— (11)}$$

The values odds orange are 0 to ∞
and the values of the probability lies 0 to 1.

Consider the equation of the straight line $Y = B_0 + B_1 x$
Now to predict the odds of success, we take log on odd formula

$$Y = B_0 + B_1 x.$$

$$\log\left(\frac{P(x)}{1-P(x)}\right) = B_0 + B_1 x. \quad \text——— (12)$$

Date:

exponential on both sides.

$$e^{\ln\left(\frac{P(x)}{1-P(x)}\right)} = e^{\beta_0+\beta_1 x}$$

$$e^{\ln(20)} = x.$$

$$\frac{P(x)}{1-P(x)} = e^{\beta_0+\beta_1 x} \quad (13)$$

$$\frac{P(x)}{1-P(x)} = y.$$

In the above eqn we substitute y on right hand side and eqn becomes

$$\frac{P(x)}{1-P(x)} = y. \longrightarrow (14)$$

Simplify the eqn (14) $\quad \therefore y = e^{\beta_0+\beta_1 x}$

$$P(x) = y\,(1-P(x))$$

$$P(x) = y - yP(x)$$

$$P(x) + yP(x) = y.$$

$$P(x)(1+y) = y$$

$$P(x) = \frac{y}{1+y} \longrightarrow (15)$$

Substitute the value of y in these equations. (15)

$$P(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (16)$$

$$P(x) = \frac{e^{\beta_0+\beta_1 x} / e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x} / e^{\beta_0+\beta_1 x}}$$

The sigmoid function equation can be obtained as

$$P(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \longrightarrow (17)$$

We know the eqn of the straightline $Z = \beta_0 + \beta_1 x$
Substitute in the equation in (17)

$$P(x) = \frac{1}{1+e^{-z}} \longrightarrow (18)$$

Q. Let us consider the student dataset which and enterance based on the historical who are selected and non-selected based on the Logistic represent.

$b_0 = 1$

$b_1 = 8$

assuming the mark = 60

compute the resultant class

$$Y = b_0 + b_1 x_1 \qquad 1 + 8 \times 60 = 481$$

$$P(x) = \frac{1}{1+e^{-z}}$$

$$P(x) = \frac{1}{1+e^{-481}} = 1$$

Page No.:
Date:

The threshold value as 0.5 compare the result with the threshold value and interpret it.

We observed that the resultant value is greater than the threshold (0.5). Therefore, the candidate mark with 60 is selected.

Q. Let us consider the dataset having two features hours study, pass and fail.
Apply the logistic regression to answer the following question.

(i) Calculate the probability of pass of the student who studied 33 hours.

(ii) Atleast how many hours student should study that max he will pass the ~~score~~ score with the pro. max 95%.

The optimization produce the model is log(odds) = -64 + 2* hours

Assume the model suggested the ~~be~~ optimizer for odds of the passing score is $\boxed{\log(odds) = -64 + 2*h}$

hours	fail/Pass
29	0
15	0
33	1
28	1
39	1

Sol- (i)

$$P(x) = \frac{1}{1+e^{-z}}$$

$$\log(odds) = -64 + 2 * hours$$

$$z = -64 + 2 * 33$$

$$z = 2$$

$$P(x) = \frac{1}{1+e^{-2}}$$

$$P(x) = 0.88$$

(ii)

$$P(x) = \frac{1}{1+e^{-z}} = 0.95$$

$$0.95\,(1+e^{-z}) = 1$$

$$1+e^{-z} = \frac{1}{0.95}$$

$$1+e^{-z} = 1.0526$$

$$e^{-z} = 1.0526 - 1$$

$$e^{-z} = 0.0526$$

$$\ln(e^{-z}) = \ln(0.0526)$$

$$z = 2.94.$$

$$2.94 = -64 + 2 * \text{hours}.$$

$$66.94 = 2h$$

$$\frac{66.94}{2} = h$$

$$\boxed{33.47 = h}$$

L_1 Regularization [Lasso Regression]

M T W T F S S
Page No.:
Date:

Ridge Regression (L_2 Regularization)

- The objective of the Ridge regression is to prevent the over fitting to meet the models generalization ability.
- It is used when your dataset is having multi co-linearlity features.

- When the model is having high variance / low bias and high bias with low variance as well as highly correlated of one attribute with other attributes then Ridge Regression will be useful.

- The Ridge Regression can be defined as cost function of linear regression plus (+) penality term. The above concept is called as a Regularization.

Mathematically, The Ridge Regression can be defined as minimise $\left(SSE + \lambda \sum_{j=1}^{p} B_j^2 \right) \longrightarrow (1)$

$\rightarrow$ linear regression (sum of squares of error)

λ term is need to find overfitting / underfitting.

Where λ is a hyperparameter of L_2 regularization
(λ is the regularization parameter)

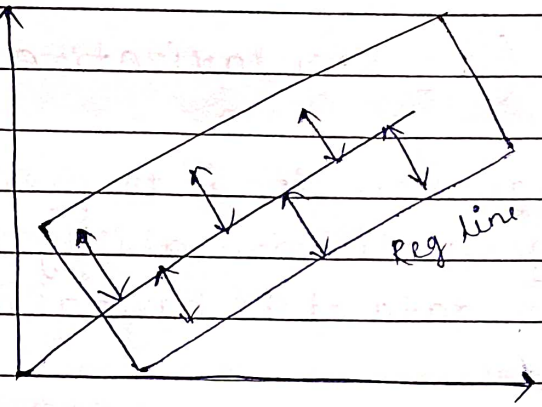
B – coefficient of linear regression model.

p – Number of features.

The above equation can also be written as

Ridge regression = Cost Function (of LR) + Hyperparameter of L_2

$\longrightarrow (2)$



Q. How to avoid overfitting ?

To avoid the overfitting issues we required the cross validation technique that helps in estimating the error over the test set and it will decide what parameters works best for the model.

The cost function of the simple LR is

$$\sum_{i=1}^{M} (y_i - \hat{y})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 \longrightarrow (3)$$

where

M is the number of instances

p is the number of features.

$\hat{y} = w[0] x[0] + w[1] * x[1] + \cdots w[N] x[N] + b$

b is the constant

This can be written as the linear model based on n number of features and considering a single feature $w[0]$ is the slope

b is the intercept.

The LR looks for optimizing w and p such that it minimizes the cost functions

Date:

In Ridge regression the cost function is altered by adding the penality term which is equivalent to square of the magnitude of the coefficient and the equation can be written as.

$$\sum_{i=1}^{M}(y_i-\hat{y})^2 + \lambda = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2 \longrightarrow (4)$$

Which is called as cost function of linear regression. Where w_j is the square of magnitude of the coefficients.

The ideal value of the λ should be 0 but not exactly zero. The RR shrinks the coefficient and it allow to reduce the complexity of the model.
So the lower the λ value the model provides good results.

if $\lambda = 0$
then RR became LR.

* Let us take one example $y = \lambda_1 + 1{\cdot}2X_1 + 20X_2 + 39X_3$
In the above case X_3 plays the important feature the magnitude or the coefficient is high and it will give less importance to X_1 and X_2.

Our objective is to minimise the high coefficient variables Otherwise it requires more expansive cost and the model becomes complex.
To avoid the above problem (Reducing the models complexity) Regularization is used.

* The ridge regression can be defined

$$\text{Ridge} = \text{Loss} + \lambda \|w\|^2 \longrightarrow (5)$$

where $\|w\| = w_1^2 + w_2^2 + \cdots w_n^2$.

Hence, Ridge = Loss + penality. $\longrightarrow (6)$

M T W T F S

Page No.:

Date:

The periality is to reduce the losses where γ is the constant. $\|w\|$ is the ~~Factor~~ vector of coefficients.

M T W T F S S
Page No.:
Date:

Lasso Regression → Feature selection technique

The main objective of Lasso regression is to minimise the difference b/w observed value and predicted value as well as penality term λ and absolute size of the coefficients.

OLS - Order list square regression.

As the coefficients are determined ~~abso~~ by the absolute size it help us to solve the overfitting problems.

LASSO - Least absolute shrinkage and selection operation.

It is a kind of LR model that use a penality term and called L1 regularisation.

* OBJECTIVE FUNCTION OF LASSO REGRESSION

In Lasso regression the objective function is to define the as t
between actual and predicted.

Mathematically,

$$\text{minimize } \frac{1}{2m} \sum_{i=1}^{N} (Y_i - \hat{Y})^2 + \lambda \sum_{j=1}^{P} ||B_j||$$

{Sum of squares of the residual (RSS)}
{↳ cost function of Linear regression}

Where,

Lasso = RSS + penalty.

n = number of observations / instances

p = number of features / predictors.

B = Coefficients.

Y_i = Actual

$\hat{y}$ = predicted

λ = hyperparameter / a constant for Lasso.

The above function is called as a objective function of Lasso regression.

The constants controls the strength of the penalty term the higher value of the λ gives more regularization.

The cost function of Lasso regression can be written as

$$\sum_{i=1}^{M} (y_i - \hat{y})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{i=1}^{p} W_i X_{ij} \right)^2 + \lambda \sum_{i=1}^{p} \| B_j \|$$

Let us take one example where,

$$y = \alpha_1 + 1.2x_1 + 20x_2 + 39x_3$$

After some Iteration the y becomes

$$y = 0.9 + 0x_1 + 0x_2 + 5x_3$$

x_1	x_2	x_3

In the above example we understood that the higher coefficient impact to the model and it requires more cost so that the model becomes complex. Inorder to avoid this situation we use the concept of regularisation.
The Lasso regression not only helps in reducing overfitting but also help us to reduce in the feature selection purpose.

* The main objective of Lasso is to use a regularization as well as feature selection. This type of regularization can leads to zero coefficients
* Some of the features are completely neglected (x_1 and x_2)

In the above example x_1 and x_2 variables are completely neglected. This values are not important for determining the value of Y. Lasso is called as a feature selection technique.

$$Y = 0.9 + 5x_3.$$

The complexity reduced because unecessary features are eliminated Overfitting problem solved because we have done feature selection. Whenwe increase the value of λ some of the variable become zero.

* ~~ORDE~~ ~~OD~~ II

* ORDER LEAST SQUARE REGRESSION

The main objective of OLS regression is to minimize the sum of square difference between the actual and the predicted values by the linear model.

The main role of OLS is to estimate the coefficients of a linear regression model.

At the end $\hat{y}$ can be estimated.

Whine we perform linear regression we use OLS method to estimate the coefficients of the Linear equation that best fits the data.

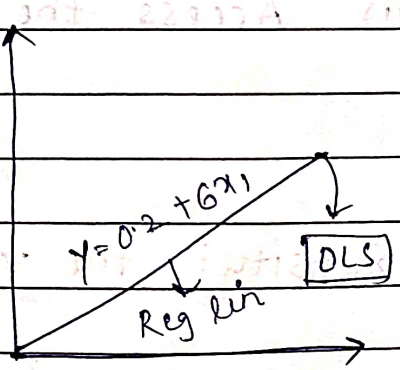
$$y = B_0 + B_1 x$$

$$y = 0.2 + 6(2)$$

$$y = 0.2 + 12$$

$$y = 12.2.$$

x
2
3
4

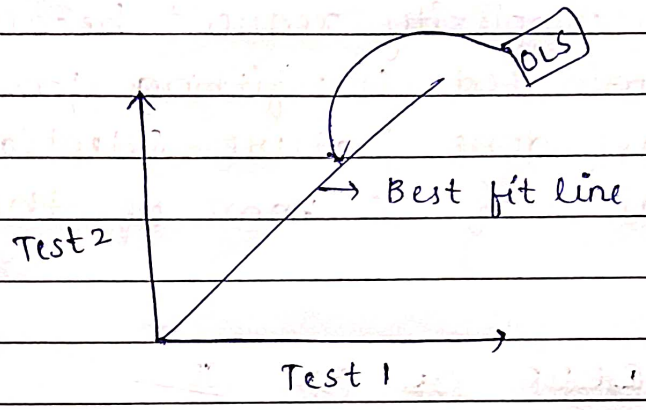


Page No.:
Date:

Steps.

(i) Let us take the dataset X and Y apply the OLS regression technique to estimate the regression equation. Use the hypothesis testing to determine wheather the estimated coefficients are statistically significant or not.

x	y
1	2
2	4
3	5
4	4
5	5



The estimated regressor line

$$Y = 2.2 + 0.6x \longrightarrow \text{regressor line}$$

(ii) $\hat{Y}$ value.

2.8
3.4
4
4.6
5.2

(iii) Access the goodness of fit

$$R^2 = 1 - \frac{\Sigma(Y_i - \hat{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

Subsitute the value of Nuemarotor and denomenator

$$R^2 = 1 - \frac{\Sigma(Y_i - \hat{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

$$R^2 = 0.60$$

Page No.:
Date:

The R^2 value indicates that 60% of the variance Y is explained with the linear regression equations.

(iv) ~~The OLS~~ In OLS we will use the hypothesis testing to determine whether statistically significant or not.

* The Null hypothesis

$H_0 : \beta_0 = 0$

β_0 = intercept.

Similarly, $H_0 : \beta_1 = 0$

β_1 = slope

* Alternate hypothesis

$H_0 : \beta_0 \neq 0$

(Intercept)

$H_0 : \beta_1 \neq 0$

(slope)

* performe the hypothesis test for β_0

(i) estimated coefficient (β_0) = 2.2

(ii) SE (Standard error of β_0) [From linear regression O/P]

(iii) DF - Degree of freedom

$DF = n-2$

where, n is the number of observation.

(iv) Calculate the t value t = distributions

$$t = \frac{\beta_0}{\text{SE of } \beta_0}$$

(v) Compare the calculate t value with a critical value from the t distribution table.

(vi)

SE(β_0)

By calculating the SE of the intercept (β_0) the formula is

$$SE = \sqrt{\frac{MSE}{n} + \frac{(\bar{x}^2)}{SS_x}}$$

Page No.:
Date:

where MSE Mean square error / residual error

n - number of observation

SS_x - Sum of square of independent variable

$\bar{x}$ - Mean of x.

In this equation Known values are

$n = 5$

$\bar{x} = 3$

$$MSE = \sqrt{\frac{\Sigma(\hat{y} - y)^2}{N-2}}$$

$$MSE = (1 - R^2) * var(y)$$

$$= (1 - 0.6) * var(y)$$

$$= 0.4 * \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{n-1}$$

$$= 0.4 * 1.5$$

$$MSE = 0.6$$

* ~~SE~~ ~~Substitute~~ To calculate SS_x we have the formula

$$SS_x = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$SS_x = 10$$

Date:

Plugged in the value of

$$SE = \sqrt{\frac{MSE}{N} + \frac{(\bar{x})^2}{SS_X}} = \sqrt{\frac{0.6}{5} + \frac{9}{10}} = 1.009$$

The degree of freedom = 5-2

$$DF = 3$$

Calculate the t value

$$\beta_0 = 2.2$$

$$t = \frac{\beta_0}{SE(\beta_0)} = \frac{2.2}{1.00} = 2.2$$

→ ~~The~~ comparision ~~the~~

The calculate t value is = 2.2

and our DF is 3

we can find the critical t value from t distribution table at level of significance (α) is 0.05 is 3.182.

Since 2.2 falls below 3.182 We donot have the avidence to reject the null hypothesis

$$H_0 : \beta_0 = 0$$

Therefore, we conclude that intercept is not statistically significant different from 0 at 0.05 level of significance.

UNIT-2

M T W T F S S
Page No.:
Date:

K-NN Algorithm :- (K-Nearest Neighbour)

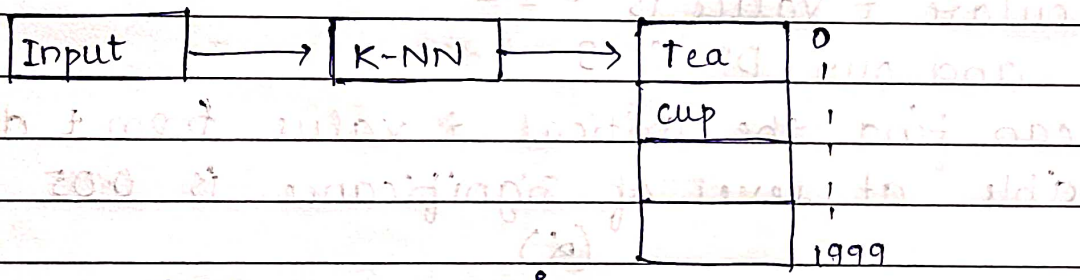
(i) KNN stands for K-Nearest Neighbour

(ii) It is a supervised Machine learning algorithm. That can be used for both regression and classification.

(iii) A classification problem has a discrete value at its output. Each row in the dataset typical called an example, observation or a datapoint while each column is called a predictor, dimension, independent variable or feature.

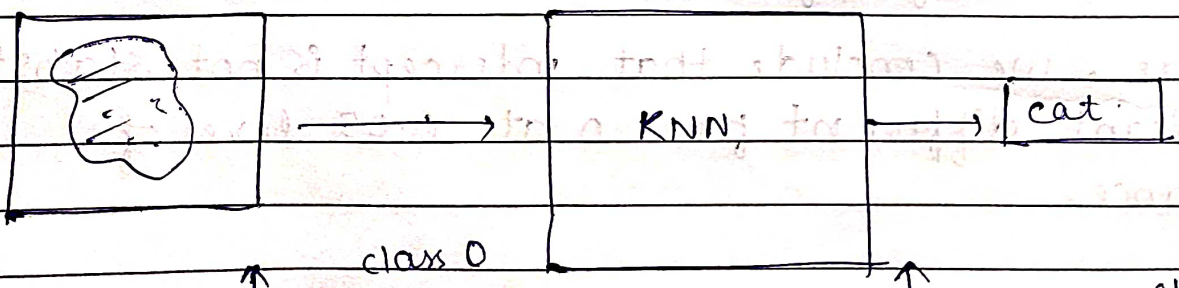
(iv) KNN algorithm captures the idea of similarity. Sometimes it is called as proximity or closeness.

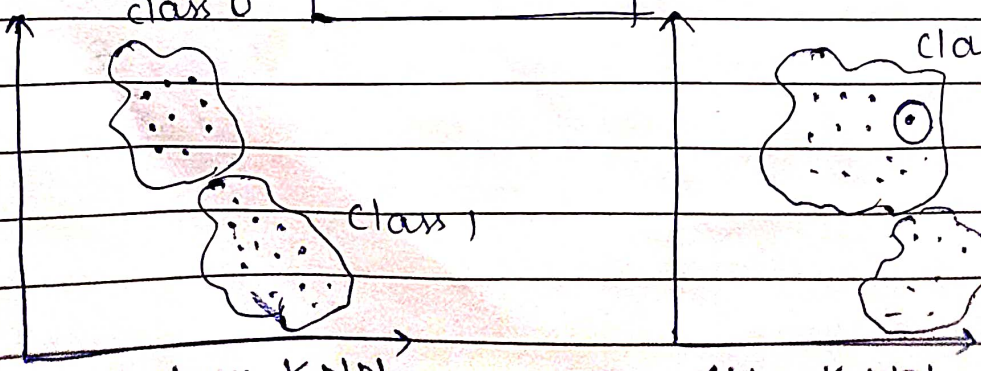
(v) KNN algorithm assumes the similarity b/w the new data and available data and put the new data into the catogory that is most similar to available catogories.



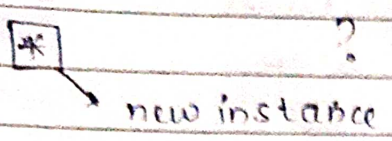
80% ⟶ Training

20% ⟶ Testing.





KNN algorithm can also used for regression as well as classification but mostly used for classification algorithm. It is a non-parametric algorithm because it does not make any assumptions to the underlined data.



promit proximity
closeness
⇓
similarity.

The new instance can be decided by using the distance measure formula

① ~~Euclidian~~ Euclidian distance.

$$\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}.$$

② Manathon

$$|x_2-x_1| + |y_2-y_1|$$

It is also called as Lazy learner algorithm It does not learn from the training set immediately instead it stores the dataset and at the time of classification it performs an action from the dataset.

* How does K-NN works ?

(i) Initialise the number of K neighbours
(ii) Calculate their distance using eqclidian distance
(iii) Take the K nearest neighbour as per euclidian distance.
(iv) Count the number of points in each category.
(v) Assign the new datapoints to the category for which the number of neighbour is maximum.

M T W T F S S

Page No.:

Date:

Q. Let us take one dataset which consist of the following information apply KNN classifier to predict their class level.

HT	wt	class	Distance
167	51	UW	6.70
182	62	N	13
176	69	N	13.41
173	64	N	7.61
172	65	N	8.2
174	56	UN	4.12
169	58	N	1.41
173	57	N	3
170	55	N	2
170	57	?	0
		N	

Find the distance b/w new instance to first instance. to estimate using distance formula.

$$\sqrt{(170-167)^2 + (57-51)^2} = 6.70$$

Idealy K=3 or K=5

M T W T F S S
Page No.:
Date:

(4) Restructure the table according to the minimum distance.

HT	Wt	class	Distance.	Rank
169	58	N	1.41	1
170	55	N	2	2
173	57	N	3.	3
174	56	UN	4.1.	4
167	51	UN	6.70.	5
173	64	N	7.6.	6
172	65	N	8.2	7
182	62	N	13	8
176	69	N	13.41	9
170	57	N		

(5) Provide the rank according to the minimum distance.

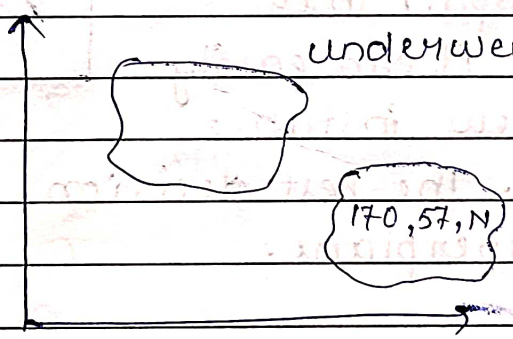
(6) Step-3

Assume K=3

K=1 ⟶ (N)

K=2 ⟶ (N)

K=3 ⟶ (N)



As 170 (Height) and 57 (Weight) is close to the three nearest values so the new instance will be of normal class.

Page No.:
Date:

SVM - Support Vector Machine

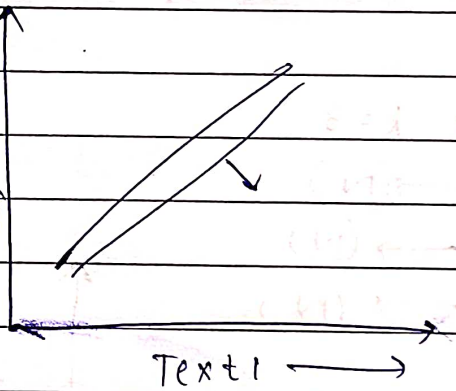
SVM is a supervisior learning. It can be possible both classification as well as regression. If we predict the class level then it is called as Support Vector for Regression (SVR). If it classify the class level then it is called as support Vector for classification (SVC).

The main objective of SVM is to predict the level class level by using bypas the concept of hyperplane.

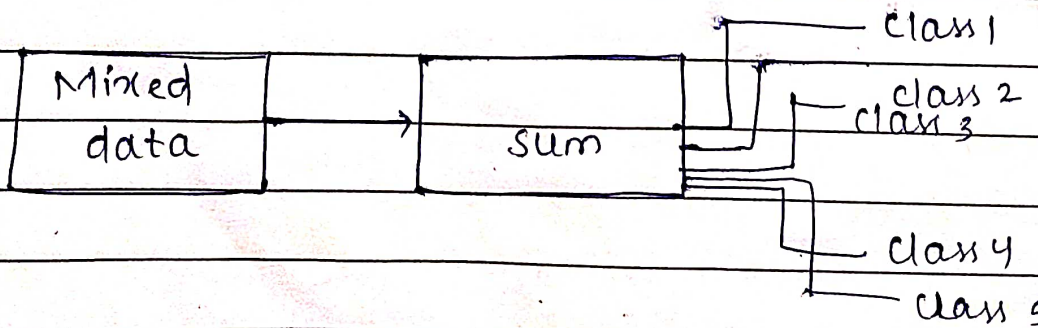
Representation of hyperplane.

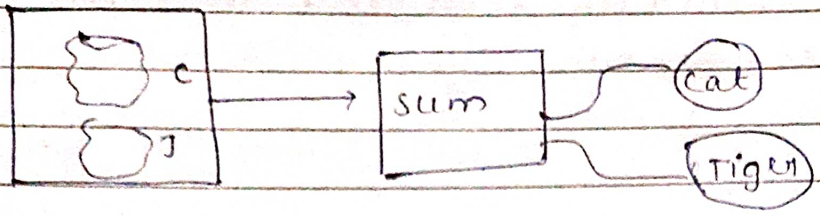
It is also called as a decision boundary which will classify both the +ve and -ve classes.

It is the goal of SVM to create the best line or decision boundary that can seggregate n dimensional space into classes so that it can easily classify the new instance in the future. The best decision boundary is called as hyperplane.



There may be 'n' number of boundaries. we choose one of the best from it we choose and call it as hyperplane.

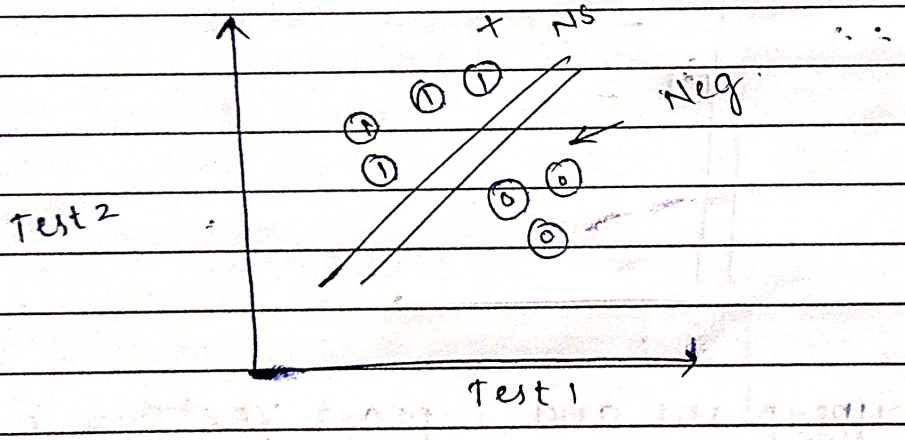




Datapoint	Test 1	Test 2	Result	
1	-	-	Yes	1
2	-	-	Yes	1
3	-	-	Yes	1
4	-	-	No	0
5	-	-	No	0
6	-	-	No	0
New instance 7	-	-	Yes.	?

Predict the class level.

By seeing the result of the above table we can generate / estimate identify the hyperplane which will divide the +ve and -ve classes.

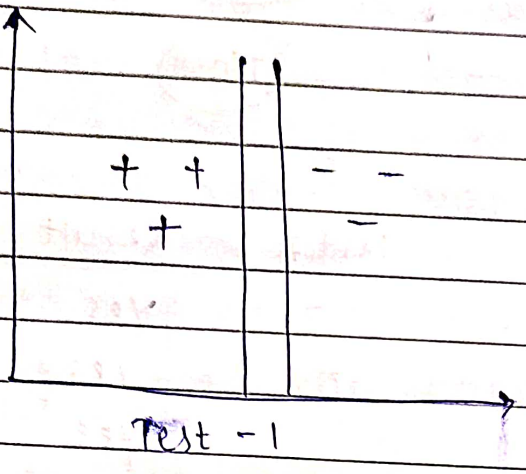


There are two types of SVM

(i) Linear SVM

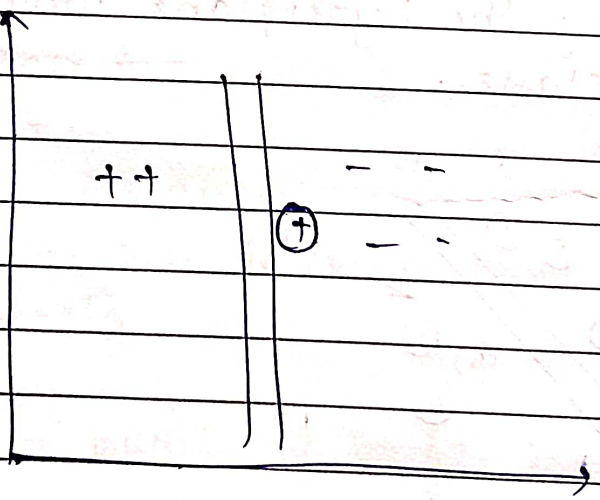
In linear SVM we use the linearly separable data which means if a dataset can be classified into two classes by using a single straight line then such

data is termed as linearly separable and the classifier is called as linear SVM.



(ii) Non linear SVM

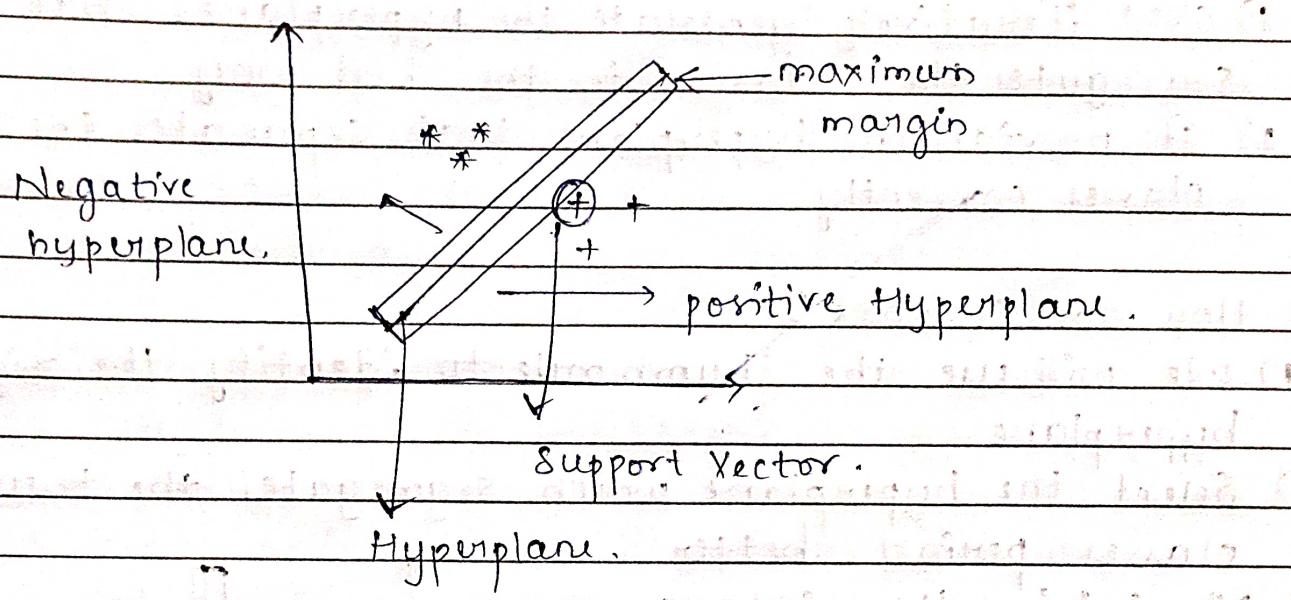
It is used for non-linearly separated data which means if a dataset cannot be classified by using a straight line then such data is termed as non-linearly separable and the classifier is called as Non-linear SVM.



* The concept of hyperplane and support vectors in SVM

Those point which touches the boundary and very closest to boundary they are called as the support vector

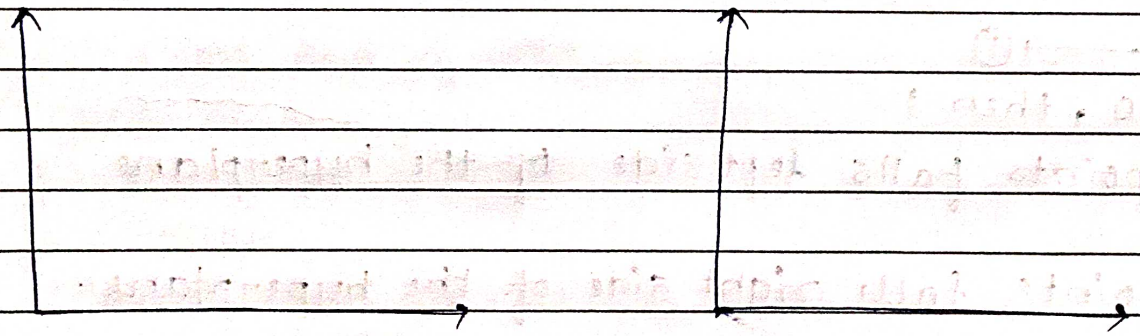
There can be multiple decision line to seggregate the classes in n dimensional p but we need to find best decision boundary that helps to classify the datapoints.



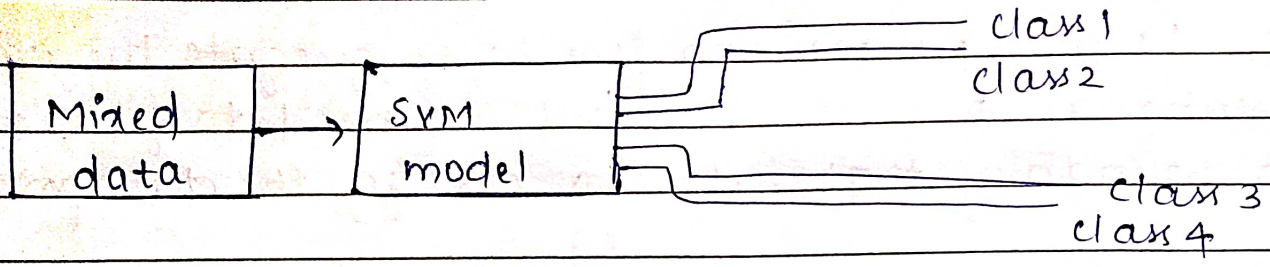
Margin :-

It may be defined as the gap between two lines on the closest data point of the different classes. It may be positive hyperplane and negative hyperplane.

It can be calculated as the perpendicular distance from the line to support vectors.



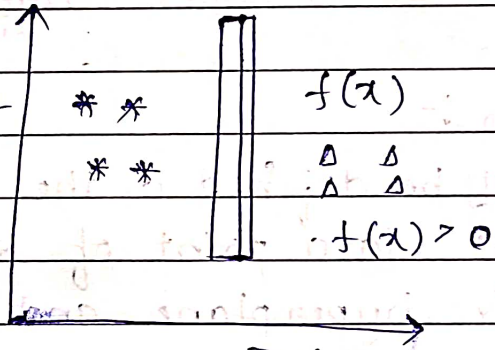
Large margin is considered as a good margin and a good hyperplane similarly small margin is considered as a bad margin not and not a accurate hyperplane between the feature.



1) SVM iteratively generate the hyperplanes that segregates the classes in the best way.
2) It provides the hyperplane that separates the classes correctly.

* How does it work?

1) We will use the thumb rule to identify the right hyperplane.
2) Select the hyperplane which segregates the two classes period better.
3) Maximizing the distance between nearest data points and hyperplane that distance is called as Margin.



$f(x)$ → SVM objective function

> 0

< 0

$f(x) = 0$ → (1)

if $f(x) < 0$, then:

data points falls left side of the hyperplane.

else:

datapoints falls right side of the hyperplane.

$f(x) = wx + b$ → (2)

[∵ Consider the 2-dimensional data points and the equation is $wx + b$]

where w is normal to the line
x is Input vector
b is known as constant (Weight Vector)

* Advantages of SVM :-
- It works better if the number of features are large
Ex - Spam filtering, disease prediction, crop recommendation etc.
- SVMs are conceptually easy to understand.

* Disadvantages of SVM :-
- It has the two major constraints
- It works only with real numbers
- All the data points in all the dimensions define numeric value only.
- It works only binary class classification

- Training the SVM is an efficient and time consuming process when the data is large.
- When we have a noise in the data set then it does not work well.

* Application Area of SVM :-
→ Regression Analysis
→ pattern Recognization
→ prediction
→ Mail is spam or not.

* Optimization Objective

→ SVM allows to minimize the classification error while maximizing the margin.

M T W T F S S

Page No.:

Date:

- SVM allows to minimize the classification error while maximizing the margin.
- The optimization objective function of SVM can be defined as

$$\text{minimize } \frac{1}{2}\|w\|^2 \longrightarrow (3)$$

W is the weight Vector, Subjective constraint that

$$y_i d_i \longrightarrow (4)$$

We know that $d_i = wx_i + b \longrightarrow (5)$

Substituting eqn (5) in (4)

$$y_i(wx_i + b) \longrightarrow (6)$$

$y_i(wx_i + b) \geq 1$ for all training examples of (x_i, y_i)

* Decision function / Boundary :-

- The Decision function for classifying the new data point x is given by

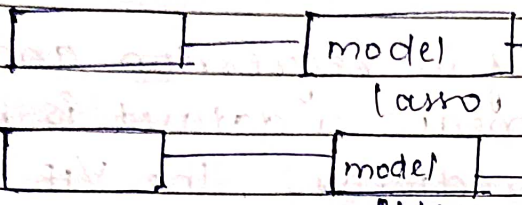
$$f(x) = wx + b$$

If $f(x) \geq 0$ the data point classify one class and if $f(x) < 0$ the data point classify other class.

Page No.:
Date:

* Elasticnet Regression

The main objective of the elasticnet regression is to improve the performance of the model Individually, the model perform well but to make the robust model we need to develop the hybridizations.



The term Ridge regression can be defined as

$$\text{Ridge} = \text{Loss} + \text{Penality}.$$
$$= \text{Loss} + \lambda \|w\|^2$$

$\|w\|^2$ = square of magnitudes.

The objective of ridge regression is to handle the regularization. Similarly, Lasso is used for feature selection because some of the features are eliminated those dataset is having multi-colinearity affect we will use lasso regression combining these two regression a new regression is formed which is called as elasticnet regression.

$$\text{Rigde} = \text{Loss} + \|w\|^2 \longrightarrow (1)$$
$$\text{Lasso} = \text{Loss} + \lambda \|w\| \longrightarrow (2)$$

Combining these two equations we get

$$\text{ENR} = \text{Loss} + \lambda_1 \|w\|^2 + \lambda_2 \|w\|$$

So the cost function of ENR is

It is called as hybrid regularization.

Page No.:
Date:

* What is VIF - Variation inflaction factor ?

$$Vif = \frac{1}{1-R^2}$$

It detects multicolinearity in regression analysis The term multicolinearity means a dataset is having correlation between the predictors. The VIF estimates how much variation of the regression coefficient is inflacted/ suffered due to multicolinearity in the models.

$$\frac{1}{1-R^2} = \frac{1}{\text{Tolerance}}$$

$$1-R^2 = \text{Tolerance}$$

Where,

R^2 = Unadjusted coefficient of determination for i^{th} independent of variance.

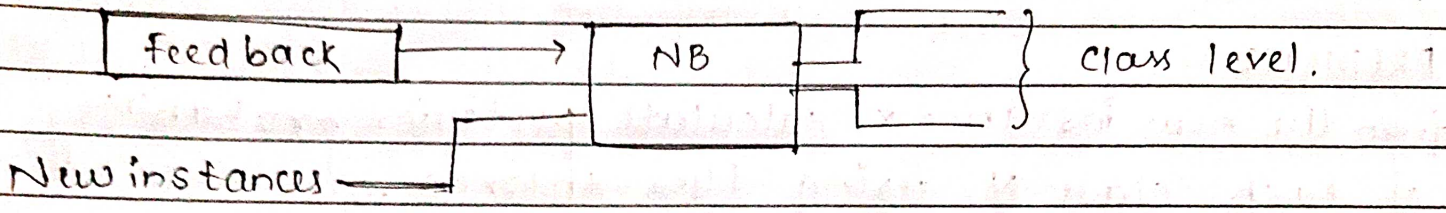
Reciprocal of VIF is called as Tolerance.

If VIF = 1 No corelated
If VIF = 5 Moderately co-related
If VIF > 5 Highly corelated.

Date:

NB (Naive Byes) classifier :-

It is a kind of supervised learning which main objective is for classification. It allow to predict the class levels.



St.	FB	Result
1		1
2		1
3		1
4		0
5		0
6		

↑ new instances

The model is trained with a training data of 5 samples when new instance comes the model has to credit the class level. It is based on the byes theorem with the assmption that features are conditionally independent given in the class level.

- Steps of the Naive byes classifier

1. Data preparation.
   Split the dataset into two parts training and testing let x and y are two features
   x → independent.
   Y → dependent.

M T W T F S S

Page No.:

Date:

2. Calculate the prior probability $P(Y)$ for each class in the training examples.

- For each feature X_i and each class Y calculate the likelihood, $P(X_i|Y)$ which is the probability distribution of feature X_i for given class Y.

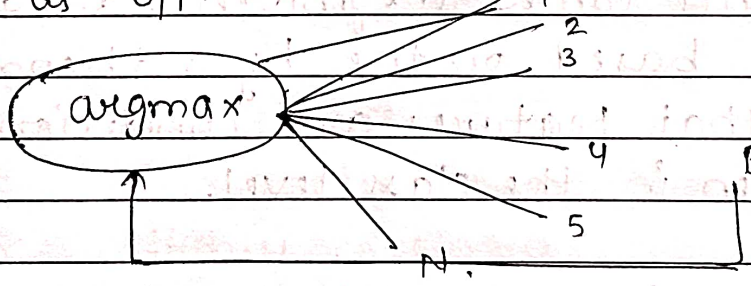
PREDICTIONS

Given the new instance x calculate posterior probability of each class Y using Byes theorem.

- $$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$ → Bayes theorem formula

- Since $P(X)$ is constant for all classes it can be ignored during comparision.

- The class with the highest posterior probabity is predicted as O/P.



optimization posterior probability.

- It can be represented as

$$\hat{y} = \text{argmax}\ P(Y|X)$$

M T W T F S S

Page No.:

Date:

4. Evaluation

Evaluate the performance of the ~~classification~~ classifier using appropriate matrix such as accuracy, precision, Recall, F1 Score, Decision etc.

* BAYE'S THEOREM

It is a fundamental concept in the probability theory it describe the probability of prior knowledge that might be related to the event.

Mathematically,

The Baye's theorem is represented as

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

$P(Y|X)$: posterior probability.

$P(X|Y)$: likelyhood

$P(X)$: ~~probability of~~ prior probability of the predictor X.

$P(Y)$: Prior probability of class Y.

* Applications of Naive Byes.

(i) Sentiment Analysis

(ii) Disease prediction

(iii) Crop recommendation

(iv) Emotional detection

(v) protein Analysis

UNIT-3

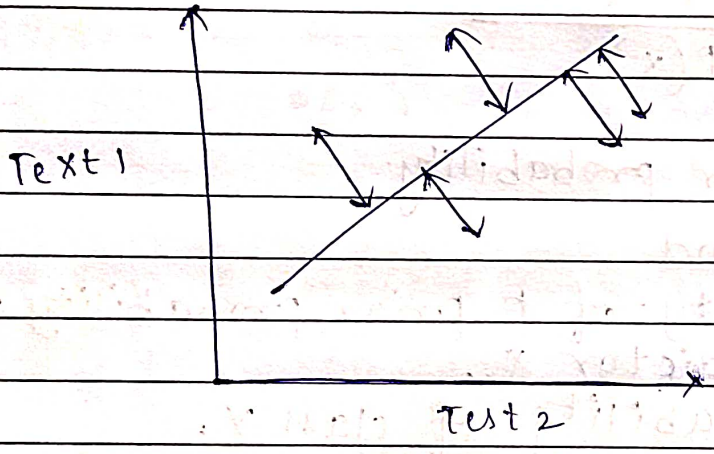
GRADIENT DESCENT

M T W T F S S

Page No.:

Date:

- It is a kind of optimization technique which is used to minimize the loss function of machine learning algorithm.
- In mathematical terminology optimization algorithm refers to the task of minimizing or maximizing an objective function $f(x)$.
- Similarly in ML, optimization is the task of minimizing the cost function.
- The main objective of Gradient descent is to minimize the loss or cost function of the Machine learning classifier.

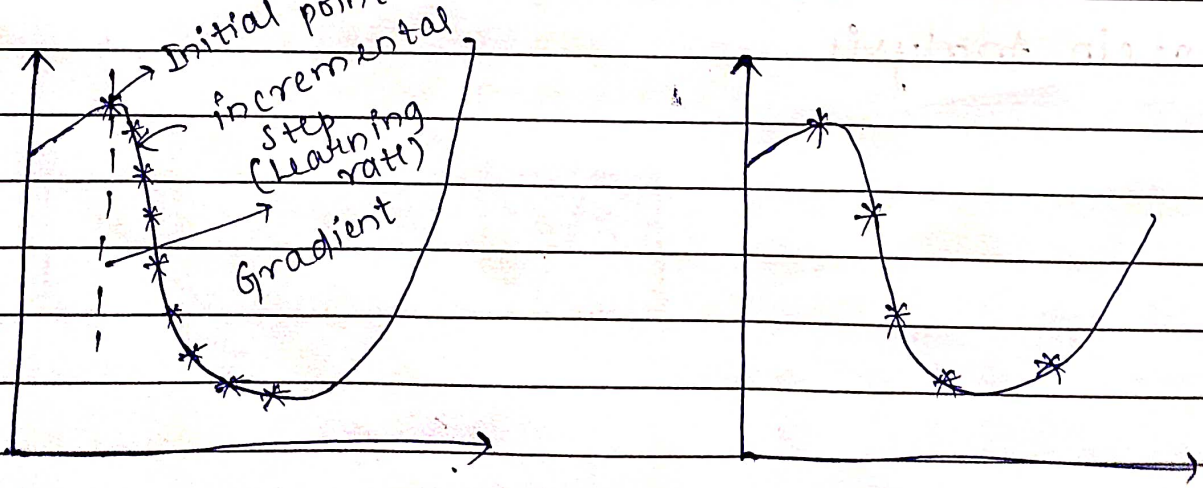


Line of best fit / Regression line..

Error = Actual - predicted

$$\sum_{i=0}^{n} (y_i - \bar{y})^2$$

$$x := x - \alpha \frac{\partial L}{\partial w}$$



Page No.:
Date:

- Gradient descent is defined as one of the most commonly used iterative optimization algorithm of machine learning to train the machine learning and deep learning model

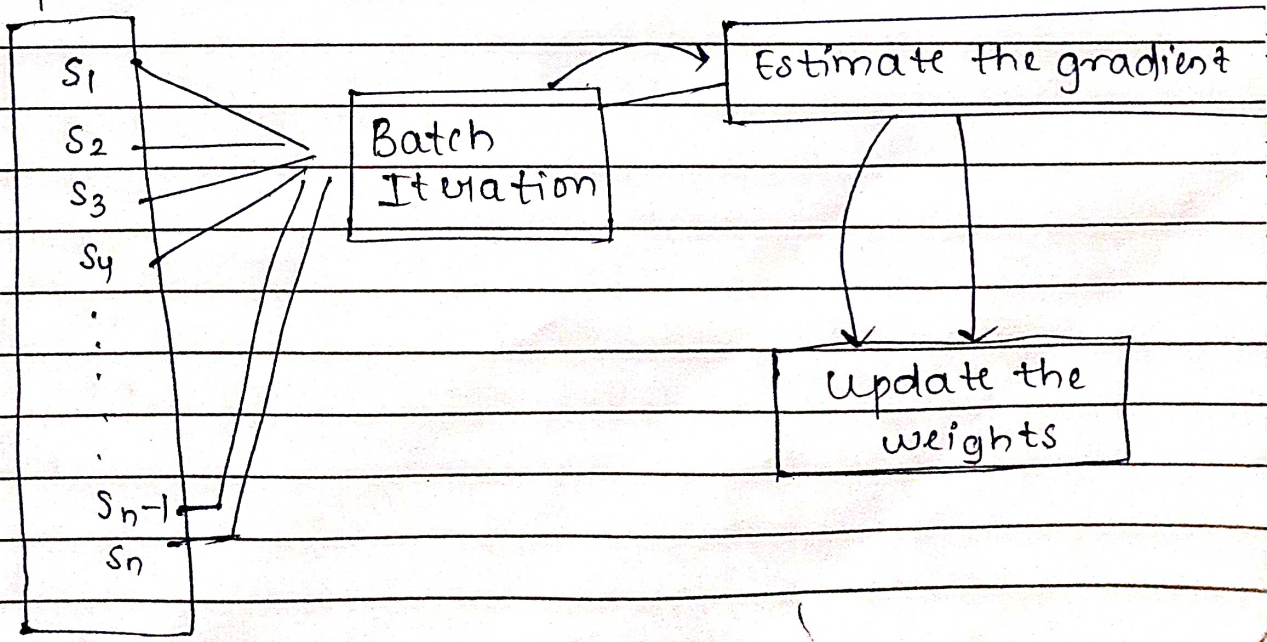
- It helps to find local minima of a function.
- If we move towards a negative gradient of the function at the current point it will give a local minima.
- If we move towards a positive direction of the function at the current point it will get a local maxima of that function.

Gradient estimation
weights update.

* Types of Gradient Descent :-

(*) There are 3 types of Gradient Descent.
(i) Batch Gradient
(ii) Mini Batch Gradient
(iii) Stochastic Gradient.

(i) BATCH GRADIENT



M T W T F S S

Page No.:

Date:

- In Batch Gradient Descent we use all the training samples to prepare a single batch and pass into the network.
- It will compute the gradient loss function for each sample.
- Then update the weights.
- If the training sample size is more than it will take much time to compute the gradient otherwise it is the best approach.
- A hypothesis of the gradient can be defined

$$h_w(x) = w_0 + w_1 x$$

- The parameters that need to be updated are w_0 and w_1 because they are unknown.
- The cost function can be defined as $J(w_0, w_1)$ in term of gradient descent it will be

$$= \frac{1}{2m} \sum_{i=1}^{m} \left(h_w(x_i) - y_i \right)^2$$

(11)

M T W T F S S

Page No.:

Date:

MINI BATCH GRADIENT

In mini batch gradient some of the training samples created a Iteration which is called as a batch. Each batch further responsible to estimate the gradient.

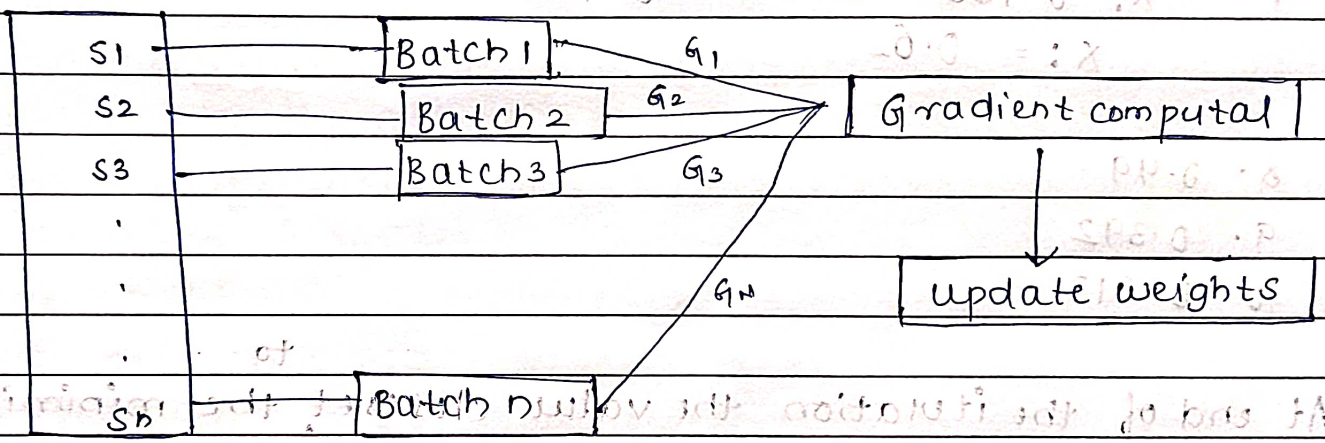
The last step of mini batch gradient is to compute the weights.

It updates the parameters using the small random subset of training data. It reduces the computational burden compared to batch gradient descent.

It combines the benefits of batch gradientation and Stockastic gradientation descent.

* STOCHASTIC GRADIENT

In this gradientation each training sample prepared a batch and each batch is responsible to find a gradient.



$$x := x - \alpha * \frac{\partial f}{\partial x}$$

$x := 3$

$\alpha = 0.1$

$f(x) = x^2$

$\frac{\partial f}{\partial x} = 2x$

	old	New
	3	2.4
	2.4	1.92
	1.92	1.536
	1.536	1.228
	1.228	0.976

1. $x := 3 - 0.1 \times 2x$

$3 - 0.1 \times 2.3$

$x := 2.4$

2. $x := 2.4 - 0.1 \times 2 \times 2.4$

$x = 1.92$

Date:

3. $X := X - \alpha \times 2x$

$1.92 - 0.1 * 2 \times 1.92$

$X = 1.536$

4. $X := 1.536 - 0.1 \times 2 * 1.536$

$X := 1.228$

5. $X := 1.228 - 0.1 \times 2 \times 1.228$

$X := 0.976.$

6. $X := 0.976 - 0.1 \times 2 \times 0.976$

$X := 0.786$

7. $X := 0.786 - 0.1 \times 2 \times 0.786$

$X := 0.62$

8. 0.49

9. 0.392

10. 0.313

At end of the iteration the value closeset to the minimization function $f(x) = 0$.

The entire procedurce will terminate either the function will converge or the maximum iterations reached.

* Linear regression in terms of gradientation descent approach.

In linear regression, our objective was to find the best fitted line.

Algorithm - Gradient descent for linear regression.

M T W T F S S

Page No.:

Date:

1. Start with the initial value for the slope m intercept b.

$y = mx + b$

m = slope m : 0

b = intercept & b : 0

2. Cost function

choose a cost function to measure the error between actual and predicted

- for example in linear regression the mean square error is commonly used

$$MSE = \frac{1}{N} * \Sigma (y_i - (mx+b))^2$$

where N is the number of observations.

3. Gradient calculation

perform the iterations to update the parameters based on gradient.

the gradient indicates the direction of steepest ascent so we subtract a fraction of it from the parameters to descent towards the minimum.

Repeat until conversions or maximum number of iterations.

4. Partial derivative of the MSE

the partial derivative of the MSE with respect to slope(m) is $\frac{\partial MSE}{\partial m}$

$$-\left(\frac{2}{n}\right) * \Sigma x(y - mx+b)$$

The partial derivative of MSE with respect to MSE

$$\frac{\partial MSE}{\partial b} = -\left(\frac{2}{n}\right) * \Sigma (y - (mx+b))$$

This equation represents the rate of change of MSE with respect to change in the intercept. It accounts for the contribution of each datapoint to the error without multiplication by the input feature since the intercept does not directly depend on x.

5. Prediction
after convergence use the learning parameter to make predictions. for new input x predict y using $y = mx + b$.

~~If it consider~~

* Gradient descent with logistic regression.

(i) Initialize the parameters
weight : 0
intercept ~~bias~~ : 0

(ii) Define the sigmoid functions

$$f(x) = \frac{1}{1+e^{-z}}$$

Sigmoid z ($z = w_1x_1 + w_2x_2 \cdots + w_nx_n$) ~~$\beta_n x_n$~~ ~~$z = \beta_0 + \beta_1 x_1 \cdots$~~

(iii) Compute the predictions
for each data points

$$z = w_1x_1 + w_2x_2 + \cdots w_nx_n + b$$

$y\_pred = sigmoid(z)$

M T W T F S S

Page No.:

Date:

(iv) Compute gradients

$$\frac{\partial cost}{\partial w_1} \quad \frac{\partial cost}{\partial w_2} \quad \frac{\partial cost}{\partial b}$$

We need to find the gradient of the costfunction (binary cross entropy) with respect to parameters weights using the following formula.

$$\frac{\partial cost}{\partial w_1} = \left(\left(\frac{1}{m}\right) * \Sigma(y_{pred} - y) * x_i\right)$$

for $i = 1$ to N

$$\frac{\partial cost}{\partial b} = \left(\left(\frac{1}{m}\right) * \Sigma(y_{pred} - y)\right)$$

Here m is the number of data points

(v) update the parameters using the gradients and learning rate α

1. $w_{i\text{-}new} = w_i - \alpha * \frac{\partial cost}{\partial w_i}$ for $i = 1$ to N

2. $b\text{-}new = b - \alpha * \frac{\partial cost}{\partial b}$

(vi) Repeat iterations.
Repeat steps 3-5 until convergence or stopping criteria met.

(vii) Convergence can be determined based on changes in the cost function or parameter value.

Date:

- Compute the gradients

The gradient of the cost function with respect to the parameters weight and intercept can be written as

$$\frac{\partial cost}{\partial w_i} = \frac{1}{4}\left(\sum_{i=1}^{4}(\hat{y}_i - y_i)x_i\right)$$

$$\frac{\partial cost}{\partial b} = \frac{1}{4}\left(\sum_{i=1}^{4}(\hat{y}_i - y_i)\right)$$

The cost function / binary cross entropy function of logistic regression given by

$$cost(y, \hat{y}) = -\frac{1}{m}\sum_{i=1}^{m}\left[(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

cost function / cross entropy

Q. Let us take a dataset having x and y. and their values are given $x = [1, 2]$ $y = [2, 4]$ Apply gradient descent technique in linear regression cost function and find the optimal parameter and update it.

① Initialize : $b = 0$ // Intercept
$m = 0$ // slope
$\alpha = 0.1$ // step size

② calculate the predicted value :

$$y\text{-predicted} = [mx + b \text{ for } i : 1 \text{ to } N]$$

$$y\text{-pred} = [0 \times 1 + 0, \; 0 \times 2 + 0]$$

$$Y\text{-pred} = [0, 0]$$

Page No.:
Date:

(3) Compute the cost function.

$$MSE = \left(\frac{1}{2}\right) * \sum (y - y\_predicted)^2$$

$$= \frac{1}{2} * \left[(2-0)^2 + (4-0)^2\right]$$

$$= \frac{1}{2} * (4+16)$$

$$= 10$$

(4) Calculate the gradients
In this step we required to calculate the partial derivate of the cost function of linear regression with required to slope is

$$\frac{\partial MSE}{\partial m} = -\left(\frac{2}{n}\right) * \sum x (y - (mx+b))$$

$$= -\left(\frac{2}{n}\right) * \left[1 * (2 - (0*10) + 2 * (4 - (0*2+0))\right]$$

$$= -\left[1 * (2-0) * 2 * (4-0)\right]$$

$$= -\left[1*2+8\right]$$

$$\frac{\partial MSE}{\partial M} = -10.$$

(5) $$\frac{\partial MSE}{\partial b} = \left(\frac{2}{n}\right) * \sum (y - (mx+b))$$

$$= \left(\frac{-2}{2}\right) * \left[2 - (0*1+0) + 4 - (0*2+0))\right]$$

$= -6$

(6) $x := x - \alpha * \frac{\partial MSE}{\partial b}$

Calculate the updated parameter.

$$m\_new = m - \alpha * \frac{\partial MSE}{\partial m}$$

$$0 - 0.1 * (-10)$$

$$m\_new = 1$$

$$b\_new = b - \alpha * \frac{\partial MSE}{\partial b}$$

$$= 0 - 0.1 * -6 = 0.6.$$

At the end of the first iteration m value changes to 1 and b value changes to 0.6

$$y = mx_1 + b$$

$$y = 1(1) + 0.6$$

$$y_1 = 1.6$$

$$y = mx_2 + b$$

$$y_2 = 1(2) + 0.6$$

$$y_2 = 2.6.$$

(2) $$MSE = \frac{1}{N} * \sum (y_i - (mx + b)^2)$$

$$= \frac{1}{2} * \left[(2 - 1.6)^2 + (4 - 1.6)^2 \right] = 1.06$$

Page No.:
Date:

• Calculate the gradients

$$\frac{\partial MSE}{\partial m} = -\left(\frac{2}{n}\right) * \sum x\,(y-(mx+b))$$

$$= \left(\frac{-2}{2}\right) * \left[1\,(2-(1\times1+0.6)) + 2 * (4-(1\times2+0.6))\right]$$

$$\frac{\partial MSE}{\partial m} = -3.2$$

$$\frac{\partial MSE}{\partial b} = \frac{-2}{n} * \sum (y-(mx+b))$$

$$= \left(\frac{-2}{2}\right) * \left[(2-(1\times1+0.6)) + 4-(1\times2+0.6))\right]$$

$$\frac{\partial MSE}{\partial b} = -1.8.$$

• update the gradients.

$$m_1\text{-new} = m - \alpha * \frac{\partial MSE}{\partial m}$$

$$= 1 - 0.1 * (-3.2) = 1.32$$

$$b\text{-new} = b - \alpha * \frac{\partial MSE}{\partial b}$$

$$= 0.6 - 0.1 * (-1.8) = 0.78.$$

We can continue this process for more iterations until parameters converged to their optimal values or until a predefined number of iterations is reached.

Q.2. $x = [1,2,3,4,5]$ $y = [2,4,5,4,5]$ $\alpha = 0.01$

① Initialize.

$m = 0$

$b = 0$

$\alpha = 0.01$

② Calculate the predicted value.

$y\text{-pred} = mx+b$

$= [0*1+0, \ldots\ldots\ldots, 0*5+0]$

$= [0,0,0,0,0]$

③ Cost function

$MSE = \frac{1}{5}\Sigma(y - y\_pred)^2$

$= \frac{1}{5}[(2)^2 + (4)^2 + (5)^2 + (4)^2 + (5)^2]$

$= 17.2$

④ Calculate $\frac{\partial MSE}{\partial m}$ and $\frac{\partial MSE}{\partial b}$

$$\frac{\partial MSE}{\partial M} = \frac{-2}{n} * \Sigma x(y-(mx+b))$$

$$= \frac{-2}{5} * [1(2-0) + 2*(4-0) + 3(5) + 4(4) + 5(5)]$$

$$= \frac{-2}{5} \times 66$$

$$\frac{\partial MSE}{\partial M} = -26.4$$

$$\frac{\partial MSE}{\partial b} = \frac{-2}{n} * \Sigma(y-(mx+b))$$

$$= \frac{-2}{5} \times 20$$

$$\frac{\partial MSE}{\partial b} = -8$$

$m\_new = m - \alpha \times \frac{\partial MSE}{\partial m}$

$= 0 - 0.01 \times (-26.4)$

$= 0.264$

$b\_new = b - \alpha \frac{\partial MSE}{\partial b}$

$= 0 - 0.01 \times (-8)$

$= 0.08.$

2nd iterations.

1. Initialize $m = 0.264$ $b = 0.08$ $\alpha = 0.01$

(2) $Y\text{-pred} = mx + b$

$= [(0.264 \times 1 + 0.08), (0.264 \times 2 + 0.08), (0.264 \times 3 + 0.08), (0.264 \times 4 + 0.08), (0.264 \times 5 + 0.08)]$

$= [0.344, 0.608, 0.872, 1.136, 1.4]$

(3) $MSE = \frac{1}{5} * \Sigma[(y - y\text{-pred})^2]$

$= \frac{1}{5} * [(2-0.344)^2 + (4-0.608)^2 + (5-0.872)^2 + (4-1.136)^2 + (5-1.4)^2]$

$MSE = 10.488.$

(4) Calculate the gradients.

$\frac{\partial MSE}{\partial m} = \left(\frac{-2}{n}\right) * \Sigma x(y - (mx+b))$

$= \left(\frac{-2}{5}\right) * [1(2 - (0.264 \times 1 + 0.08)) + \cdots + 5*(5 - (0.264 \times 5 + 0.08))]$

$\frac{\partial MSE}{\partial m} = -20.166$

$\frac{\partial MSE}{\partial b} = \frac{-2}{n} * \Sigma(y - (mx+b))$

$\frac{\partial MSE}{\partial b} = -6.25$

Page No.:
Date:

⑤ update the gradients

$$m\_new1 = m - \alpha * \frac{\partial MSE}{\partial m}$$

$$= 0.264 - 0.01 \times (-20.106)$$

$$= 0.465$$

$$b\_new = b - \alpha * \frac{\partial MSE}{\partial b}$$

$$= 0.08 - 0.01 * (-6.25)$$

$$= 0.14$$

We can continue this process for more iterations until parameters converged to their optimal values or until a predefined number of iterations is reached.

* MINI BATCH GRADIENT.

Let us consider the dataset contains one feature and target variable apply mini batch gradient descent to optimization linear regression cost functions.

Q.1 X : [1, 2, 3, 4, 5] // features
Y : [2, 4, 6, 8, 10] // Target

steps to calculate the mini Batch

① Initilization
$y = mx + b$
$m = 0$
$b = 0$

② create a mini batch (For one mini batch)
X - batch : [1, 2]

Y - batch : [2, 4]

Page No.:
Date:

(3) Calculation/ compute the prediction using correct parameters.

$Y\_pred = wx + b$

$= w \cdot x\_batch + b$

(4) Calculate the gradients of the cost function with respect to weights and bias

$$\frac{\partial MSE}{\partial w} = \frac{-2}{n} \sum_{i=1}^{N} (y_i - (wx_i + b) * x_i)$$

$$\frac{\partial MSE}{\partial m} = \frac{-2}{n} \sum_{i=1}^{N} (y_i - (wx_i + b))$$

To calculate ∂w we have :-

$$\partial w = \left(\frac{-2}{N}\right) * sum((y\_batch(-w * x\_batch + b)) * x\_batch)$$

$$\partial b = \left(\frac{-2}{N}\right) * sum(y\_batch(-w * x\_batch + b))$$

(5) update the parameters

$w\_new\ i = w - \alpha * \partial w$

$b\_new\ i = b - \alpha * \partial b$

$\boxed{\alpha = 0.01}$

(6) Repeat steps 2-4 until the convergence.

Q. X: [1,2,3,4,5] Y: [2,4,6,8,10]

$\alpha = 0.01$

(1) Initialization

$w = 0$

$b = 0$

$x\_batch = [1,2]$ $y\_batch = [2,4]$

Page No.:
Date:

② $y\_pred = w * x\_batch + b$

$y\_pred = 0 * 1 + 0 \;,\; 0 * 2 + 0$

$[0, 0]$

③ Compute.

$$\frac{\partial MSE}{\partial w} = \frac{-2}{N} \sum_{i=1}^{N} (y_i - (wx_i + b)) * x_i$$

$$= \frac{-2}{2} [1 * 2$$

$$= -10$$

$$\frac{\partial MSE}{\partial b} = \frac{-2}{N} \sum_{i=1}^{N} (y_i - wx_i + b)$$

$$= \frac{-2}{2} \sum_{i=1}^{N} [(2 - 0) + (4 - 0)]$$

$$\frac{\partial MSE}{\partial b} = -6$$

$w\_new = w - \alpha * \partial w$

$= 0.1$

$b\_new = b - \alpha * \partial b = 0.06$

Select the mini batch 2

$x : [3, 4]$ and $w\_new = 0.1$

$y : [6, 8]$ $b\_new = 0.06$

§ ① Intialize.

$w = 0.1$

$b = 0.06$

$x = [3, 4]$ $y = [6, 8]$

② $y\_pred = wx + b$

$= [0.1 * 3 + 0.06 \;,\; 0.1 * 4 + 0.06]$

$= [0.36, 0.46]$

③ $\frac{\partial MSE}{\partial w} = \frac{-2}{N} \sum_{i=1}^{N} (y_i - (wx_i + b) * x_i)$

$\frac{\partial MSE}{\partial w} = -47.08.$

$\frac{\partial MSE}{\partial b} = \frac{-2}{N} \sum_{i=1}^{N} (y_i - wx_i + b)$

$\frac{\partial MSE}{\partial b} = -13.18.$

$w\_new = 0.1 - 0.01 * -47.08 = 0.5768.$

$b\_new = 0.06 - 0.01 * -13.18 = 0.1918.$

Page No.:
Date:

* Multiple Linear Regression with Gradient Descent Approach :-

- It involves predicting the target variables based on multiple input features.
- The main objective of this algorithm is to update the multiple parameters corresponding to each input features along with the intercept.
- Let us consider a dataset having two features along with the target variables our objective is to find multiple linear regression model to find the optimal values of the coefficient m_1 and m_2 and the intercept of linear regression.
- The equation we will use for multiple linear regression is

$$y = m_1x_1 + m_2x_2 + b \longrightarrow (1)$$

Algorithm

(1) Initialize the parameters.

Lets proceed the Gradient descent algorithm with initialization parameter.

Set

$m_1 = 0$

$m_2 = 0$

$b = 0$

$\alpha = 0.01$

(2) Find the cost function

The cost function of Multiple Linear regression is termed as MSE.

$$MSE = \left(\frac{1}{N}\right) * \sum \left(y - \left(m_1x_1 + m_2x_2 + b\right)^2\right)^2$$

Date:

③ Gradient Descent Iteration

perform the iterations to update the parameters m_1, m_2, b using the Gradient descent of the cost function with respect to each parameter.

(i) $\frac{\partial MSE}{\partial m_1}$ (ii) $\frac{\partial MSE}{\partial m_2}$ (iii) $\frac{\partial MSE}{\partial b}$

④ Update the parameters

The parameters will be updated by using the formula

$X := X - \alpha * \text{cost function}$.

⑤ prediction

After convergence we the learning parameters to make prediction on new data.

predicted valu.

$y\text{-pred} = m_1 x_1 + m_2 x_2 + b$.

Q. $x_1 = [1, 2]$

$x_2 = [2, 3]$

$y = [3, 4]$

① $m_1 = 0$ $b = 0$

$m_2 = 0$ $\alpha = 0.01$

② $y\text{-pred} = [0, 0]$

③ $MSE = \left(\frac{1}{N}\right) * \Sigma \left(y - (m_1 x_1 + m_2 x_2 + b)\right)^2$

$= \frac{1}{2} * \left[(3-0)^2 + (4-0)^2\right]$

$MSE = 12.5$

④ $\frac{\partial MSE}{\partial M_1} = \left(\frac{-2}{N}\right) * \Sigma \left(x_1 \left(y - (m_1 x_1 + m_2 x_2 + b)\right)\right)$

$$\frac{\partial MSE}{\partial m_2} = \frac{-2}{N} * \Sigma\left(x_2\left(y - (m_1x_1 + m_2x_2 + b)\right)\right)$$

$$\frac{\partial MSE}{\partial b} = \frac{-2}{N} * \Sigma\left(y - (m_1x_1 + m_2x_2 + b)\right)$$

$$\frac{\partial MSE}{\partial m_1} = \frac{-2}{2} * \left[(1 * (3-0)) + (2 * (4-0))\right]$$

$$= -11$$

$$\frac{\partial MSE}{\partial m_2} = \frac{-2}{2}\left[(2 * (3-0)) + (3 * (4-0))\right]$$

$$= -18$$

$$\frac{\partial MSE}{\partial b} = \frac{-2}{2}(3+4)$$

$$= -7.$$

$$m_1\text{-new} = m_1 - \alpha \frac{\partial MSE}{\partial m_1}$$

$$= 0 - 0.01 \times -11$$

$$= 0.11$$

$$m_2\text{-new} = m_2 - \alpha \frac{\partial MSE}{\partial m_2}$$

$$= 0 - 0.01 \times -18$$

$$= 0.18.$$

$$b\text{\_new} = b - \alpha \frac{\partial MSE}{\partial b}$$

$$= 0 - 0.01 \times (-7)$$

$$= 0.07$$

The new updated parameters are.

① Initialize $m_1 = 0.11$ $b = 0.07$

$m_2 = 0.18$

Page No.:
Date:

(2) $y\text{-pred} = m_1 x_1 + m_2 x_2 + b$

$= [(0.11 \times 1 + 0.18 \times 2 + 0.07), (0.11 \times 2 + 0.18 \times 3 + 0.07)]$

$= [0.54, 0.83]$

(3) $MSE = \frac{1}{N} * \Sigma (y - (m_1 x_1 + m_2 x + b))^2$

$= \frac{1}{2} * ((3 - 0.54)^2 + (4 - 0.83)^2)$

$= \frac{1}{2} (6.05 + 10.04) = 8.045$

(4) $\frac{\partial MSE}{\partial m_1} = \frac{-2}{N} * \Sigma x_1 (y - (m_1 x_1 + m_2 x_2 + b))$

$= \frac{-2}{2} * (1 * (3 - 0.54) + 2 * (4 - 0.83))$

$= -1 * 8.8$

$= -8.8$

$\frac{\partial MSE}{\partial m_2} = \frac{-2}{N} * \Sigma x_2 (y - (m_1 x_1 + m_2 x_2 + b))$

$= \frac{-2}{2} [2 * (3 - 0.54) + 3 * (4 - 0.83)]$

$= -14.43$

$\frac{\partial MSE}{\partial b} = \frac{-2}{N} * [y - (m_1 x_1 + m_2 x_2 + b)]$

$= \frac{-2}{2} * [(3 - 0.54) + (4 - 0.83)]$

$= -5.63$

$m_1\text{-new} = m_1 - \alpha \cdot \frac{\partial MSE}{\partial m_1}$

$= 0.11 - 0.01 \times (-8.8) = 0.19$

Date:

$$m_2\text{-new} = m_2 - \alpha \cdot \frac{\partial MSE}{\partial m_2}$$

$$= 0.18 - 0.01 \times (-14.43)$$

$$= 0.32$$

$$b\text{\_new} = b - \alpha \cdot \frac{\partial MSE}{\partial b}$$

$$= 0.07 - 0.01 \times (-5.63)$$

$$= 0.12$$

* Non-parametric Test :-

- Spearman correlation coefficient Rank :-

It is a non-parametric test which is used to calculate the value of $r(r)$.

The spearman Rank corelation can be written by the formula

$$\gamma_R = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}$$

where N is the number of observations

D is the difference b/w Rank x and Rank y.

The values of γ_R lies b'etween $[-1, 1]$

If $\gamma_R = 1 \longrightarrow$ Perfect positive corelation

$\gamma_R = -1 \longrightarrow$ perfect Negative correlation

$\gamma_R = 0. \longrightarrow$ No Correlation

If $\gamma_R = 0.08 \rightarrow$ Moderate correlation

Date:

Q. Let us take a dataset. Apply spearman Rank correlation coefficient. Interpret the results.

x	y	R_x	R_y	D	D^2
10	7	3.5	3.5	0	0
8	4	5.5	6	-0.5	0.25
12	6	2	5	-3	9
15	7	1	3.5	-2.5	6.25
8	9	5.5	1	4.5	20.25
10	8	3.5	2	1.5	2.25

$D = R_x - R_y$

$$r_R = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$1 - \frac{6 \times 38}{6(36-1)}$$

$$= -0.085.$$

* Wilicoxon- Rank sum Test :-

- Wilicoxon Rank sum Test also known as Mann-Whitney (U) test
- It is a non-parametric test used to access whether two independent samples come from the same distribution.
- The main objective of this test is to compare the medians of two independent groups, it is also called as alternative to two-sample t-test.

① Formulate the Hypothesis

Define the null and alternative hypothesis

H_0 : Null hypothesis

H_1 : Alternative hypothesis

The Null hypothesis tells that there is no difference b/w distribution of two groups.

Date:

The Alternative hypothesis tells that there is a difference between distribution of two groups.
The Alternative hypothesis tells that there is a difference between distribution of two groups.

(2) Rank the Data
Combine the data from both the groups.
Rank them in the ascending order.
Assign the ranks with the smallest value.
If there is same value then assign the average rank.

(3) Calculate the sum of Rank
Calculate the sum of Rank x (Tx)
Calculate the sum of Rank y (Ty)

(4) Calculate the statistics U prior to this Calculate

$$U_x = T_x - \frac{n_1(n_1+1)}{2}$$

$$U_y = T_y - \frac{n_2(n_2+1)}{2}$$

$$U = \min(U_x, U_y)$$

(5) Determine the critical value
Look the wilcoxon rank sum table where the critical value depends on sample sizes and level of significance.
For example - $n_1 = n_2 = 3$

$\alpha = 0.01$

critical value = 9

$U = 7.8$

M T W T F S S
Page No.:
Date:

⑥ Compare the test statistics to the normal value
If U is less than or equal to critical value than reject the Null hypothesis else accept the alternative hypothesis.

⑦ Interpretation
If the Null hypothesis is rejected then it suggests that there is a significance between distribution of two groups.
If the Null hypothesis is not rejected then it indicates that there is not enough evidence to conclude a siognificance difference.

Q. Let us consider the two different groups and their values are available apply wilcoxon Rank sum test and interpret it.

x	60		45		23		32				
y	10		25		20		54		32	65	8.

① Hypothesis
$n_1 = 4$
$n_2 = 7$
H_0 : Both are not same
H_1 : Both are same

Level	Y	Y	Y	X	Y	X	Y	X	Y	X	Y
Data	8	10	20	23	25	32	32	45	54	60	65
Rank	1	2	3	4	5	6.5	6.5	8	9	10	11

$T_x = 4 + 6.5 + 8 + 10 = 28.5$
$T_y = 1 + 2 + 3 + 5 + 6.5 + 9 + 11 = 37.5$

$U_x = 28.5 - \frac{4(5)}{2}$ $\quad U_y = 37.5 - \frac{7(8)}{2}$

$U_x = 18.5$ $\quad U_y = 9.5.$

Page No.:
Date:

Cross Verification.

Test

1) $U_x + U_y = n_1 n_2$

$18.5 + 9.5 = 4 \times 7$

$28 = 28$

2) $T_x + T_y = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$

$28.5 + 37.5 = \frac{(4+7)(4+7+1)}{2}$

$66 = \frac{11 \times 12}{2}$

$66 = 66$

$U_{min} = (U_x, U_y)$

$U = (18.5, 9.5)$

$U_{min} = 9.5$ ✓

Q. A = 78, 82, 85, 88, 90

B = 72, 75, 80, 83, 86

Critical value = 8

$\alpha = 0.05$

Sol :- ① $n_1 = 5$

$n_2 = 5$

$\alpha = 0.05$

$n_1 = n_2$

H_0 : Both are not same

H_1 : Both are same

Level	B	B	A	B	A	B	A	B	A	A
Data	72	75	78	80	82	83	85	86	88	90
Rank	1	2	3	4	5	6	7	8	9	10

$T_x = 3+5+7+9+10 = 34$

$T_y = 1+2+4+6+8 = 21$

$$U_x = T_x - \frac{n_1(n_1+1)}{2}$$

$$= 34 - \frac{5(5+1)}{2}$$

$$= 34 - \frac{5(6)^3}{7}$$

$$= 34 - 15$$

$$\boxed{U_x = 19}$$

$$U_y = T_y - \frac{n_2(n_2+1)}{2}$$

$$= 21 - \frac{5(6)^3}{7}$$

$$= 21 - 15$$

$$\boxed{U_y = 6}$$

$U_{min}(U_x, U_y)$

$U_{min} = 6.$

$6 <= 8$

Reject the Null hypothesis

M T W T F S S

Page No.:

Date:

Mann Whitney Test for large Sample.

When the sample size is more i.e. $n_1 >= 10$ or $n_2 >= 10$ then we called it as a Mann whitney Test for large sample.

- Steps

1. Test the hypothesis
2. Compute U_x and U_y
3. compute Test statistics

where

$$z = \frac{U - U_0}{\sigma}$$

where $U = \min(U_x, U_y)$

$$U_0 = \frac{(n_1)(n_2)}{2}$$

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

4. compare the test statistics to the critical value

If compared $z >$ critical value at level of given significance,

We reject H_0 otherwise we fail to reject $[H_0]$

Q. The below data below show the salaries at randomly selected advertisement in two different occupation [education and Marketing]

education	Marketing.		
22	28	16	26
40	40	19	55
18	20	21	48
25	45	30	41
15	50		42
23	39		

Date:

Use mann whitney test at 1% level of significance at the median and check the salary in the education is lower than the median salary in the marketing.

① Test the Hypothesis

$n_1 = 10$

$n_2 = 11$

H_0 = Both are not same

H_1 = Both are same

Lable	E	E	E	E	M	E	E	E	E	M	M	E
Data	15	16	18	19	20	21	22	23	25	26	28	30
Rank	1	2	3	4	5	6	7	8	9	10	11	12

②

$E_x = 66.5$

$E_y = 165 \rightarrow 164.5 \quad \alpha = -2.33$

$$U_x = E_x - \frac{n_1(n_1+1)}{2}$$

$$= 66.5 - \frac{10(11)}{2}$$

$$= 66.5 - 55$$

$$\boxed{U_x = 11.5}$$

$$U_y = E_y - \frac{11(12)}{2}$$

$$= 165 - 66$$

$$\boxed{U_y = 99}$$

Page No.:
Date:

$U = \min(U_x, U_y)$

$\boxed{U = 11.5}$

③ $U_0 = \frac{10(11)}{2} = 55$

$z = \frac{11.5 - 55}{\sigma}$ $\qquad$ $\frac{11.5 - 55}{14.20} = -3.063.$

$= \boxed{-3.09}$

$\sigma = \sqrt{\frac{11(10)(10+11+1)}{12}} = 14.20$

M	M	E	M	M	M	M	M	M
39	40	40	41	42	45	48	50	55
13	14.5	14.5	16	17	18	19	20	21

④ The critical value at 1% level of significance is

$-3.09 < -2.33$ so we reject the

Therefore we conclude that

Q. A 42 20 51 39 57 60 23 24 28 30

B 30 42 25 29 35

① Test the Hypothesis

$n_1 = 10 \quad n_2 = 5$

Lable	A	A	A	B	A	B	A	B	B	A	A	B	A	A	A
Data	20	23	24	25	28	29	30	30	35	39	42	42	51	57	60
Rank	1	2	3	4	5	6	7	8	9	10	11	11.5	11.5	14	15

$T_x = 80 \quad T_y = 38.$

M T W T F S S

Page No.:

Date:

$$U_x = 80 - \frac{5(11)}{2}$$

$$= 80 - 55 = 25$$

$$U_y = 38 - \frac{5(6)^3}{2}$$

$$38 - 15$$

$$U_y = 23$$

$$U = \min(U_x, U_y)$$

$$\boxed{U = 23}$$

$$U_0 = \frac{5(5)}{2} = 25$$

$$\sigma = \sqrt{\frac{10(5)\ (10+5+1)}{12}}$$

$$= 8.164$$

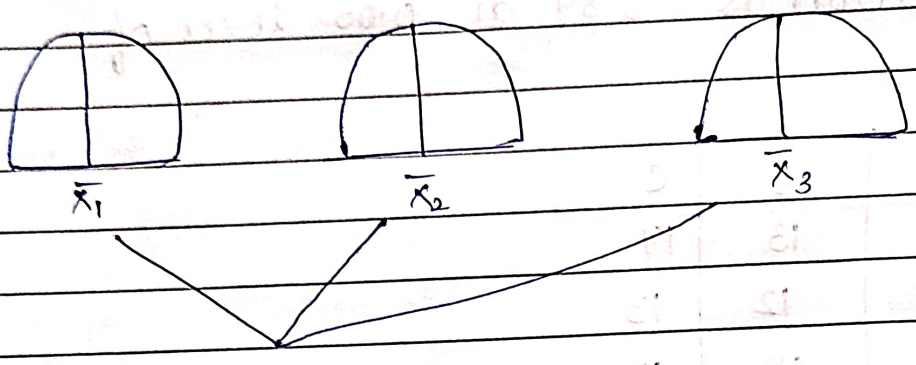
$$Z = \frac{23-25}{8.164} = -0.244$$

$$\boxed{\alpha = -5.24}$$

* Analysis of Variance (ANOVA)

Variance - It is defined as expectation of the square deviation of a random variable of their mean.

- For comparision of more than 2 population or population having more than two sub groups of the same population we use the same technique?



Formula for ANOVA is

$$\text{ANOVA (F)} = \frac{\text{Variability between the mean (X)}}{\text{Variability within the mean (Y)}}$$

Total Variance = X + Y

- Assumption :-

1. Each population is having normal distribution
2. The population from which the samples are drawn have the equal variance.
3. Each sample is drawn randomly and they are independent.

$$S_1^2 = S_2^2 = \cdots S_K^2$$

Null Hypothesis $\mu_1^2 = \mu_2^2 \cdots$

Alternate $\mu_1^2 \neq \mu_2^2$

Page No.:
Date:

Q. To access the significance of possible variation in performance in a certain test between the convent School of a city.
A common entrance test is given to a no. of students taken randomly from 5th class of there different schools.
Their concerned results are given below :-
Perform the 1 way ANOVA test, the critical value of distribution is 3.89 at 0.05 level of significance.

A	B	C
9	13	14
11	12	13
13	10	17
9	15	7
8	5	9

$\bar{X}_A = \frac{50}{5} = 10$ $\bar{X}_B = \frac{55}{5}$ $\bar{X}_C = \frac{60}{5} = 12$

① Calculate the mean $\bar{X}_A = 11$, $\bar{X}_B$, $\bar{X}_C$

② $\bar{\bar{X}} = \frac{\bar{X}_A + \bar{X}_B + \bar{X}_C}{3}$

$= \frac{10 + 11 + 12}{3} = 11$

③ Calculate SSE (Variation b/w the sample)

$SSE = \sum \left[(\bar{X}_A - \bar{\bar{X}})^2 + (\bar{X}_B - \bar{\bar{X}})^2 + (\bar{X}_C - \bar{\bar{X}})^2 \right]$

$\bar{X}_A - \bar{\bar{X}}$	$(\bar{X}_A - \bar{\bar{X}})^2$	$\bar{X}_B - \bar{\bar{X}}$	$(\bar{X}_B - \bar{\bar{X}})^2$	$\bar{X}_C - \bar{\bar{X}}$	$(\bar{X}_C - \bar{\bar{X}})^2$
-1	1	0	0	1	1
-1	1	0	0	1	1
-1	1	0	0	1	1
-1	1	0	0	1	1
-1	1	0	0	1	1

$SSC = 5 + 0 + 5 = 10$

So the value of $SSC = 10$

④ Degree of freedom

$$V = C - 1$$

C is the number of colamns.

$$V = 2$$

⑤ calculate the Mean square

Mean square $= SSC / Df(V)$

$= \frac{10}{2} = 5$

⑥ Now calculate SSE (sum of square within the samples)

$$SSE = \sum (A - \bar{X}_A)^2 + (B - \bar{X}_B)^2 + (C - \bar{X}_C)^2$$

$A - \bar{X}_A$	$(A - \bar{X}_A)^2$	$(B - \bar{X}_B)$	$(B - \bar{X}_B)^2$	$(C - \bar{X}_C)^2$	$(C - \bar{X}_C)^2$
-1	1	2	4	2	4
1	1	1	1	1	1
3	9	-1	1	5	25
-1	1	4	16	-5	25
-2	4	-6	36	-3	9
	16		58		64

SSE = 138

Page No.:
Date:

(7) Degree of freedom

$V_2 = n - c$ $\quad$ $V_2 = n - c$ $\quad$ $V_2 = n - c$

$= 5 - 2$ $\quad$ $= 5 - 2$ $\quad$ $V_2 = (5 \times 3) - 3$

$V_2 = 3.$ $\quad$ $V_2 = 3.$

$\boxed{V_2 = 12}$

(8)

$$\text{Mean square} = \frac{SSE}{V_2} = \frac{138}{12} = 11.5$$

(9)

$$F = \frac{5}{11.5} = 0.434$$

The calculated value f value = 0.435 and f distribution tabular value $\underset{0.05}{\circledast} = 3.89$

(10) Compare these two values so our tabulated value is greater than 0.435 Hence Null hypothesis is passed and no significant variation in the schools marks sheet.

Q.

A	B	C	
2	3	4	Critical value
4	5	6	3.189
6	7	8	

(1) $\bar{X}_A = \frac{12}{3} = 4$

$\bar{X}_B = \frac{15}{3} = 5$

$\bar{X}_C = \frac{18}{3} = 6.$

Date:

(2) $\bar{\bar{X}} = \frac{4+5+6}{3} = \frac{15}{3} = 5$

(3) $SSC = \Sigma\left[(\bar{X}_A - \bar{\bar{X}})^2 + (\bar{X}_B - \bar{\bar{X}})^2 + (\bar{X}_C - \bar{\bar{X}})^2\right]$

$(\bar{X}_A - \bar{\bar{X}})$	$(\bar{X}_A - \bar{\bar{X}})^2$	$(\bar{X}_B - \bar{\bar{X}})$	$(\bar{X}_B - \bar{\bar{X}})^2$	$(\bar{X}_C - \bar{\bar{X}})$	$(\bar{X}_C - \bar{\bar{X}})^2$
-1	1	0	0	1	1
+1	1	0	0	1	1
-1	1	0	0	1	1
	3		0		3

$SSC = 3 + 0 + 3 = 6.$

(4) $V = C - 1$

$V = 3 - 1 = 2$

(5) Mean square $= \frac{SSC}{Df(V)} = \frac{6}{2} = 3.$

(6) $SSE = \Sigma(A - \bar{X}_A)^2 + (B - \bar{X}_B)^2 + (C - \bar{X}_C)^2.$

$(A - \bar{X}_A)$	$(A - \bar{X}_A)^2$	$(B - \bar{X}_B)$	$(B - \bar{X}_B)^2$	$(C - \bar{X}_C)$	$(C - \bar{X}_C)^2$
-2	4	-2	4	-2	4
0	0	0	0	0	0
2	4	2	4	2	4
	8		8		8

$SSE = 8 + 8 + 8 = 24$

(7) $V_2 = n - C$

$V_2 = (3 \times 3) - 3$

$V_2 = 6.$

(8) Meansquare $= \frac{SSE}{V_2} = \frac{24}{6} = 4.$

(9) $f = \frac{3}{4} = 0.75.$

* ~~ANN~~ (~~Artificial Neural Network~~)

* Performance Model of Regression and classification.
In regression there are different approaches available which exactly tells how is your model.

- RMSE - Root Mean Square Error
- MSE - Mean square Error
- MAE - Mean Absolute eq error
- RMSLE - Root mean square log Error.

① RMSE = $\sqrt{\frac{\Sigma(y_i - \hat{y})^2}{N}}$

Let us take one example, find RMSE of the above dataset using LR

BA	TA (y_i)	predicted TA $\hat{y}$	E	E
500	50	52	2	4
550	48	57	-9	81
600	60	62	-2	4
700	65	72	-7	49
800	75	82	-7	49
				187

RMSE = $\sqrt{\frac{187}{85}} = 6.115$

RMSE ↑ model is biased
RMSE ↓ accuracy ↑

Based on the bill amount we have to identify the tip amount the training data tax as a model and identy the function.

② MSE - It is one of the most preferred matrix for the regression task.

Date:

It is simply the average of square difference between actual and predicted value. As it squares the differences it penalizes even a small error which lead to over estimate how bad the model is.

The formula for MSE is

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

Let us take a dataset which consist of dependent and independent variable find performance of a model using MSE.

HT	WT (y_i)	$\hat{y}$	E	E^2
160	72	74.8448	−2.8448	+8.0928
171	76	77.1328	−1.1328	1.2832
182	77	79.4208	−2.4208	5.8602
180	83	79.0048	3.9952	15.9616
154	76.	73.5968	2.4032	5.7753
				36.9731

The regressor line is $\hat{y} = 41.5648 + 0.208\,x_i$

$$\text{MSE} = \frac{36.9731}{5} = 7.3946.$$

MSE is always positive and the value closer to 0 or lower value is better.
if error is zero then model is overfitted.

(3) $$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

N = Number of obs.
y_i = actual
$\hat{y}$ = predicted

It is the absolute difference b/w the actual and the predicted value MAE is the most robust to outliers and doesnot penalize the errors as extremly as MSE. ~~It can be ∞~~

M T W T F S S

Page No.:

Date:

Q. Find out the models performance by using MAE of the following dataset.

Act	predicted	Error	|Error|
100	130	-30	30
150	170	-20	20
200	220	-20	20
250	260	-10	10
300	325	-25	25

$$MAE = \frac{105}{5} = 21$$

(4) RMSLE

(Root Mean square log error) - It is calculated at logarithimic scale while calculating

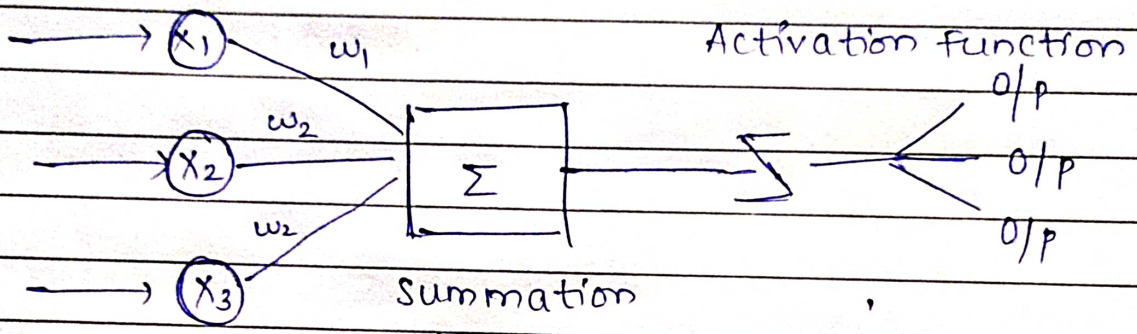
UNIT-4

M T W T F S S
Page No.:
Date:

- ANN - Artificial Neural Network.

The ANN stands for Artificial Neural Network which is part of ML. The Human brain is a complex structure and it is interconnected network of simple processing elements called neurons.

The behaviour of neuron can be captured by using the following model.



The total input can be calculated as

$$Y_{in} = w_1x_1 + w_2x_2 + w_3x_3$$

$$Y_{in} = \sum_{i=1}^{n} x_i w_i$$

Where x_1 x_2 x_3 are the inputs

w_1 w_2 w_3 are the weights

Input signals are verified through waits to generated the final output by Y.

The sum is passed to the activation function which is known as Non-linear activation function ϕ or filter or squash function.

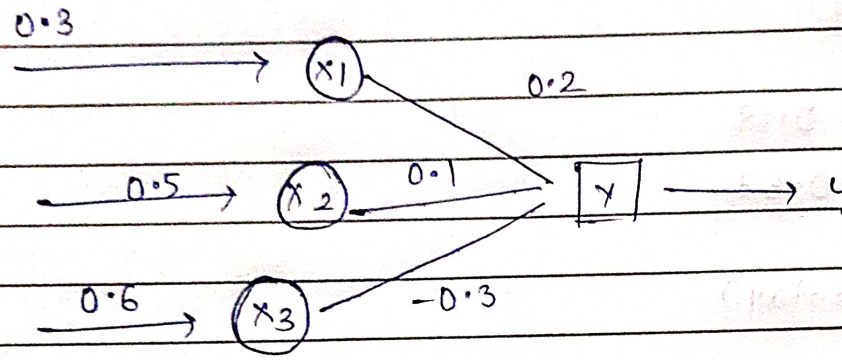
M T W T F S S

Page No.:

Date:

A very commonly used activation function is the thresholding function compared with a threshold value ϕ if the value of $\phi > 0$ then the o/p is 1 → Otherwise 0.

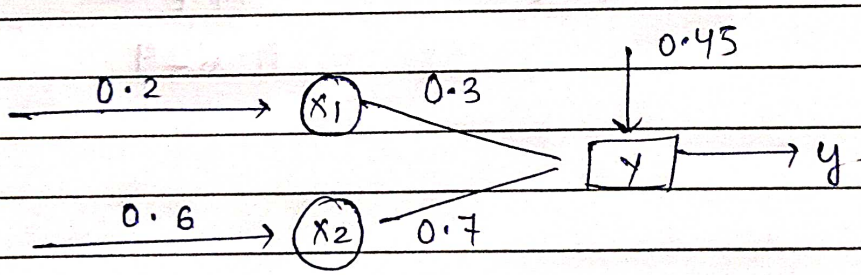
Q. From the given model calculate the total input and weight.



$$Y_{in} = \sum_{i=1}^{N} x_i w_i$$

$$\boxed{Y_{in} = -0.07}$$

Q.



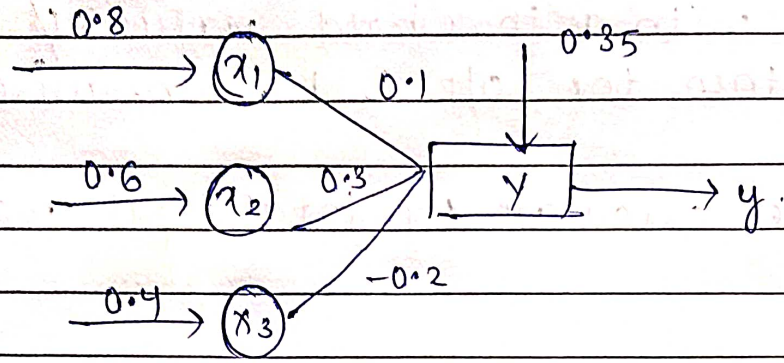
$Y_{in} = 0.48.$

$y = 0.48 + 0.45 = 0.93.$

Q.

Obtain the o/p of the neuron Y for the network shown in figure using activation function

1. binary Sigmoidal
2. Biploar Sigmoidal.



$Y_{in} = 0.18$

$Y = 0.53.$

(2) sigmoidal (Bipolar)

$$Y = f(Y_M) = \frac{2}{1+e^{-Y_{in}}} - 1 = \frac{2}{1+e^{-0.53}} - 1 = 1.258 - 1$$

$$= 0.259$$

(1) $$y = f(Y_M) = \frac{1}{1+e^{-y_{in}}} = \frac{1}{1+e^{-0.53}}$$

(binary)

$$= 0.629$$