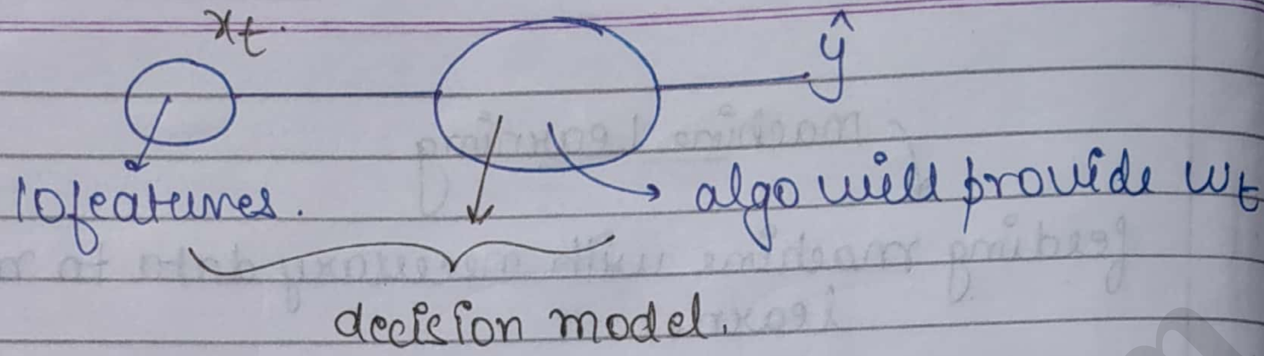


$$w_0 + w_1x_1 + w_2x_2 + \cdots\cdots + w_{10}x_{10} = \hat{y}$$
where $w_0 + w_1 + \cdots w_{10} = 1$

Supervised learning

Classification/ pattern recogonition

- class label is categorical.
- discrete values of class labels to be predicted.
- here we can differentiate objects into diff categories/ discrete values & not range of values.

eg:- Types of fishes (A, B, C, ---)

- y is an integral value

Regression/ estimate.

- used to predict continous scale
- continous value of class labels.

eg:- predicting price of a property based upon some features.

eg:- weight (cont. variable).

- y is a real valued data
  (y: value to be predicted.)

CLASSTIME Page No.
Date / /

NOTE :- Classifiers to be used in supervised learning are :-

(1) Naive Bayes classifier
(2) linear Reg.
(3) logistic
(4) Neural Net.

- Unsupervised learning :- (Clustering).

Here we aren't given labels but some useful patterns. (association)

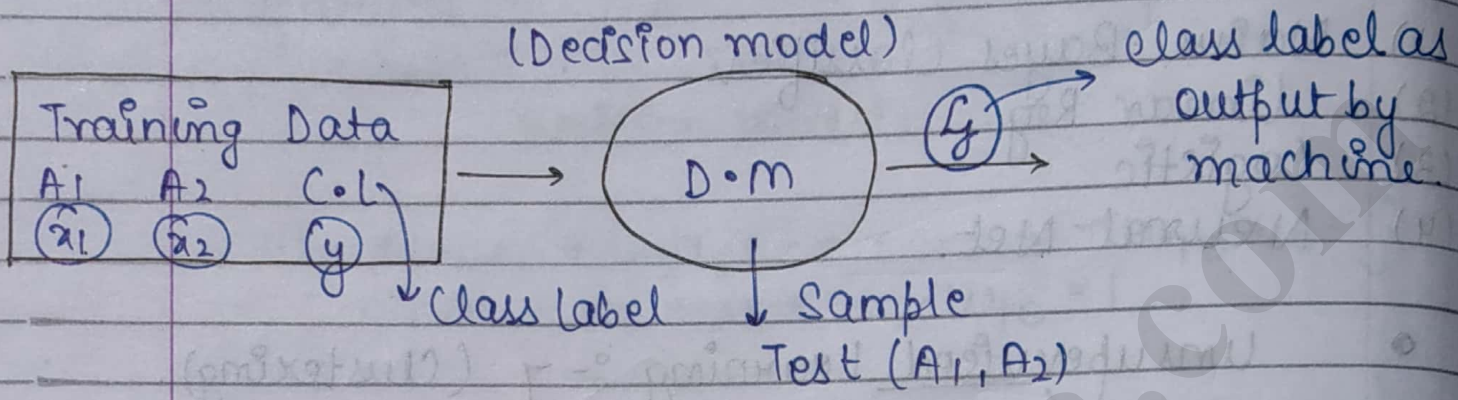
grouping in some patterns

eg :- Basket-market example / Bread butter clustering.

CLASSTIME Page No.
Date / /

4th Jan, 19

- **Naive Bayes :-** {supervised Learning}



Suppose samples = 500
Out of 500 outputs, 480 are correct.

$\therefore$ accuracy $= \frac{480}{500}$ and error $= \frac{20}{500}$

$P(A \cap B) = P(A|B) \cdot P(B) \longrightarrow$ probability of B.

Joint probability.

probability of A given B (conditional probability)

$= P(B|A) \cdot P(A)$

probability of B given A.

$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

$$\boxed{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

Bayes formula.

CLASSTIME Page No:
Date / /

→ Suppose an object has $x_1 \cdots\cdots x_n$ features.

$$X = x_1, x_2 \cdots\cdots x_n$$

and can be classified into k classes

$$C_1 C_2 \cdots C_k.$$

$P(C_i | X) \longrightarrow$ probability of class C_i for given values of x.

Using Bayes formula

$$\frac{P(X|C_i)\ P(C_i)}{P(X)}$$

$P(X) \longrightarrow$ evidence

for $i = 1, 2, 3, \cdots\cdots$

Denominator is fixed.
and numerator varies.

$$\boxed{P(C_i|X) \propto P(X|C_i)\ P(C_i)}$$

↓ Posterior ↓ likelihood → Priori

Find

Let

$0.6 = P(C = C_1 | \langle x_1, x_2 \rangle)$
$0.3 = P(C = C_2 | \langle x_1, x_2 \rangle)$
$0.1 = P(C = C_3 | \langle x_1, x_2 \rangle)$

Sum = 1

max. argument

$$\hat{y} = \text{max. arg } P(C_i|X)$$

for $P(C_i) \rightarrow$ If nothing is given, probability is distributed equally among all C_i's.

priori

(Derivation not imp.)

or

Calculate it from $\underbrace{\text{training data}}_{\text{Sample}}$

1000 — A (400/1000)
1000 — B (500/1000)
1000 — C (100/1000)

$X = x_1, x_2 \cdots x_n$

$$P(X|c_i) = P(<x_1, x_2, x_3 \cdots x_n>|c_i)$$

$$P(X|c_i)\,P(c_i) = P(X \cap c_i)$$

$$= P(\underbrace{x_1}_{\text{breaking}}\,\underbrace{x_2 \cdots x_n c_i}_{\text{every variable in intersection}})$$

$$P(x_1|x_2 \cdots x_n c_i)\,P(x_2 \cdots x_n c_i)$$

$$= P(x_1|x_2 \cdots x_n c_i)\,P(x_2|x_3 \cdots x_n c_i)\,P(x_3 \cdots x_n c_i)$$

let $X = x_1 \cdot x_2 \cdot x_3$

$$= P(x_1|x_2 x_3 c_i)\,P(x_2|x_3 c_i)\,P(x_3 c_i)$$

$$= P(x_1|x_2 x_3 c_i)\,P(x_2|x_3 c_i) \cdot P(x_3|c_i)\,P(c_i) \quad (1)$$

NOTE:- Naive Bayes classification applies on those attributes that are independent of each other.

CLASSTIME Page No.
Date / /

(Length and weight are not independent) and increase in length increases weight.

$$P(A \cap B) = \underbrace{P(A|B) \cdot P(B)}$$

Given - If A and B are independent events then $P(A \cap B) = P(A) \cdot P(B)$.

eqn ① can be written as :-

$$= P(x_1|c_i) \cdot P(x_2|c_i)\, P(x_3|c_i)\, P(c_i)$$

(as x_1 is independent of x_2 and x_3) {for Naive's Baye's classification}

$$= \prod_{j=1}^{n} (P(x_j|c_i))\, P(c_i)$$

Imp

$$\boxed{V_{NB} = \text{argmax}\, P(v_j) \prod_i P(a_i/v_j)}$$

(Here $v_j = c_i$ and $a_i = x_j$)

Example

	x_1	x_2	x_3	x_4	C.L (y)	
Day	Outlook	Temp	Humidity	Wind	PlayTennis	
D1	Sunny	Hot	High	Weak	No	
D2	Sunny	Hot	High	Strong	No	
D3	Overcast	Hot	High	Weak	Yes	1
D4	Rain	Mild	High	Weak	Yes	2
D5	Rain	Cool	Normal	Weak	Yes	3
D6	Rain	Cool	Normal	Strong	No	

D7	Overcast	Cool	Normal	Strong	Yes	4
D8	Sunny	Mild	High	Weak	No	
D9	Sunny	Cool	Normal	Weak	Yes	5
D10	Rain	Mild	Normal	Weak	Yes	6
D11	Sunny	Mild	Normal	Strong	Yes	7
D12	Overcast	Mild	High	Strong	Yes	8
D13	Overcast	Hot	Normal	Weak	Yes	9
D14	Rain	Mild	High	Strong	NO	

$$V_{NB} = \underset{V_j \in \{Yes, No\}}{\arg\max} \; P(V_j) \prod_i P(a_i | V_j)$$

Here from Training Data,

$P(Yes) = \frac{9}{14}$ ⟸ max.

$P(No) = 5/14$.

Given → (Outlook = Sunny
Humidity = High
Wind = Strong
Temp = Cool)

→ Decision model will have 22 parameters irrespective of any size of data.

P(Yes), P(No), P(Sunny | Yes). P(Sunny | No)
P(Overcast | Yes). P(Overcast | No)
P(Rain | Yes). P(Rain | No)

This way we will have
$2 + 6 + 6 + 4 + 4 = \underline{22}$.

CLASSTIME Page No.
Date / /

$P(Yes) = \frac{9}{14}$ $P(No) = \frac{5}{14}$

$P(outlook = sunny \mid Yes) = \frac{2}{9}$

$P(Sunny \mid No) = \frac{3}{5}$ $5 \longrightarrow (14-9)$

$P(outlook = sunny) = \frac{5}{14}$

$P(Yes \mid outlook = sunny) = \frac{2}{5}$

$\frac{P(Sunny \mid Yes) \cdot P(Yes)}{P(Sunny)} = \frac{\frac{2}{9} \times \frac{9}{14}}{5/14} = \frac{2}{5}$

$P(Temp = cool \mid Yes) = \frac{3}{9}$

Total (Yes = 9)

$P(cool \mid No) = \frac{1}{5}$

$P(Wind = Strong \mid Yes) = \frac{3}{9}$

$P(Wind \mid No) = \frac{3}{5}$

$P(High \mid Yes) = \frac{3}{9}$ $P(High \mid No) = \frac{4}{5}$.

(Yes)

$P(Yes \mid X') = \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{2}{378}$

proportion (α) = 0.005291

$P(No \mid X') = \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{4}{5} = \frac{18}{875}$

(no)

Proportion (α) = 0.0205

Final answer ↓

$$V_{NB} = \text{max of (Yes or no)}$$

$$= \underline{No}$$

$$\boxed{\text{class label} = \text{No}}$$

Probability of No and Yes

$$P(X) = \sum_{i=1}^{n} P(X|C_i)\,P(C_i)$$

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

$$P(\text{Yes} \mid X) = \frac{0.00529}{0.00529 + 0.0205} = \underline{0.205}$$

$$P(\text{No} \mid X) = \frac{0.0205}{0.00529 + 0.0205} = \underline{0.794}$$

4 Jan, 19

Practical

> help.start()

> a = 1:5 (1D array of length 5)
> a[1] = 7

1 to 5

> a = seq(from=1, to=10, by=2 {1 3 5 7 9}
> a = seq(1, 10, 2)

> ? seq (Help) // ? name

> a // print
[1] 7 2 3 4 5.

→ If we don't know exact function.
?? name ⇒ displays everything related to this.

To know about fname's declaration and definition.

> fnname
{outputs full function}.

- For initializing 1D array / column vector.

> a = c(20, 25, 24, 23) {no need to define datatype but only one type of array}
> a
[1] 20 25 24 23

- For sequential array {1D array}

CLASSTIME Page No.
Date / /

① > a = seq (from = 1, to = 10, by 2)
> a
[1] 1 3 5 7 9

or

② > a = 1:5 (1 to 5)
> a
[1] 1 2 3 4 5

> a = 1:100
> a
[1] 1 2 3 4 5 - - - - - - - 100.

> b = c ("Everyone", "likes", "machines")
> b
[1] "Everyone" "likes" "machines"

if > b = c (2, "Everyone")
> b
[1] "2" "everyone".

> example ("seq") // displays usage of seq.

> help (seq) // tells about sequence & its methods.

> v = c (0, 1, 1, 2, 3, 5, 8, 13, 21, 34)
> x = v [3]
> x
[1] 1

// fetching sub array from index 1 to 5.

> v [1:5]
[1] 0 1 1 2 3

```
> v [c(1,5,2,8)]      // extract elements of
                          these indexes.
  [1] 0 3 1 1 3

> v [-1]              // removes index value from
  [1] 1 1 2 3 5 8 13 21 34   v and doesnt
                          make change in v.

> v [-1:-2]   // v [-(1:2)]   { remove index
  [1] 2 3 5 8 13 21 34        value from 1 to 2 }

> v [-c(1,3)]   // remove these particular
                   index values.
  [1] 3 18 13 21 24.

> v1 = c(1,2,4)
> v2 = c(34,56,2)
> v3 = c(v1,v2)     // combine v1 and v2.
> v3
  [1] 1 2 4 34 56 2

> c(1,2,3.1)        // automatic conversion
  [1] 1.0 2.0 3.1      to higher DT.

> v1 = c(1,2,3)     } comparing index
> v2 = c(1,2,4)       value of 1 with
> v1 == v2            index value of 2
  [1] True True false  and then returning
                        true.

v1 = 1
v2 = "1"
v1 == v2
  [1] True.
```

⟹ Addition]
① If size is same for both arrays add corresponding index values.

② If size mismatches, then first add till same size and then start again adding from start.]
and also display warning.

→ > v_1 = c(1,2,3)
> v_2 = c(1,2,4)
> v_1 + v_2
[1] 2 4 7.

> v_1 = c(1,2,3)
> v_2 = c(1,2,3,4)
> v_1 + v_2 → Warning message.
[1] 2 4 6 5 → 4+1

> a]
[1] 1 2 3.
> b]
[1] 3 4 5
> a * b]
[1] 3 8 15
(1x3) (2x4) (5x3)

multiply & divide in similar manner

$\frac{5}{3}$.

> a / b]

[1] 0.33 0.500 0.6000

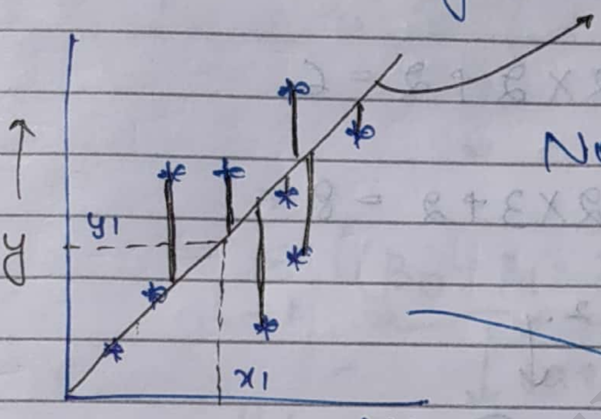
CLASSTIME Page No:
Date / /

11 Jan, 19

* Linear Regression -

Forming a line given the training data, we need to know slope and intercept

$y = mx + c$



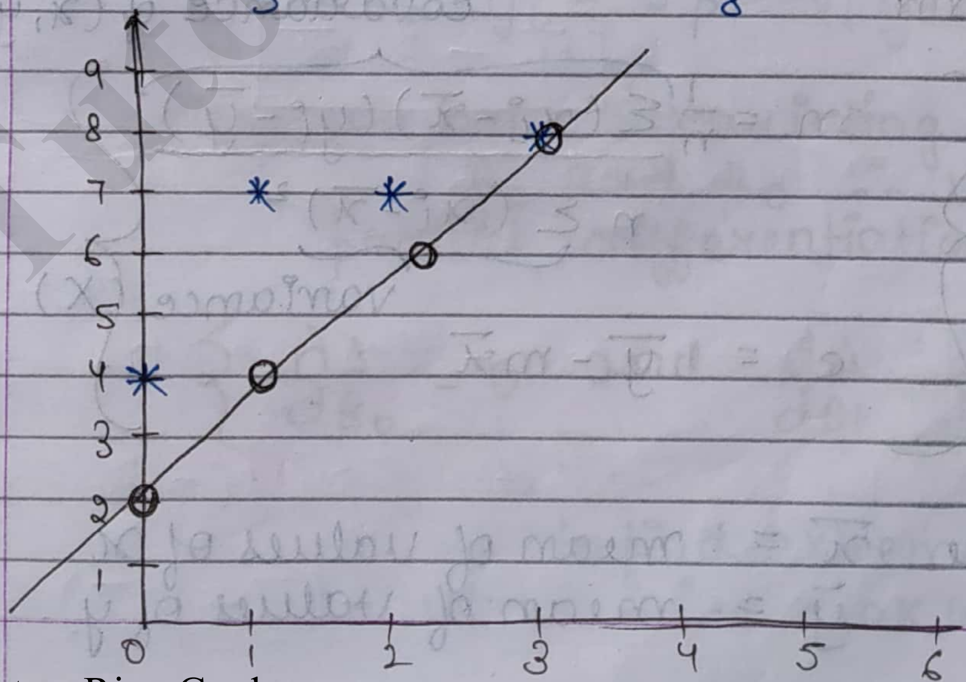
Now if some x_1 comes, we can estimate a price of y. denoted by y_1.

line must be such that sum of vertical distances must be least.

Ex-

To predict the best suitable line

I/P x	O/P y
0	4
1	7
2	7
3	8



CLASSTIME Page No.
Date / /

Assuming $m=2$ and $c=2$

$y = mx+c$

(estimated val.) $\hat{y}$

(0,4) $\hat{y} = 2\times 0+2 = 2$

(1,7) $\hat{y} = 2\times 1+2 = 4$

(2,7) $\hat{y} = 2\times 2+2 = 6$

(3,8) $\hat{y} = 2\times 3+2 = 8$

Now $(\hat{y}-y)^2$

we can either take | | or do $(\ )^2$ of errors.

4
9
1
0

adding them = 14.

$$\boxed{\Delta = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2}$$

error

$$m = \frac{\overbrace{\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})}^{\text{covariance of }(x,y)}}{\underbrace{\frac{1}{n}\Sigma(x_i-\bar{x})^2}_{\text{variance }(X)}}$$

$$c = \bar{y} - m\bar{x}$$

where $\bar{x}$ = mean of values of x

$\bar{y}$ = mean of values of y.

CLASSTIME Page No:
Date / /

Derivation -

$$\Delta = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2$$

Considering $y = mx + c$

Put $c = \beta_0$

and $m = \beta_1$

$$\therefore \Delta = \sum_{i=1}^{n} ((\beta_0 + \beta_1 x_i) - y_i)^2$$

$$= \sum_{i=1}^{n} (\underbrace{(\beta_0 + \beta_1 x_i)^2}_{(a+b)^2} + y_i^2 - 2y_i(\beta_0 + \beta_1 x_i))$$

$$= \sum_{i=1}^{n} (\beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0\beta_1 x_i + y_i^2 - 2y_i\beta_0 - 2y_i x_i \beta_1)$$

$$= \sum_{i=1}^{n} \beta_0^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2 + 2\beta_0\beta_1 \sum x_i + \sum y_i^2 - 2\beta_0 \sum y_i - 2\beta_1 \sum x_i y_i .$$

Here we have 2 variating quantities β_0 and β_0 $\therefore$ we perform partial differentiation

$$\left\{ \frac{\partial \Delta}{\partial \beta_0} = 0 \text{ and } \frac{\partial \Delta}{\partial \beta_1} = 0 \right\}$$

to find extreme pts. i.e max or min. pt.

CLASSTIME Page No.
Date / /

Diff. wrt β_0

$$\frac{d\Delta}{d\beta_0} = 2\sum_{i=1}^{n} \beta_0 + 2\beta_1 \Sigma x_i - 2\Sigma y_i$$

(diff. wrt β_1)

$$\frac{d\Delta}{d\beta_1} = 2\sum_{i=1}^{n} \beta_1 x_i^2 + 2\beta_0 \Sigma x_i - 2\Sigma x_i y_i$$

Now $\frac{d\Delta}{d\beta_0} = 0$

$$\sum_{i=1}^{n} \beta_0 + \beta_1 \Sigma x_i - \Sigma y_i = 0$$

$$n\beta_0 + \beta_1 \Sigma x_i - \Sigma y_i = 0$$

$$\beta_0 = \frac{\Sigma y_i - \beta_1 \Sigma x_i}{n}$$

(mean of y)

$$= \frac{\Sigma y_i}{n} - \frac{\beta_1 \Sigma x_i}{n}$$

$$= \boxed{\bar{y} - \beta_1 \bar{x}}$$

Now $\frac{d\Delta}{d\beta_1} = 0$

$$\beta_1 \sum_{i=1}^{n} x_i^2 + \beta_0 \Sigma x_i - \Sigma x_i y_i = 0.$$

$$\beta_1 \sum_{i=1}^{n} x_i^2 = \Sigma x_i y_i - \beta_0 \Sigma x_i$$

$$= \Sigma x_i y_i - (\bar{y} - \beta_1 \bar{x}) \Sigma x_i$$

$$\beta_1 \sum x_i^2 + \left(\frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}\right)\sum x_i - \sum x_i y_i = 0$$

$$\beta_1\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) + \frac{\sum x_i \sum y_i}{n} - \sum x_i y_i = 0.$$

$$\beta_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

(Slope=m)

$$\boxed{= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}}$$

Now $\beta_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$$n\left(\frac{\sum x_i}{n} \cdot \frac{\sum y_i}{n}\right)$$

$$\frac{\sum x_i y_i - \sum x_i \bar{y} - \sum \bar{x} y_i + \sum \bar{x}\bar{y}}{\sum x_i^2 + \sum \bar{x}^2 - 2\sum x_i \bar{x}}$$

$$\frac{\sum x_i y_i - \cancel{\frac{\sum x_i \sum y_i}{n}} - \frac{\sum x_i}{n} y_i + \cancel{\frac{\sum x_i \sum y_i}{n}}}{\sum x_i^2 + \underset{n}{\sum}\frac{x_i^2}{n^2} - 2\frac{\sum x_i \sum x_i}{n}}$$

$$\cancel{= \frac{\sum x_i y_i \left(1 - \frac{1}{n}\right) + \sum x_i \sum y_i \left[\frac{1}{n} \sum\right]}{\sum x_i^2 \left(1 + \frac{1}{n}\right) - 2\frac{(\sum x_i)^2}{n}}}$$

$$= \frac{\sum x_i y_i \left(1 - \frac{1}{n}\right)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

CLASSTIME Page No.
Date / /

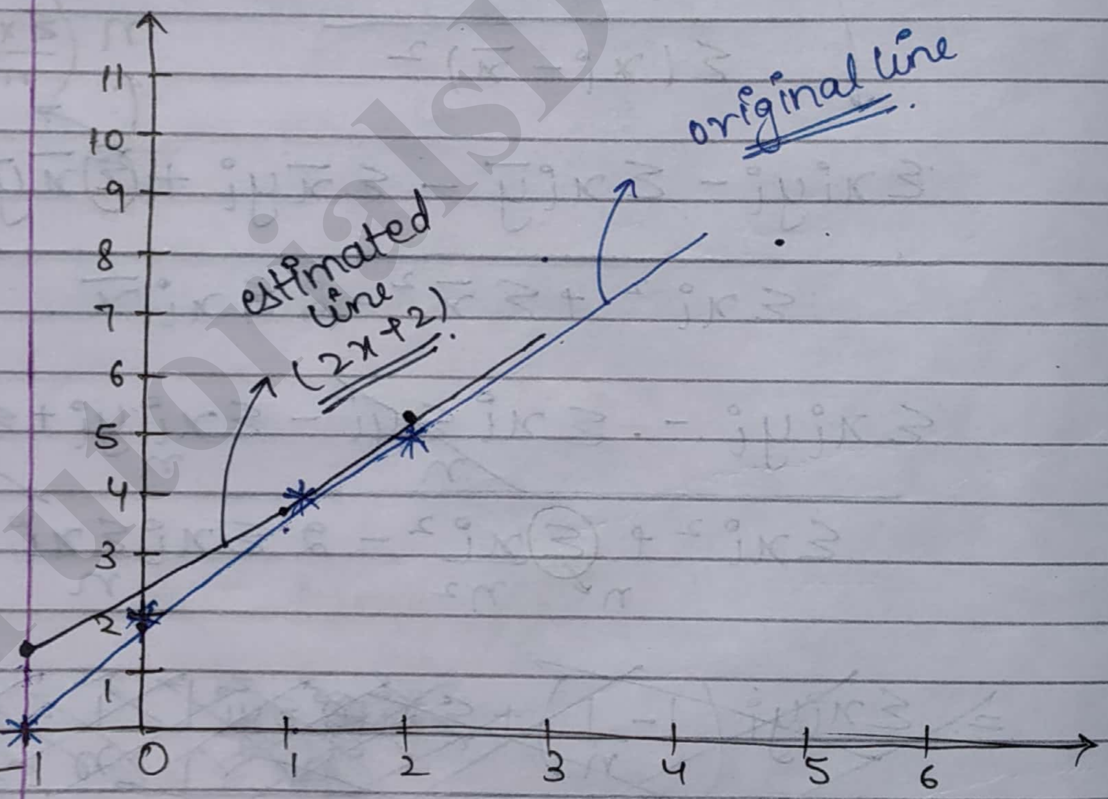
$$\boxed{= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}}$$

Q) (-1, 0), (0, 2), (1, 4), (2, 5)

↓

Find the least regression line

x	y
-1	0
0	2
1	4
2	5



Choosing $m = 2$ and $c = 2$

$$\hat{y} = 2x + 2$$

x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$\Sigma(x_i-\bar{x})(y_i-\bar{y})$
-1	0	-3/2	-11/4	33/8
0	2	-1/2	-3/4	3/8
1	4	1/2	5/4	5/8
2	5	3/2	9/4	27/8

$\bar{x} = \frac{1}{2}$, $\bar{y} = \frac{11}{4}$

$\therefore \frac{33+3+5+27}{8} = \frac{68}{8}$

$\Sigma(x_i-\bar{x})^2$

$\left.\begin{matrix} 9/4 \\ 1/4 \\ 1/4 \\ 9/4 \end{matrix}\right\} \frac{20}{4} = 5$

$\therefore m = \frac{\Sigma(x_i-x)(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{68}{8\times5} = \frac{17}{10}$

$= 1.7$

$c = \bar{y} - m\bar{x}$

$= \frac{11}{4} - \frac{17}{10}\left(\frac{1}{2}\right) = \frac{38}{20} = 1.9$

Now $\hat{y} = 2x+2$

$\therefore$ Plotting line, we get

x	$\hat{y}$
-1	1.2
1	3.6
0	1.9
2	5.3

For predicting value of y for $x=3$

$\hat{y} = 2(3)+2$

$= 8$

{Vector - 1D array}
{Matrix - 2D array}

a = c(2, 3, 5, 56, 7, 4)
↓
Vector

→ a = matrix(c(1,2,3,4), $\underbrace{2, 2}_{m \times n}$)
↓

	[,1]	[,2]
[1,]	1	3
[2,]	2	4

a = matrix(1:12, 2, 6)

Starts filling rowwise

	[,1]	[,2]	-	-	-	[,6]
[1,]	1	3	5	7	9	11
[2,]	2	4	6	8	10	12

→ a = 1:12
$\underbrace{\text{dim(a) = c(2, 2, 3)}}_{\text{3 dimensions}}$

, , 1

1	3
2	4

, , 2

5	7
6	8

, , 3

9	11
10	12

3 matrix of 2X2

CLASSTIME Page No.
Date / /

* Control Loops -

help ("control") in sourcepanel.
↓
Syntax of all loops.

print (1:i)
↓
print values from 1 to 'i'

eg:- for (i in 1:5) print (1:i)

[1]	1					(1 to 1)
[1]	1	2				(1 to 2)
[1]	1	2	3			(1 to 3)
[1]	1	2	3	4		(1 to 4)
[1]	1	2	3	4	5	(1 to 5)

> if (a == 9) a = a + 2. { a = 28 initially }

> if (a == 9) a = a + 2 else a = a + 3
∴ a = 31

If and else

• For loop ↓

for (i in 1:5) print (1:i).

for (i in c(2,1,3,5)) print (1:i)

```
[1] 1 2
[1] 1
[1] 1 2 3
[1] 1 2 3 4 5
```

- While Loop :-

```
while ( i<=25) { print (i) ; i=i+5}
```

For 2 stmts in same lines, use (;)

For ending infinite loop, use Esc.

```
repeat { if ( i>25) break else
           { print (i) ; i = i+5 } }
```

func. name

```
> gcd = function (a, b) {
+     if (b==0) return (a)
+     else return (gcd(b, a%%b))
+ }
```

+ expects more stmts.

%% remainder division/ modulo division.

```
> gcd (34,15)
    [1] 1
> gcd (12,34)
    [1] 2
```

This is made on command prompt ∴ can be directly used.

CLASSTIME Page No:
Date / /

→ gcd 2 = function(a,b)
{
if (b==0) return(a)
else return (gcd (b, a%%b))
}

If we write this in Source panel, then it is first necessary to run (compile) this code so that gcd2 comes into the library.

Now on Console panel → gcd2(34,12)
↓
[1] 2

Q f1 = function(a,b) {
if(a==0) return (b+2)
else return (b+3)
}

> f1(0,34) → 36

Q Factorial of a no.

fact = function(f) {
if (f==0) return(1)
else return(f * fact(f-1))
}

fact(5) → 120

CLASSTIME Page No:
Date / /

```
gcd2 = function(a,b)
{
    if (b==0) return(a)
    else return(gcd(b, a%%b))
}
```

If we write this in Source panel, then it is first necessary to run (compile) this code so that gcd2 comes into the library.

Now on Console panel → gcd2(34,12)
[1] 2

Q

```
f1 = function(a,b) {
    if(a==0) return(b+2)
    else return(b+3)
}
```

> f1(0,34)
36

Q Factorial of a no.

```
fact = function(f) {
    if(f==0) return(1)
    else return(f * fact(f-1))
}
```

fact(5) → 120

CLASSTIME Page No.
Date / /

Q Find whether a no. is prime

```
ch1 = "Y"
ch2 = "N"
prime = function(p) {
    n = p/2
    for (i in 2:n) if((p%%i)==0)
        return(ch2)
    return(ch1)
}
```

prime(4)
[1] "N"

prime(3)
[1] "Y".

CLASSTIME Page No:
Date / /

15 Jan, 19

Machine Learning

Q

x_1	x_2	x_3	x_4	y
size (feet)^2	no. of bedrooms	no. of floors	Age of home (yrs)	Price (\$1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178

$m = 47$

⟶ $x_j^{(i)}$ → training example

j → j^{th} feature.

eg:- $x_3^{(2)}$ → 2^{nd} training set and 3^{rd} feature.

$$x^{(1)} = \begin{bmatrix} 2104 \\ 5 \\ 1 \\ 45 \\ 460 \end{bmatrix}$$

$$\text{Suppose } X = \begin{bmatrix} 2104 & 5 & 1 & 45 \\ 1416 & 3 & 2 & 40 \\ 1534 & 3 & 2 & 30 \\ 852 & 2 & 1 & 36 \end{bmatrix}$$

$$\underbrace{y = \beta_0 + \beta_1 x}_{\text{for 1 variable}}$$

Now for multiple features that are to predicted is

↓

coeff. with features
(weightage of features)

$$\rightarrow \quad y = \underbrace{\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4}_{\text{4 features}}$$

θ_0 → constant
(represents a base value)

(same as $ax^2 + bx + \underline{c}$)

$$y = \theta_0 + \underbrace{\frac{dy}{dx_1}}_{\theta_1} + \underbrace{\frac{dy}{dx_2}}_{\theta_2} + \cdots + \underbrace{\frac{dy}{dx_4}}_{\theta_4}$$

$$\therefore \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}_{(5\times1)} \text{ and } X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix}_{(4\times5)}$$

assuming an x_0 for uniformity $(\theta_0 x_0)$

$$\therefore y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}_{(4\times1)}$$

$$\boxed{\underset{(4\times1)}{y} = \underset{(4\times5)}{X}\, \underset{(5\times1)}{\theta}}$$

$$\rightarrow \quad ① \quad \boxed{\theta = (X^T X)^{-1} X^T Y}$$

$$\left\{ \begin{aligned} Y &= X\theta \\ X^{-1}Y &= X^{-1}X\theta \\ X^{-1}Y &= I\theta \\ \underline{X^{-1}Y} &\underline{= \theta} \end{aligned} \right\}$$

This is not possible as 47X5 is not a sq. matrix.

Proof of 1 :-

$Y = X\theta$

$YX^T = \underbrace{XX^T}\theta$

becomes a sq. matrix as $X \rightarrow 47\times5 \longrightarrow 47\times47$

$X^T \rightarrow 5\times47$

CLASSTIME Page No.
Date / /

Assuming, inverse can be found $(X^TX)^{-1}X^TY = (X^TX)^{-1}X^TX\theta$

$\underline{I}$

$\therefore \theta = (X^TX)^{-1}X^TY$

- If 2 features are related to each other, then $\Delta = 0$ $\therefore$ inverse can't be calculated.

Limitation of this method.

- The inverse operation on XX^T is quite expensive as θ becomes of $O(n^3)$.

→ To overcome these problems, we have another method.

* Gradient Descent method :- (Used in neural network)

(slope) (reducing).

Calculating for 1 variable

$\hat{y} = \theta_0 + \theta_1 x$

$h_\theta(x) = \theta_0 + \theta_1 x.$ // Hypothesis func.

$J(\theta_0, \theta_1) = \boxed{\min_{\theta_0, \theta_1} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2}$

(cost fn)

where $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

Average error

↓

$\left\{ \frac{\text{error}}{m} \right\}$

predicted value.

original value

and m = no. of features = 1. (rows)

and $J(\theta_0, \theta_1)$

↓

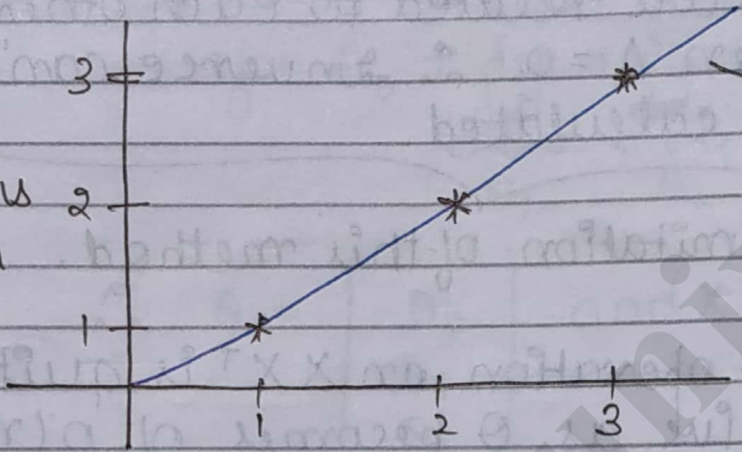
Cost function / Penalty func.

$$= \min_{\theta_0 \theta_1} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)} \right)^2$$

$$h_\theta(x) = \theta_1 x$$

{assuming line passes through origin}

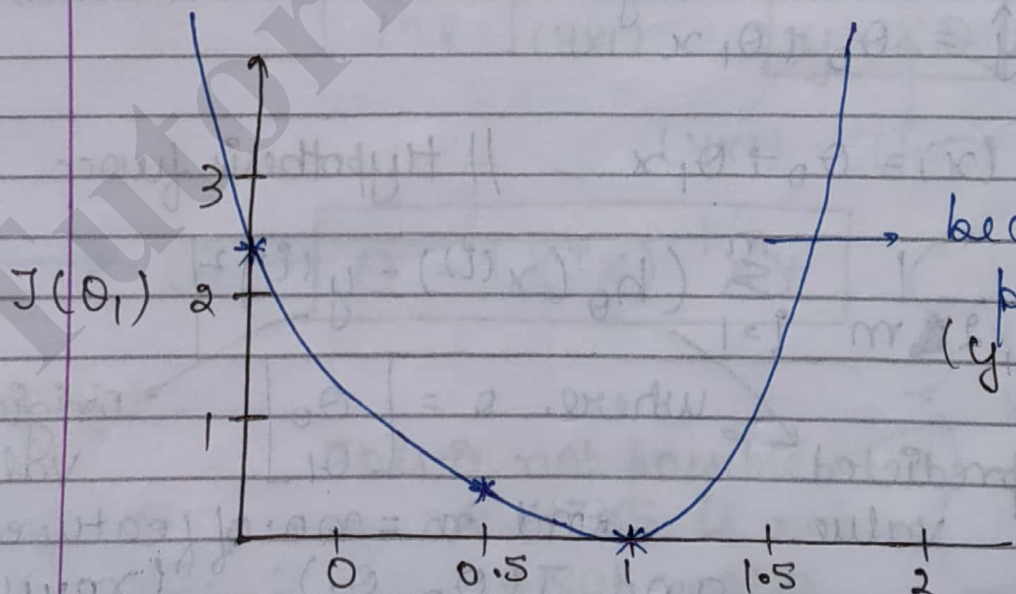
graph of hypothesis function



no error here.

$$\left\{ \underbrace{\theta_1}_{(m)} = 1 = \tan 45^\circ \right\}$$

$$\therefore \min_{\theta_0 \theta_1} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_1 (x)^{(i)} - y^{(i)} \right)^2$$



becomes a parabola. $(y = x^2)$

θ_1 (cost func)

(case1)

x	y
1	1
2	2
3	3

$\theta_1 = 1$

$$(\theta_1 (x)^{(i)} - y^{(i)})^2$$

$$\left.\begin{array}{l} 1(1) - (1) = 0 \\ 1(2) - (2) = 0 \\ 1(3) - (3) = 0 \end{array}\right\} \frac{0+0+0}{2} = 0$$

(case2)

Now if $\theta_1 = 0.5$

$$(\theta_1 (x)^i - y^{(i)})^2$$

$$= \frac{1}{2m}\left((0.5\times1)-1)^2 + (.5\times2-2)^2 + (.5\times3-3)^2\right)$$

$$= \underline{0.76}$$

For $\theta_1 = 0$ value $= \underline{2.3}$ (case3)

→ from graph, we need to choose min value of θ_1

23/1/19

$$y = \underbrace{\theta_0 + \theta_1 x + \theta_2 x_2 + \theta_3 x_3}_{\text{linear relation}}$$

weightage

As we keep on ↑ing degree of variables, it becomes a polynomial curve.

polynomial regression

x_i → where i represents feature no.
For n features, there will be x_{n+1} weightage var.

$$y^{(1)} = \theta_0 + \theta_1 x_1^{(1)} + \theta_2 x_2^{(1)} + \theta_3 x_3^{(1)}$$

superscript is sample no.

Subscript denotes feature no.

$$\underbrace{y^{(2)} = \theta_0 + \theta_1 x_1^{(2)} + \theta_2 x_2^{(2)} + \theta_3 x_3^{(2)}}$$

Given 2 samples

Given →

x_1	x_2	x_3	y
1	2	-1	5
0	3	2	7

To find → $\theta_0, \theta_1, \theta_2, \theta_3$.

Rewriting, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}_{(4\times1)}$ and

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} \end{bmatrix}_{(2\times4)}$$

$$Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix} \to (2\times1)$$

$$\therefore Y = \underset{(2\times4)}{X}\ \underset{(4\times1)}{\theta}$$

{ Here inverse of X is not possible as X is not a sq. matrix

CLASSTIME Page No.
Date / /

$$\therefore \theta = (X^T X)^{-1} X^T Y$$

Thus, if features increases, complexity increases.

$O(n^3)$ → n = no. of features.

- To find min. in graph → use gradient descent method.
  finds nearest local min.
- To find max. in graph → use gradient ascend method.
  finds nearest local max.

has a very low convergence rate

Gradient Descent method :-

$$\underset{(new)}{\theta_j} = \underset{(old)}{\theta_j} - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

$J(\theta)$ → func. to be minimized

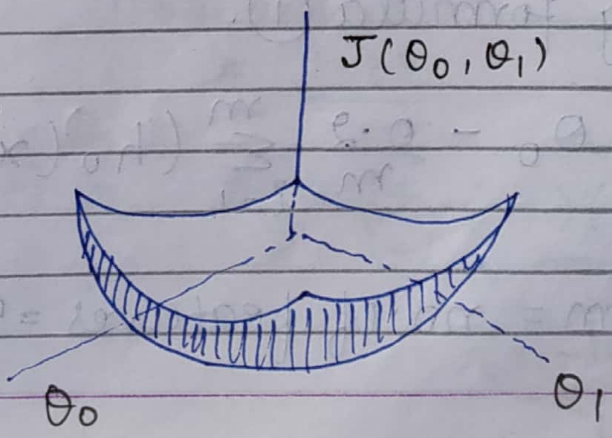
α → gives step size.

{same parabola graph as drawn previously}.

repeat this process until it converges.

NOTE: α shouldn't be too large as the eq^n will diverge.

For 3-D ↓



$$\theta_0 = \theta_0 - \alpha \frac{\partial}{\partial \theta_0} \left(\underbrace{\frac{1}{2m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2}_{J(\theta_0, \theta_1).} \right)$$

Using chain rule

$$(1) \quad \boxed{\theta_0 = \theta_0 - \frac{2\alpha}{2m} \sum_{i=1}^{m} (h_0(x^{(i)}) - y^{(i)})}$$

Similarly for θ_1

$$(2) \quad \boxed{\theta_1 = \theta_1 - \frac{\alpha}{m} \left(\sum_{i=1}^{m} (h_0(x^{(i)}) - y^{(i)}) \cdot x^{(i)} \right)}$$

Previous Year Question PPr

X	Y
0	3
1	4
2	5
3	4
4	6

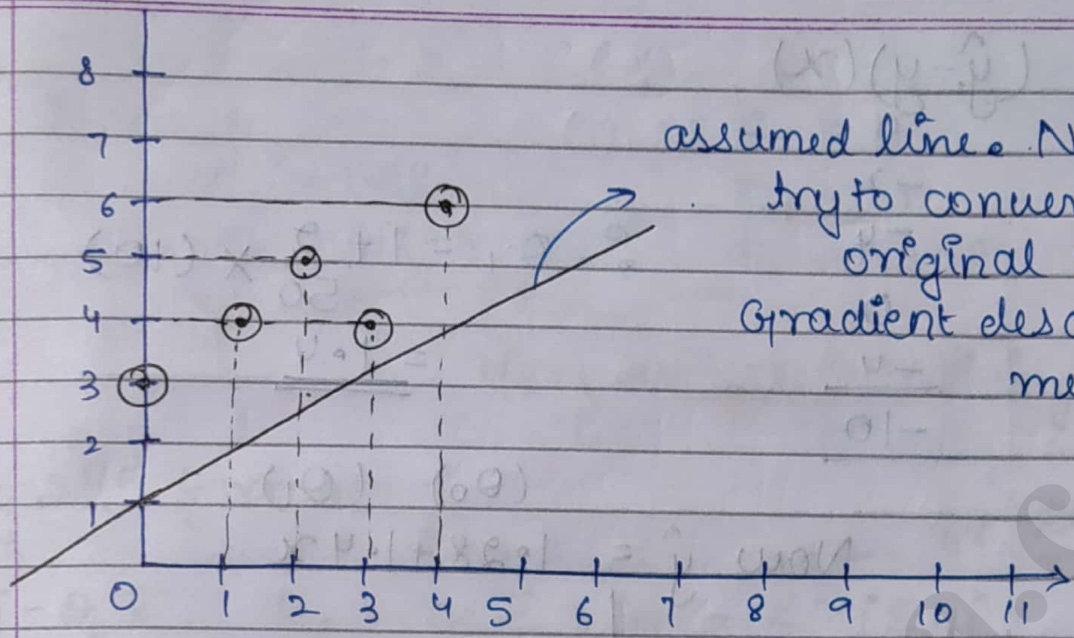
Let $\alpha = 0.2$ and initial value of $\theta_0 = 1$ and $\theta_1 = 1$

We show upto 2 iterations

Using formula (1).

$$\theta_0 = \theta_0 - \frac{0.2}{m} \sum_{i=1}^{m} (h_0(x^{(i)}) - y^{(i)})$$

m = no. of features = 5



assumed line. Now we try to converge to original values using Gradient descent method.

1st iteration

$\hat{y} \rightarrow \theta_0 + \theta_1 x$

$= (1+x)$

1
2
3
4
5

$\therefore \theta_0 = \theta_0 - \frac{0.2}{5} \sum_{i=1}^{5} (\hat{y}-y)$

$\hat{y}-y$
-2
-2
-2
0
-1

$\therefore \theta_0 = 1 - \frac{2}{50}(-2-2-2+0-1)$

$= \underline{1.28}$

Now $\theta_1 = \theta_1 - \frac{0.2}{5} \sum_{i=1}^{5} \underbrace{(-7)(x^{(i)})}_{(\hat{y}-y)x}$

CLASSTIME Page No.
Date / /

$(\hat{y}-y)(x)$

0
-2
-4
0
-4
-10

$\therefore \theta_1 = 1 + \frac{2}{50} \times (10)$

$= 1.4$

$(\theta_0) \quad (\theta_1)$

Now $\hat{y} = 1.28 + 1.4x$

x	y	$\hat{y} = 1.28 + 1.4x$
0	3	1.28
1	4	2.68
2	5	4.08
3	4	5.48
4	6	6.88
		20.4

2nd iteration

x	y	$\hat{y}$	$\hat{y}-y$	$(\hat{y}-y)x$
0	3	1.28	-1.72	0
1	4	2.68	-1.32	-1.32
2	5	4.08	-0.92	1.84
3	4	5.48	+1.48	
4	6	6.88	0.88	

$\Sigma(\hat{y}-y)$

CLASSTIME Page No.
Date / /

Now $\theta_0 = 1$ $\quad$ $\theta_1 = 1$
$\theta_0 = 1.28$ $\quad$ $\theta_1 = 1.4$
$\theta_0 = 1.344$ $\quad$ $\theta_1 = 1.208$

Here we see θ is ↑ing

Using 2nd formula ↓

$\theta_0 = \bar{y} - \theta_1 \bar{x}$ $\quad$ $\theta_1 = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x)^2}$ or $\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$

x	y	$\Sigma x_i y_i$	Σx_i^2
0	3	0	0
1	4	4	1
2	5	10	4
3	4	12	9
4	6	24	16
		50	30

∴ Putting values in formula.

$n = 5$

$\Sigma x_i \rightarrow 10$
$\Sigma y_i \rightarrow 22$

Now $\bar{x} = \frac{10}{5} = 2$

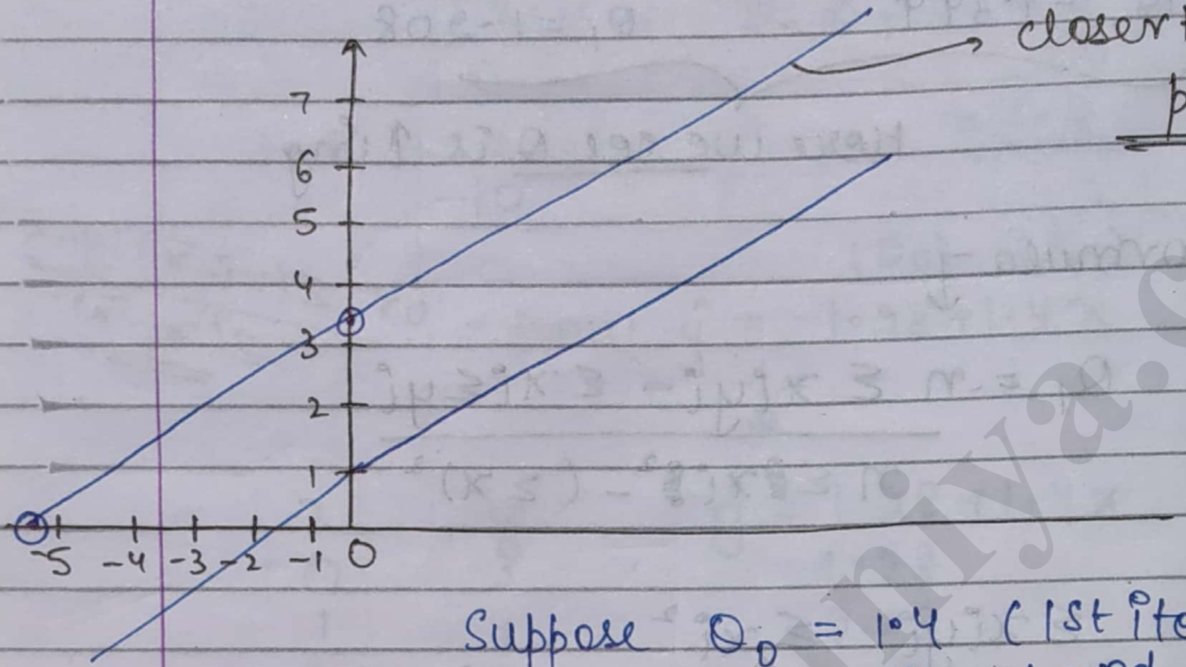
$\bar{y} = \frac{22}{5}$

$\therefore \theta_1 = \frac{5(50) - 220}{5(30) - 100} = 0.6$

$\theta_0 = \frac{22}{5} - (0.6)(2)$
$= 4.4 - 1.2$
$= 3.2$

CLASSTIME Page No.
Date / /

$\therefore$ $(0, 3.2)$
$(-5.33, 0)$



Suppose $\theta_0 = 1.4$ (1st iteration)
$\theta_0 = 5$ (2nd ")
$\theta_0 = 2$ (3rd ")

Here θ_0 is not in a particular dir^n
It keeps on moving back & forth. Thus this curve diverges.
$\therefore$ Try to reduce 'α'.

→ To plot points on graph in practical.

- plot (x, y, type = 'point') → in form of pts.

- abline (1, 1) ⟶ to plot this pt. on the line.

(1 → θ_0, 1 → θ_1)

CLASSTIME Page No:
Date / /

- Operators :-

+ → addition.
− → subtraction
* → multiplication.
%*% → matrix multiplication
/ → real division.
%/% → integer division.
^ → exponential.
%% → modulo.
& → and.
Xor → for XOR (exclusive OR).
→ comments.

- For displaying variables :-

> a = 2 // a gets value 2 but won't get displayed

>(a = 2)
2 → value
or assigned as well as displayed

> a = 2
> a → 2
> print (a) → 2

* Multiple Regression -

$$h_\theta(x^{i}) = y = \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \theta_2 x_2^{(1)} + \theta_3 x_3^{(1)}$$

eqⁿ of multiple regression.

n variables ∴ (n+1) coeff.

CLASSTIME Page No.
Date / /

Applying Gradient Descend method

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

$$\boxed{\theta_j = \theta_j - \frac{\alpha}{m}\Sigma\left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}}$$

multiple linear regression

partial derivative wrt to $\theta_0, \theta_1 - - - \theta_3$.

NOTE:

- $\boxed{\frac{x - \bar{x}}{S}}$ If this standard deviation, then it is standardization.

- If S is range, then this method becomes mean normalization.

Q Normalize given ~~of~~ data using mean normalization method

1, 2, 34, 45, 15, 40, 34.

CLASSTIME	Page No.
Date / /	

1 Feb, 19

Plotting Graphs

- $y = f(x)$

$$y' = af(bx'+c)+d \quad \text{--- (1)}$$

$$\frac{y'-d}{a} = f(bx'+c)$$

$$(x,y) \longmapsto \left(bx'+c, \frac{y'-d}{a}\right)$$

$$x = bx'+c \qquad y = \frac{y'-d}{a}$$

$\downarrow$

Finding inverse

$$x' = \frac{x-c}{b} = \frac{x}{b} - \frac{c}{b}$$

multiply by a scale of $\frac{1}{b}$.

$$\underbrace{y' = ay+d}$$

eq^n of y' has no change. It is somewhat same as (1)

$$y' = ay+d$$

Q Draw graph for $y = -2(3-2x)^2+5$.

$$\frac{y-5}{-2} = f(3-2x) \qquad \{\because f(x) = x^2\}$$

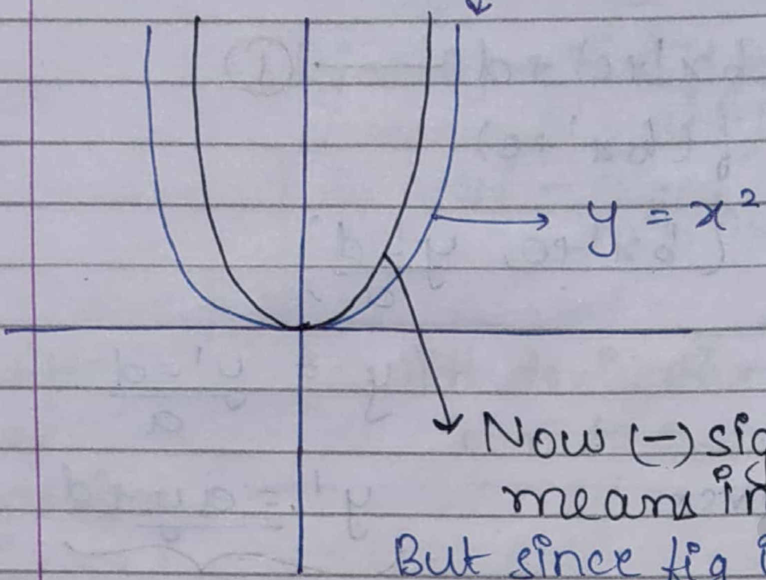
$$(x',y') \longmapsto \left(3-2x, \frac{y-5}{-2}\right)$$

$$\frac{x'-3}{-2}, \quad -2y'+5$$

$\downarrow \qquad\qquad \downarrow$

$$\frac{-x'}{2}+1.5 \qquad -2y'+5$$

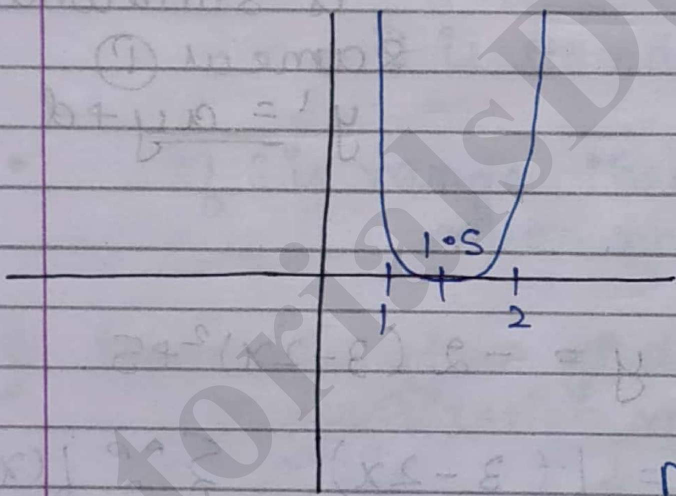
Plotting for $\frac{-x'}{2} + 1.5$



Now half the value of x i.e. reduce its width.

Now (–) sign means invert x and –x. But since fig is symmetric, it would have no effect.

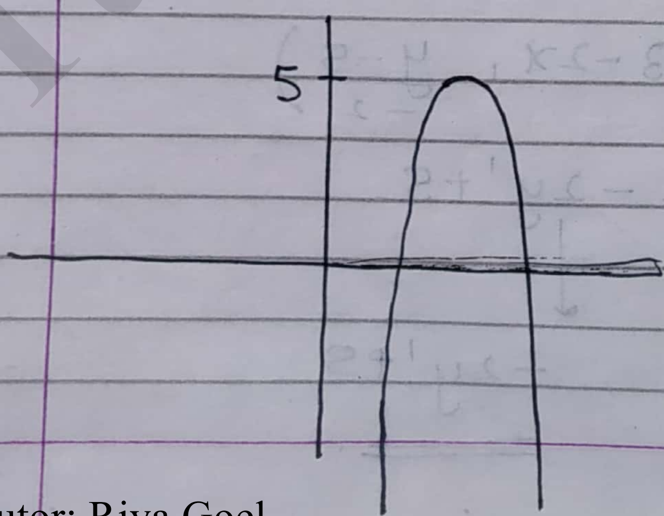
Now we perform translation of +1.5



Now for $-2y' + 5$

multiply y' by 2 and take into –ve

Multiplying y' by 2 means increasing its amplitude.



Now we need an increment of +5 in upward direction.

Logistic Regression

$$h_\theta(x) = g(\theta^T x)$$

logistic function.

sigmoid func.

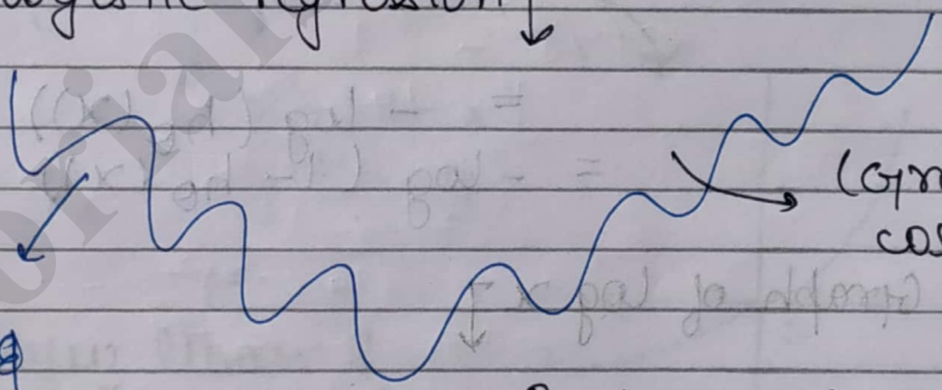
{We can't use linear regression directly ∴ we use logistic regression}.

$$g(z) = \frac{1}{1+e^{-z}} \qquad -z \rightarrow -\theta^T x$$

→ Cost function (to be minimized) → $\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

In linear regression, graph of this cost func. was a parabola where local minima = global minima using gradient Descent.

For logistic regression ↓



multiple maxima & minimas

(Graph of cost func)

∴ We define a new cost func.

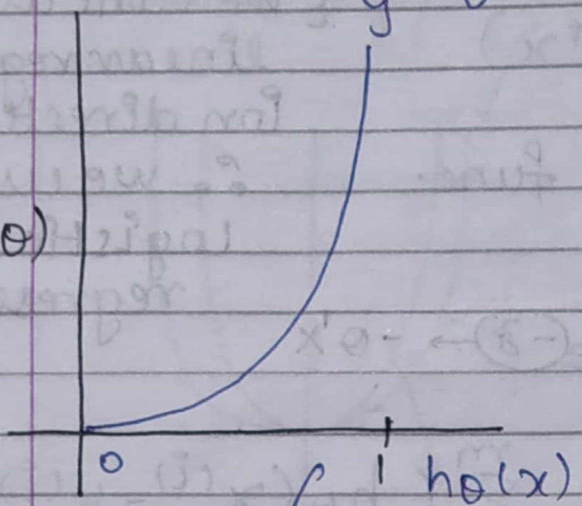
$$0 \leq h_\theta(x) \leq 1$$

(assumption)

CLASSTIME Page No.
Date / /

y = 0

J(θ)

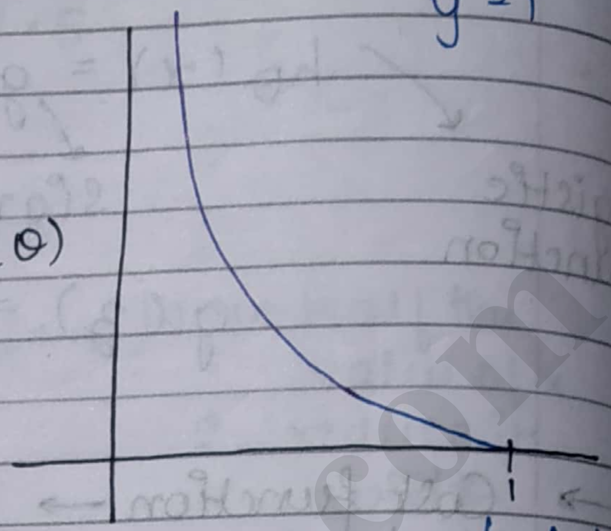


cost func./ error.

Suppose $h_\theta(x) = 1$
y = 0 ↓
predicted value
∴ error is $\underline{\infty}$

y = 1

J(θ)



Here y = 1
and $h_\theta(x) = 1$
∴ no $\underline{\text{error}}$

- $\underline{\text{Cost func}} \rightarrow \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

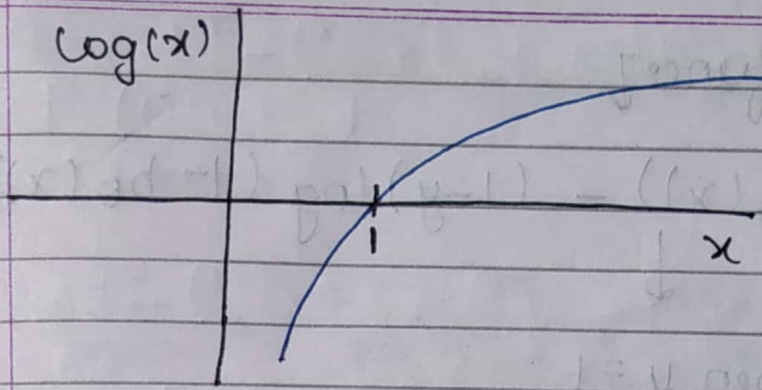
$$= -\log(h_\theta(x)) \quad \text{when } y = 1$$
$$= -\log(1 - h_\theta(x)) \quad \text{when } y = 0.$$

* Graph of log x]

$$y = \log_{10} x$$

$$10^y = x$$

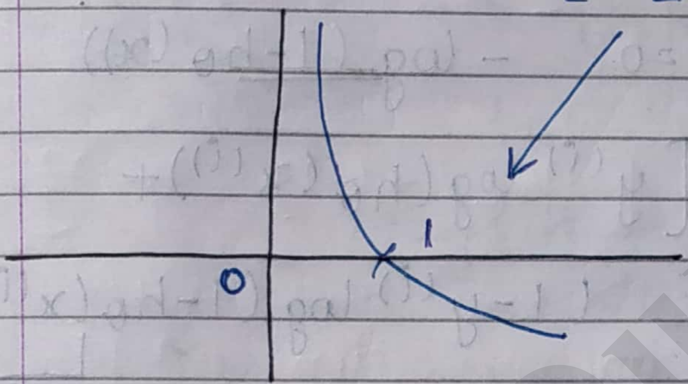
$x = 0,\ y = 1 \qquad (10, 1), (0{\cdot}1, -1)$
$(3, 1/2)$



→ $y = f(x)$

$y = -f(x)$

$= -\log x.$

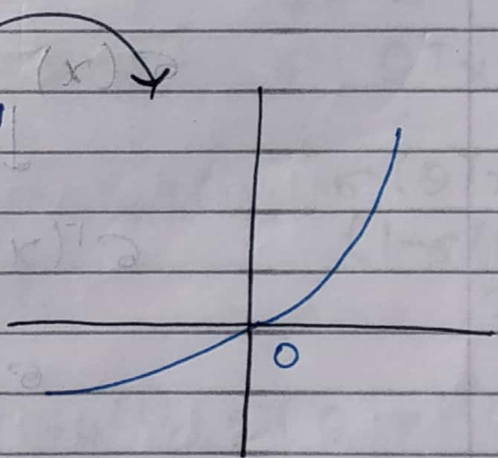


→ $y = f(x)$

$y = -f(1-x)$

$x' = 1 - x'$

$y' = -y'$



Q Diff. b/w linear & logistic regression.

Q Why linear regression can't be defined for classification.

Q What is logistic regression.

Q Derivation of cost func.

Q Numerical on logistic regression.

CLASSTIME Page No.
Date / /

* Combined cost func^n

- $-y \log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$

when $y=1$

$-\log(h_\theta(x))$ (0)

When $y=0$ $-\log(1-h_\theta(x))$

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]$$ ①

$$h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$$ sample i

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

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

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

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

$$= \sigma(x)\left[\frac{\cancel{1+e^{-x}}}{\cancel{(1+e^{-x})}} - \frac{1}{(1+e^{-x})}\right]$$

CLASSTIME Page No.
Date / /

$$\boxed{\sigma'(x) = \sigma(x)(1-\sigma(x))}$$

From ①,

$$\frac{\partial J}{\partial \theta_j} = \frac{-1}{m}\sum_{i=1}^{m} y^{(i)} \frac{\partial}{\partial \theta_j}\log(h_\theta(x^{(i)}) + (1-y^{(i)})\frac{\partial}{\partial \theta_j}\log(1-h_\theta(x^{(i)}))$$

$$= \frac{-1}{m}\sum_{i=1}^{m} \frac{y^{(i)}\frac{\partial}{\partial \theta_j}(h_\theta(x^{(i)}))}{h_\theta(x^{(i)})} + \frac{(1-y^{(i)})\frac{\partial}{\partial \theta_j}(1-h_\theta(x^{(i)}))}{1-h_\theta(x^{(i)})}$$

$\frac{d}{dx}(\log x)$

Now

$$h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}} \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$

$$\therefore h_\theta(x^{(i)}) = \sigma(\theta^T x^{(i)})$$

$$= \frac{-1}{m}\sum_{i=1}^{m} \frac{y^{(i)}\frac{\partial}{\partial \theta_j}(\sigma(\theta^T x^{(i)}))}{h_\theta(x^{(i)})} \qquad \sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)}))\, x_j^{(i)}$$

$$+ \frac{(1-y^{(i)})\,\sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)}))\, x_j^{(i)}}{1-h_\theta(x^{(i)})}$$

$$= \frac{-1}{m}\sum_{i=1}^{m} y^{(i)}(1-h_\theta(x))\,\underline{x_j^{(i)}} - (1-y^{(i)})\,h_\theta(x^{(i)})\,\underline{(x_j^{(i)})}$$

common

CLASSTIME Page No.
Date / /

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} (1 - h_\theta(x^{(i)})) - (1 - y^{(i)}) h_\theta(x^{(i)}) \right] x_j^{(i)}$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} - \cancel{y^{(i)} h_\theta(x^{(i)})} - h_\theta(x^{(i)}) + \cancel{y^{(i)} h_\theta(x^{(i)})} \right] x_j^{(i)}$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right] x_j^{(i)}$$

$$\boxed{J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right] x_j^{(i)}}$$

same as linear regression cost func.

$$\frac{1}{2m} \sum \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

upon derivative $\frac{1}{m} \sum \left(h_\theta(x^{(i)}) - y^{(i)} \right) (x_j)^i$

→ The only diff. b/w the 2 cost functions is that in linear regression, $h_\theta(x) = \theta^T x$ but in logistic regression we have $h_\theta(x) = g(\theta^T x)$

CLASSTIME Page No:
Date / /

8 Feb, 19

Q Apply logistic regression to find ↓

$\theta_0 = -2.16, \ \theta_1 = 0.425$

$\hat{Pr}(y = yes \mid x = yes)$

$Pr(y = yes \mid x = no)$ (1) (0)

$h_\theta(x) = P(y=1) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

probability of $y=1$

$= \frac{1}{1+e^{-(-2.16+0.425x)}}$ — ①

$(\theta_0 + \theta_1 x)$

$Pr(y=1 \mid x=1)$

∴ Using ①

$\frac{1}{1+e^{-(-2.16+0.425)}} = \cancel{0.85}\ \underline{0.149}$

∴ $P(y=1 \mid x=1) = \cancel{0.85}\ \underline{0.149}$

Similarly $P(yes \mid X=no)$

$P(y=1 \mid x=0)$

$\Rightarrow \frac{1}{1+e^{-(-2.16)}} = \frac{1}{1+e^{2.16}} = \underline{0.104}$

$Pr(y = no \mid x = no/yes)$

$\rightarrow 1 - P(y = yes)$

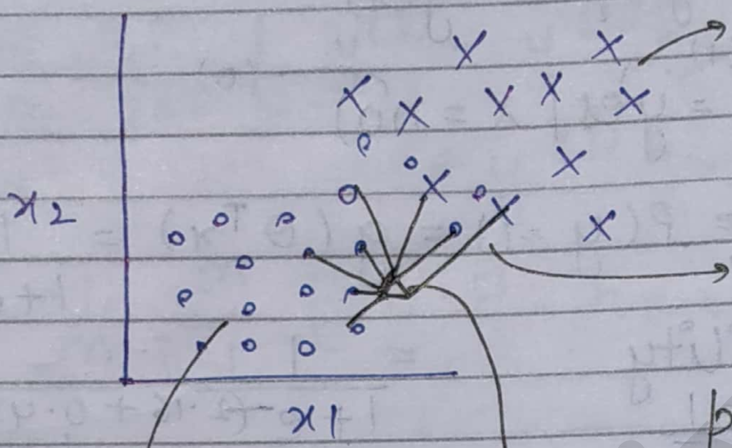
$= 1 - 0.149$

$= \underline{0.851}$

→ K means classifier (eg - for iris data)

For a data with 2 features (x_1 and x_2)

Clusters on basis of class label.



There are some outliers present as clusters are merging.

belongs to same class.

This is a new test sample

(Using decision model, we need to decide to which class it belongs)

We calculate distance of this training set from all the other objects present in data.
(using dist. formula).

Now sort all these pts. in increasing order.

Now if k = 5, then output must be 5 nearest neighbours.

Now if 4 of them belong to class 1 and 1 to class 2, then the training sample belongs to class 1.

NOTE:- Take 'k' as odd so that there is no ambiguity relating to class.

CLASSTIME Page No.
Date / /

If k = 2 (even), then 1 may belong to class 1 & other to class 2.
Then, it is difficult to decide which class does training sample belongs.