Advance Machine Learning.

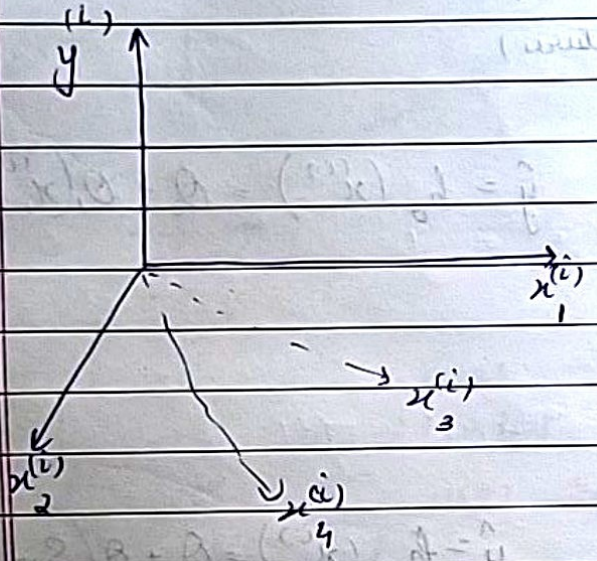
$X^{\text{example}}_{\text{feature}} = x^{(i)}_{(n)}$

i = no. of training example = no. of rows = 1 to m.

n = no. of features = no. of columns = 1 to n.

→ The more features we bring, the more complex it becomes to visualize.

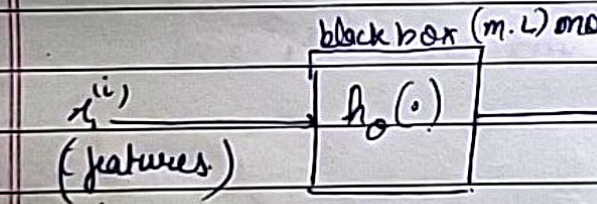
If shape of data frame = (m, n) (Actual size)



	x_1	x_2	x_3	x_4	y (true label)
①	10	20	11.7	3	7
②	8	6	7	1	11

$x^{(2)}_2 = 6$ $\quad y^{(1)} = 7$

$x^{(2)}_4 = 1$ $\quad y^{(2)} = 11$



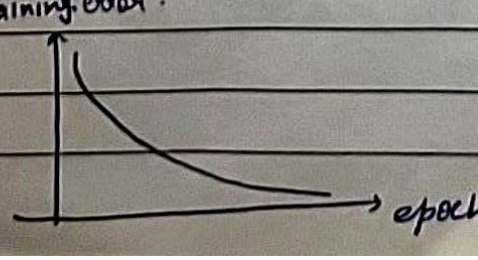
but y is the actual o/p.

we are trying to minimize the difference b/w $h_\theta(x)$ & y.

$$\text{So } L = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta x^{(i)} - y^{(i)}\right)^2 = \text{Loss func}^n = L_2 = \text{Squared Loss.}$$

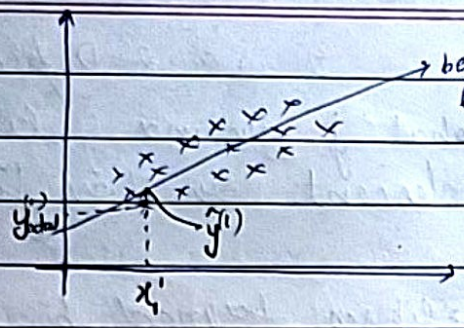
m = training examples

$$L_1 = \frac{1}{2m}\sum_{i=1}^{m}\left|h_\theta x^{(i)} - y^{(i)}\right| = L_1 \text{ loss.}$$

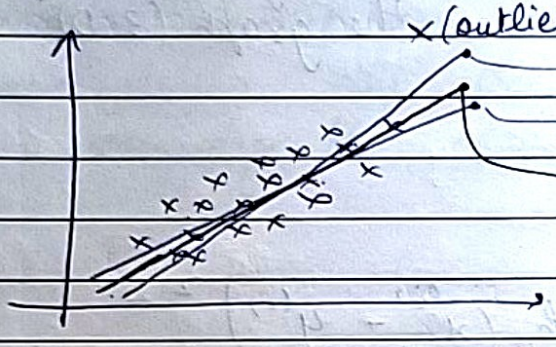


→ epoch is the no. of times we give the feature (x) to a model

→ epoch is the no. of time we made the M.L model aware of the dataset



here we can calculate L_1 or L_2 (both) to calculate the mean square error.



L_1 loss func^n uses mod func^n so the line is less deviated

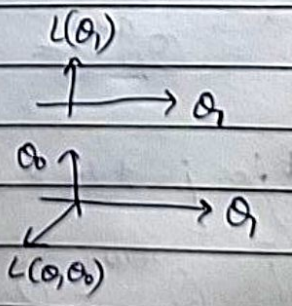
L_2 loss func^n uses square so the line is more deviated.

It is better to use L_1 but L_2 is more preferred bcz it eases our calculation of gradient descent

θ_0 & θ_1 are hyperparameters (learnable parameters) of our model.

$L(\theta_1)$ => for 2-D

$L(\theta_1, \theta_0)$ => for 3-D



Bcz we could have a problem of local minimas, so we do Regularisation.

Our main motive is to find global minima

Contour Plots :-

→ It is a 2-D plot to find global & minima by gradient descent on Loss func".

→ ellipses formed by cutting the graph (3-D)

θ_1

gradient descent

slope is high

slope is less

θ_0

$$\Rightarrow L = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)} - y^{(i)}\right)^2$$

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

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

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

$$\theta_1^{new} = \theta_1^{old} + \alpha\frac{\partial L}{\partial \theta_1}$$

Algorithm of gradient descent

for $j = 1$ to 2 {

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

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

}

α = learning rate = 0.1

RANKA
DATE / /
PAGE

Eg

x	y
1	2
2	3
3	5

$h_\theta(x^{(i)}) = y_{calculated} = \theta_0 + \theta_1 x_1^{(i)} = \hat{y}^{(i)}$ (i: 1 to 3)

$$L = \frac{1}{m} \sum_{i=1}^{3} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

Let take $\theta_0 = 0$, $\theta_1 = 0$ } initially.

$$L = \frac{1}{m} \sum_{i=1}^{3} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial L}{\partial \theta_0} = \frac{2}{m} \sum_{i=1}^{3} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)$$

$$\frac{\partial L}{\partial \theta_1} = \frac{2}{m} \sum_{i=1}^{3} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)}$$

$$\theta_0^{new} = \theta_0^{old} - 0.1 \frac{\partial L}{\partial \theta_0} \qquad \theta_1 = \theta_1 - \alpha \frac{\partial L}{\partial \theta_1}$$

Iteration

$\hat{y}^{(i)} = 0$

$$L = \frac{1}{m} \sum_{i=1}^{3} \left(\hat{y}^{(i)} - y^{(i)}\right)^2 = \frac{1}{3}\left[\left[\hat{y}^{(1)} - y^{(1)}\right]^2 + \left[\hat{y}^{(2)} - y^{(2)}\right]^2 + \left[\hat{y}^{(3)} - y^{(3)}\right]^2\right]$$

$$= \frac{1}{3}\left[(-2)^2 + (-3)^2 + (-5)^2\right] = -\frac{38}{3}$$

$$\frac{\partial L}{\partial \theta_0} = \frac{2}{3}\left[-2 + (-3) + (-5)\right] = -\frac{20}{3}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{2}{3}\left[-(2)(1) + -(2)(3) + -(3)(5)\right] = -\frac{46}{3}$$

So $\theta_0^{new} = \theta_0^{old} - \alpha \frac{\partial L}{\partial \theta_0} = 0 - (0.1)\left(\frac{-20}{3}\right) = \frac{2}{3}$

$\theta_1^{new} = \theta_1^{old} - \alpha \frac{\partial L}{\partial \theta_0} = 0 - (0.1)\left(\frac{-46}{3}\right) = \frac{4.6}{3}$

iteration 2 now $\theta_0 = \frac{2}{3} = 0.6667$ $\theta_1 = \frac{4.6}{3} = 1.533$

$\hat{y}^{(1)} = \theta_0 + \theta_1 x_1' = \frac{2}{3} + \frac{4.6}{3}(1) = \frac{6.6}{3} = 2.2$

$\hat{y}^{(2)} = \theta_0 + \theta_1 x_2^{(1)} = \frac{2}{3} + \frac{4.6}{3}(2) = \frac{11.2}{3} = 3.733$

$\hat{y}^{(3)} = \theta_0 + \theta_1 x_3^{(1)} = \frac{2}{3} + \frac{4.6}{3}(3) = \frac{15.8}{3} = 5.267$

$L = \frac{1}{3}\left[(\hat{y}^{(1)} - y^{(1)})^2 + (\hat{y}^{(2)} - y^{(2)})^2 + (\hat{y}^{(3)} - y^{(3)})^2\right]$

$= \frac{1}{3}\left[(2.2 - 2)^2 + (3.734 - 3)^2 + (5.267 - 5)^2\right]$

$= \frac{1}{3}\left[0.04 + 0.5388 + 0.0712\right] = 0.216$

$\frac{\partial L}{\partial \theta_0} = \frac{2}{3}\left[(2.2 - 2) + (3.734 - 3) + (5.267 - 5)\right]$

$= \frac{2}{3}\left[0.2 + 0.734 + 0.267\right] = 0.8$

$\frac{\partial L}{\partial \theta_1} = \frac{2}{3}\left[(2.2 - 2)(1) + (3.734 - 3)(2) + (5.267 - 5)(3)\right]$

$= \frac{2}{3}\left[0.2 + 1.468 + 0.801\right]$

$= 1.646.$

$\theta_0^{new} = \theta_0^{old} - \alpha \frac{\partial L}{\partial \theta_0}$

$= \frac{2}{3} - 0.1(0.8)$

$= 0.586$

$\theta_1^{new} = \theta_1^{old} - \alpha \frac{\partial L}{\partial \theta_1}$

$= \frac{4.6}{3} - 0.1(1.646)$

$= 1.368$

Iteration 3 $\theta_0 = 0.586$ $\theta_1 = 1.368$

$\hat{y}^{(1)} = \theta_0 + \theta_1 x_1' = 0.586 + 1.368(1) = 1.954$

$\hat{y}^{(2)} = \theta_0 + \theta_1 x_2' = 0.586 + 1.368(2) = 3.322$

$\hat{y}^{(3)} = \theta_0 + \theta_1 x_3^{(1)} = 0.586 + 1.368(3) = 4.69$

$\frac{\partial L}{\partial \theta_0} = \frac{2}{3}\left[(1.954-2) + (3.322-3) + (4.691-5)\right]$

$= \frac{2}{3}\left[-0.046 + 0.322 + -0.309\right]$

$= -0.022$

$\frac{\partial L}{\partial \theta_1} = \frac{2}{3}\left[(1.954-2)1 + (3.322-3)2 + (4.69-5)3\right]$

$= \frac{2}{3}\left[(-0.046)1 + (.322)2 + (-0.309)3\right]$

$= -0.2817$

$\theta_0^{new} = \theta_0^{old} - \alpha \frac{\partial L}{\partial \theta_0}$

$= 0.586 - (0.1)0.022$

$= 0.5891$

$\theta_1^{new} = \theta_1^{old} - \alpha \frac{\partial L}{\partial \theta_1}$

$= 1.368 - (0.1)0.217$

$= 1.389$

hyperparameters → These are variable which are in your hand, we can change is value on the basis of trial & error.
Eg α & θ_j

Dependency of m being large.

⇒ Matrix form of linear Regression (Vectorisation)

X = design matrix $\quad X \in \mathbb{R}^{m \times (n+1)}$

θ = 1-D vector $\quad \theta \in \mathbb{R}^{(n+1) \times 1}$

Y = 1-D vector $\quad Y \in \mathbb{R}^{m \times 1}$

One way to calculate optimal θ →

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

$$\theta_0 := \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_0^{(i)}$$

$$L = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2 = \frac{1}{2m} \|X\theta - Y\|_2^2$$

$$= \frac{1}{2m} (X\theta - Y)^T (X\theta - Y)$$

$$= \frac{1}{2m} \left[(X\theta)^T - (Y)^T\right] (X\theta - Y)$$

$$= \frac{1}{2m} \left[\theta^T X^T - Y^T\right] (X\theta - Y)$$

$$= \frac{1}{2m} \left[\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y\right]$$

$$L = \frac{1}{2m} \left[\theta^T X^T X\theta - 2\theta^T X^T Y + Y^T Y\right]$$

Another way to calculate optimal θ (differentiating & equating to zero) →

$$\frac{\partial L}{\partial \theta} = \frac{2}{2m} \left[X^T X\theta - X^T Y\right] = 0$$

$$X^T X\theta - X^T Y = 0$$

θ is a column vector.

RANKA
DATE / /
PAGE

$$X^T X \theta = X^T Y$$

Pre multiply with $(X^T X)^{-1}$

$$\underbrace{(X^T X)^{-1} X^T X \theta}_{\text{optimal } \theta} = (X^T X)^{-1} X^T Y$$

Pseudo inverse $\Longrightarrow$

$$\theta^* = \underbrace{(X^T X)^{-1}}_{(n+1)\times(n+1)} \underbrace{X^T}_{(n+1)\times m} Y \quad (m\times 1)$$

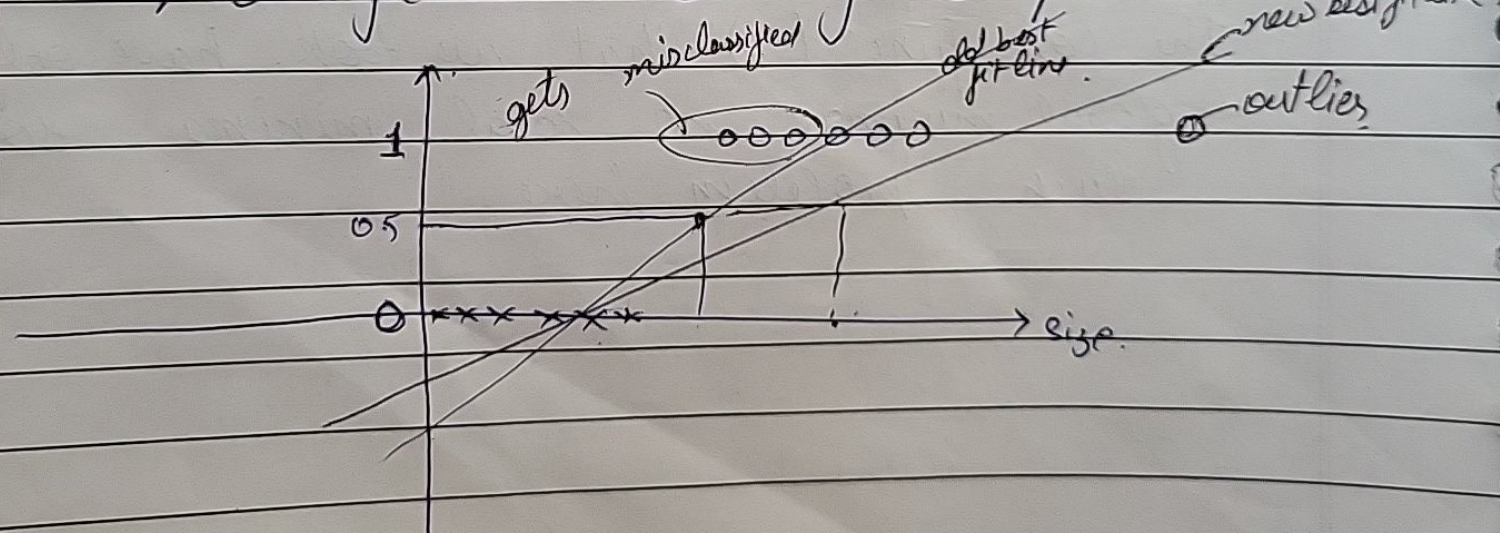
Normal eqn.

Classification :-

⇒ Logistic Regression

Y (output) will have discrete values.

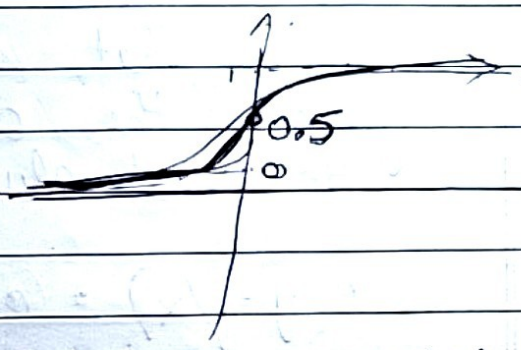
If we use linear regression in classification & use threshhold value then even one outlier will cause a big problem & change the class of the points.



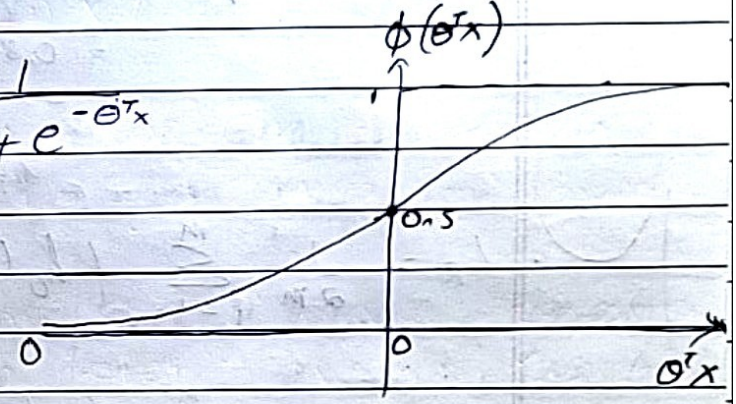
$h_\theta(x)$ — range in linear regression = $-\infty$ to ∞

$h_\theta(x)$ — ~~range~~ in classification problem = 0 & 1
└ can be done using sigmoid func^n.
~~Type of~~ activation func^n.

$\phi(z)$ = Sigmoid func^n = $\frac{1}{1+e^{-z}}$



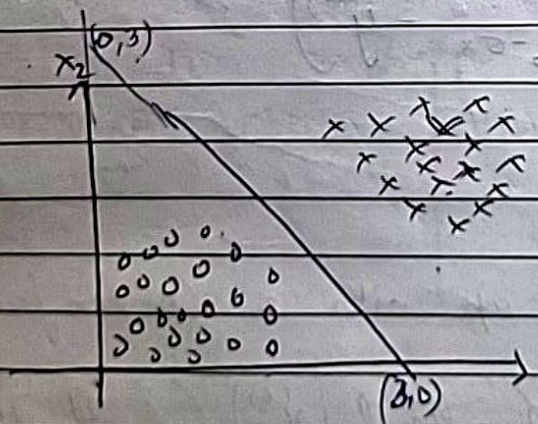
$h_\theta(x) = \phi(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$



$\theta^T x \geq 0$
$\Rightarrow \phi(\theta^T x) \geq 0.5$

decision boundary

$\theta^T x < 0$
$\Rightarrow \phi(\theta^T x) < 0.5$



Linear regression won't work here for classification.

eq^n $h_\theta(x) \Rightarrow \phi(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

By gradient descent we found out $\theta_0 = -3$ $\theta_1 = 1$ $\theta_2 = 1$

$h_\theta(x) = \phi(-3 + 1x_1 + 1x_2)$

$-3 + x_1 + x_2 > 0 \Rightarrow y = 1$

$-3 + x_1 + x_2 < 0 \Rightarrow y = 0$

The model that we'll be training will be giving us the output in terms of probability. It won't always be 0 or 1.

$P(y=1/x) = 0.95 \Rightarrow$ 95% chance of $y=1$, 5% chance of $y=0$

$= 0.88 \Rightarrow$ 88% chance of $y=1$, 12% chance of $y=0$

$L_R = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2$ = loss func^n for linear regression

$J_\theta = \frac{1}{2m} \sum_{i=1}^{m} (\phi(\theta^T x) - y^{(i)})^2$ = loss func^n for classification (logistic regression)

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

many local minimas & non-differentiable. & Non-convex.

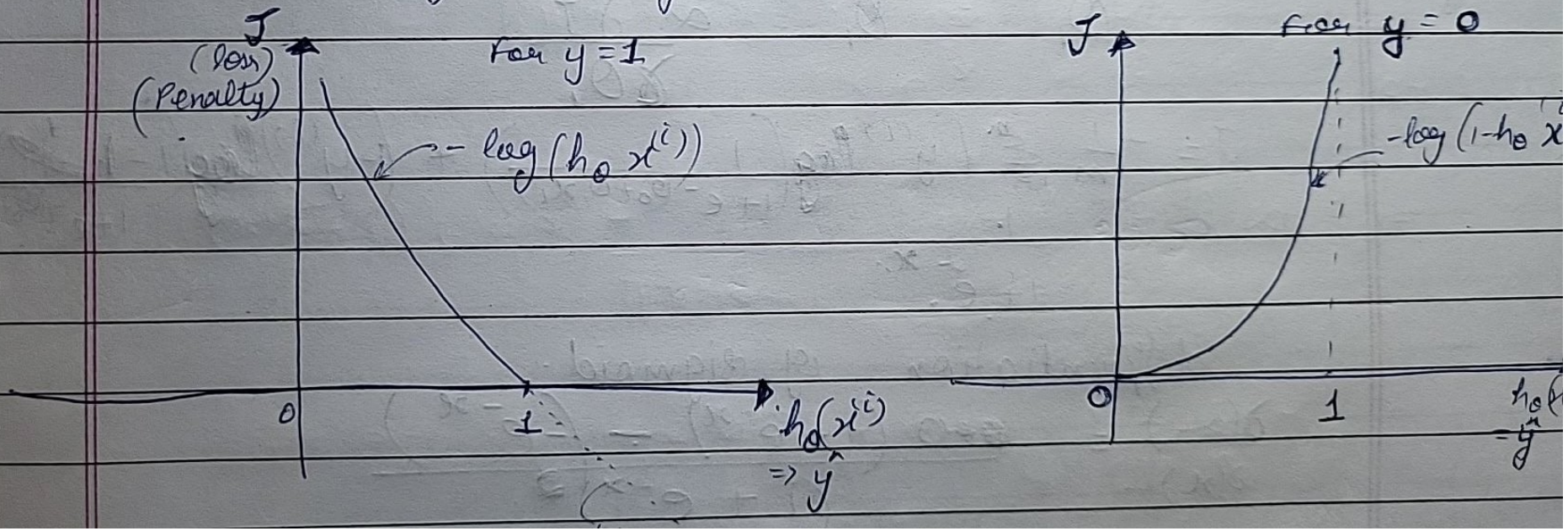
Sol^n ⇒ to come up with a Convex func^n.

$$Cost(h_\theta x^{(i)}, y^{(i)}) = \begin{cases} -\log h_\theta(x^{(i)}) & \text{if } y^{(i)} = 1 \\ -\log(1 - h_\theta(x^{(i)})) & \text{if } y^{(i)} = 0 \end{cases}$$

so that curve can be descending & not ascending for gradient descent

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

→ Binary cross entropy loss.
used for classification problem



Performing gradient descent on J

i.e $\theta_0^{new} = \theta_0^{old} - \alpha \frac{\partial J}{\partial \theta_0}$

$$\theta_1^{new} = \theta_1^{old} - \alpha \frac{\partial J}{\partial \theta_1}$$

$$J = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_i)}}\right) + (1-y^{(i)}) \log\left(1 - \frac{1}{1+e^{-(\theta_0+\theta_1 x_i)}}\right) \right]$$

$$\Rightarrow a^{(i)} = \frac{1}{1+e^{-x_i}}$$

differentiation of sigmoid.

$$\frac{\partial(a)}{\partial(x)} = \frac{0\,(1+e^{-x}) - (-e^{-x})}{(1+e^{-x})^2}$$

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

do +1 -1 to solve

$$= a(x)\,(1 - a(x))$$

$$\frac{\partial J}{\partial \theta_0} = -\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}} \left(1 - \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}\right) + \right.$$

$$\left. (1-y^{(i)}) \left(1 - \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}\right) \left(1 - \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}\right) \right)$$

$$\frac{\partial J}{\partial \theta_1} = -\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}} \left(1 - \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}\right) \right.$$

$\log x$

$\log x^{-1}$

$\log\left(1 - \frac{1}{x}\right)$

$\log\left(\frac{x-1}{x}\right)$

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

$$\frac{\partial J}{\partial \theta_0} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\frac{1}{a(x^{(i)})}\cdot a(x^{(i)})\left(1-a(x^{(i)})\right) + \left(1-y^{(i)}\right)\frac{1}{1-a(x^{(i)})}\cdot -\left(a(x^{(i)})\left(1-a(x^{(i)})\right)\right)\right]$$

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

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

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

$$\frac{\partial J}{\partial \theta_1} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\frac{1}{a(x^{(i)})}\cdot a(x^{(i)})\left(1-a(x^{(i)})\cdot x_1 + \left(1-y^{(i)}\right)\frac{1}{1-a(x^{(i)})}\cdot -\left(a(x^{(i)})\left(1-a(x^{(i)})\cdot x_1^{(i)}\right)\right)\right]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}x_1 - a(x^{(i)})y^{(i)}x_1 + -a(x^{(i)})x_1 + y^{(i)}a(x^{(i)})x_1\right]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[x_1\left(y^{(i)} - a(x^{(i)})\right)\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left[\left(a x_1^{(i)} - y^{(i)}\right)x_1\right]$$

$\Rightarrow$ The eq^{n} of $\frac{\partial L}{\partial \theta_0}$ & $\frac{\partial L}{\partial \theta_1}$ are same as that of linear.

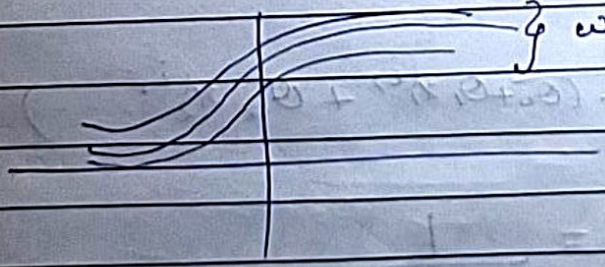
regression, the only difference is the eq" of $h_\theta(x)$ or $a(x)$

$h_\theta(x) = \theta_0 + \theta_1 x_1$ = for linear regression

$h_\theta(x) = a_\theta(x) = \frac{1}{1+e^{-(\theta^T x)}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}$ =

for logistic regression

θ as a subscript bcz it is parameterised wrt θ meaning with varying θ the value of $a(x)$ will change.



$\Rightarrow$ with varying θ the value of sigmoid func changes

$\Rightarrow$ Generalised $\frac{\partial L}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$

$x_0^{(i)} = 1$

DATE / /
PAGE

Vectorisation

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

$y \in \mathbb{R}^{m \times 1}$

$\hat{y} \in \mathbb{R}^{m \times 1}$

$X \in \mathbb{R}^{m \times (n+1)}$

$\theta \in \mathbb{R}^{(n+1) \times 1}$

$$h_\theta x^{(i)} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} \cdots)}}$$

$$\hat{y} = h_\theta(x) = \frac{1}{1 + e^{-X\theta}}$$

$$J = -\frac{1}{m} \left[y^T \log(\hat{y}) + (1 - y)^T \log(1 - \hat{y}) \right]$$

$1 \times m$ $m \times 1$ $1 \times m$ $m \times 1$

due to Python broadcasting it will itself convert to $1 \times m$ $m \times 1$.

1×1

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

$$\frac{\partial L}{\partial \theta_j} = \frac{1}{m} \left[X^T \left(\hat{Y} - Y \right) \right]$$

$m \times 1$ $\quad$ $m \times 1$

$(n+1) \times 1$ $\quad$ $(n+1) \times m$

$(n+1) \times m$

This will calculate $\frac{\partial L}{\partial \theta_j}$ in one shot.

$$\underset{(n+1)\times 1}{\theta^{new}} = \underset{(n+1)\times 1}{\theta^{old}} - \frac{\alpha}{m} X^T \left(\hat{Y} - Y \right)$$

Eg Predict Pass / Fail based on no. of hours of study

no. of hours	Outcome
1	0
2	0
3	0
4	1
5	1
6	1

Features = no. of hours = $n = 1$

no. of samples = $m = 16$

True label = Y = outcome.

$$h_\theta(x^{(i)}) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1^{(i)})}} = \hat{y}^{(i)}$$

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

$$J = -\frac{1}{6}\sum_{i=1}^{6}\left[y^{(i)}\log(\hat{y})^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right]$$

$$\theta_j^{new} = \theta_j^{old} - \frac{\alpha}{6}\sum_{i=1}^{6}(\hat{y}^{(i)} - y^{(i)})\,x_j^{(i)}$$

eration I

$\theta_0 = \theta_1 = 0$

$\alpha = 0.1$

$$\hat{y} = h_\theta x^{(i)} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1^{(i)})}}$$

$$\Rightarrow \hat{y}_1 = \frac{1}{1+e^{0}} = \frac{1}{2} = 0.5$$

(Predicted probability. So using threshold we can classify it as 0.)

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{6}\left[(0.5-0)+(0.5-0)+(0.5-0)+(0.5-1)+(0.5-1)+(0.5-1)\right]$$

$$= \frac{1}{6}(0) = 0$$

$$\frac{\partial J}{\partial \theta_1} = \frac{1}{6}\left[(0.5-0)*1+(0.5-0)*2+(0.5-0)*3+(0.5-1)*4+(0.5-0)*5+(0.5-0)*6\right]$$

$$= \frac{1}{6}\left[0.5+1+1.5+2+(-2.5)-3\right]$$

$$= \frac{1}{6}\left[-4.5\right] = -0.75$$ Ans.

$\theta_0^{new} = \theta_0^{old} - \frac{0.1}{6}(0) = 0$

$\theta_1^{new} = \theta_1^{old} - \frac{0.1}{6}(-0.75) = 0.075$

Iteration 2 $\quad \theta_0 = 0$

$\theta_1 = 0.075$

$\alpha = 0.1$

$\hat{y}_i = \sigma(x_i) = \frac{1}{1+e^{-(0.075)x_i}} \quad x_i = 1 \text{ to } 6$

$\hat{y}_1 = \sigma(x_1) = \frac{1}{1+e^{-(0.075)}} = 0.518$

$\hat{y}_2 = \sigma(x_2) = \frac{1}{1+e^{-(0.075)2}} = 0.537$

$\hat{y}_3 = \sigma(x_3) = \frac{1}{1+e^{-(0.075)3}} = 0.556$

$\hat{y}_4 = \sigma(x_4) = \frac{1}{1+e^{-(0.075)4}} = 0.574$

$\hat{y}_5 = \sigma(x_5) = \frac{1}{1+e^{-(0.075)5}} = 0.592$

$\hat{y}_6 = \sigma(x_6) = \frac{1}{1+e^{-(0.075)6}} = 0.610$

$\frac{\partial J}{\partial \theta_0} = \frac{1}{6}\left[(0.518-0) + (0.537-0) + (0.556-0) + (0.574-0.426) + (0.592-1) + (0.610-1)\right]$

$= \frac{1}{6}[0.387] = 0.0645$

$$\frac{\partial J}{\partial \theta_1} = \frac{1}{6}\Big[(0.518 * 1) + (0.537 * 2) + (0.556 * 3) + (-0.426 * 4) + (-0.408) * 5 + (-0.39 * 6)\Big]$$

$$= -0.470$$

$$\theta_0^{new} = \theta_0^{old} - 0.1\,(0.0645) = -(0.00645)$$

$$\theta_1^{new} = \theta_1^{old} - 0.1\,(-0.470) = 0.075 -$$

Gradient descent on regularised eq^n:

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

$$\theta_j^{new} = \left(\theta_j^{old} - \frac{\alpha\lambda}{m}\theta_j^{old}\right) - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta x^{(i)} - y^{(i)}\right)x_j^{(i)}$$

$$\theta_j^{new} = \theta_j^{old}\left(1 - \frac{\alpha\lambda}{m}\right) - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta x^{(i)} - y^{(i)}\right)x_j^{(i)}$$

$$\theta_j^{new} = \beta\,\theta_j^{old} - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta x^{(i)} - y^{(i)}\right)x_j^{(i)}$$

β → extra weight that is added to the θ_j^{old}.

$\beta < 1$

$\because \alpha$ is very small.

→ Normal Eq^n modified (Regularised)

$$J = \left\{\frac{1}{2m}(X\theta - Y)^T(X\theta - Y) + \frac{\lambda}{2m}\theta^T I\theta\right\}$$

$$\frac{\partial J}{\partial\theta} = \left[\frac{1}{m}X^T(X\theta - Y) + \frac{\lambda}{m}I\theta\right] = 0$$

$$\frac{1}{m}\left(X^TX\theta - X^TY + \lambda I\theta\right) = 0$$

$$X^TX\theta - X^TY + \lambda I\theta = 0$$

$$(X^TX + \lambda I)\theta = X^TY$$

$$\theta = \frac{X^TY}{X^TX + \lambda I} \Rightarrow \theta = (X^TX + \lambda I)^{-1}X^TY$$

$$\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y$$

for $\lambda = 0$

$\hat{\theta}$ = eq of ordinary (normal) linear regression

for $\lambda \neq 0$

$\hat{\theta}$ = Regularized θ

$\Rightarrow$ For Logistic Regression

$$J_{\text{Regularised}} = \frac{1}{m}\left[y^T (\log \hat{y}) + (1-y)^T \log(1-\hat{y}) \right] + \frac{\lambda}{m} \theta^T I \theta$$

$$\hat{y} = \frac{1}{1+e^{-x\theta}}$$

$$\frac{\partial J_R}{\partial \theta} =$$

ularised gistic kegression

$$J = \theta^* = \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)}) \right] + \frac{\lambda}{m} \sum_{j=1} \theta_j^2$$

Gradient descent on regularised eq^n:

$$\theta_j^{\text{new}} = \theta_j^{\text{old}}$$

$$\frac{\partial J}{\partial \theta} = \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \frac{1}{\hat{y}^{(i)}} \hat{y}^{(i)}(1-\hat{y}^{(i)}) + \left(1-y^{(i)}\right) \frac{1}{(1-\hat{y}^{(i)})} (-\hat{y}^{(i)})(1-\hat{y}^{(i)}) \right] + \frac{2\lambda}{m}(\theta_j)$$