

Advance Machine Learning.

RANKA
DATE / /
PAGE

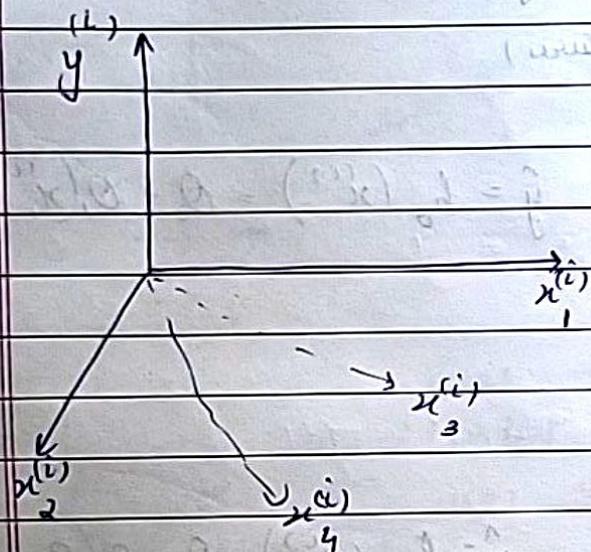
$$x_{\text{feature}}^{\text{example}} = x_{(n)}^{(i)}$$

$i = \text{no. of training example}$
 $= \text{no. of rows} = 1 \text{ to } m$.

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

$n = \text{no. of features}$
 $= \text{no. of columns} = 1 \text{ to } n$.

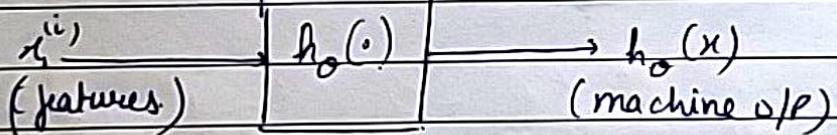
∴ Shape of data frame = (m, n)
(Actual Price)



	x_1	x_2	x_3	x_4	y
①	10	20	11.7	3	7
②	8	6	7	1	11

$$\begin{aligned} x_2^{(2)} &= 6 & y_0^{(1)} &= 7 \\ x_4^{(2)} &= 1 & y_0^{(2)} &= 11 \end{aligned}$$

block box (m.l) model



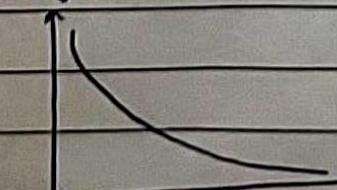
but y is the actual o/p.

we are trying to minimize the difference b/w $h_0(x)$ & y .

$$\text{So } L = \frac{1}{2m} \sum_{i=1}^m (h_0(x^{(i)}) - y^{(i)})^2 = \text{Loss func} = L_2 = \text{Quadratic Loss.}$$

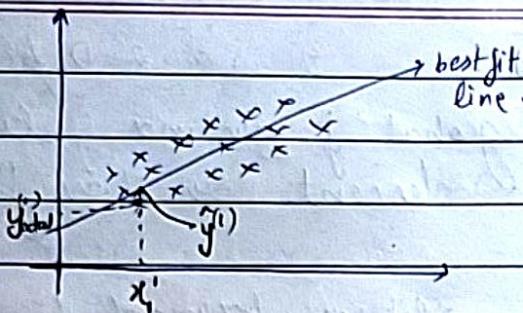
$$L_1 = \frac{1}{2m} \sum_{i=1}^m |h_0(x^{(i)}) - y^{(i)}| = L_1 \text{ loss.}$$

training error.

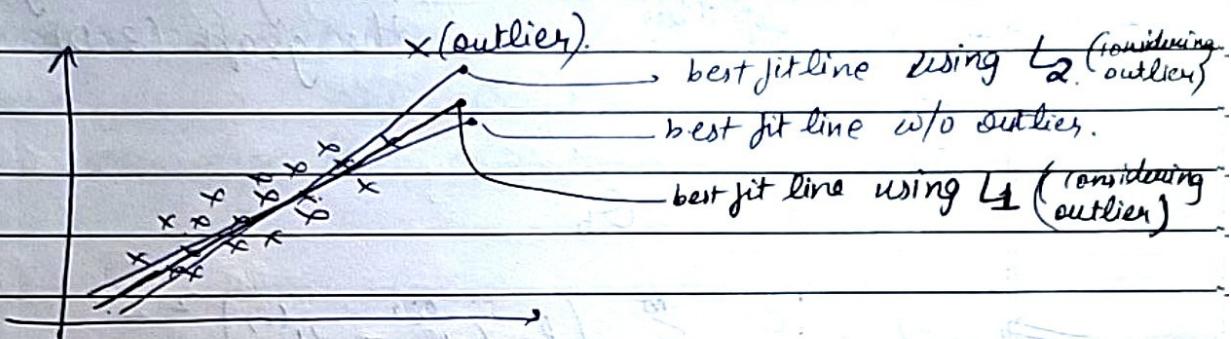


→ epoch is the no. of times we give the feature (x) to a model

→ epoch is the no. of times 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" uses mod func" so the line is less deviated

L_2 loss func" 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.

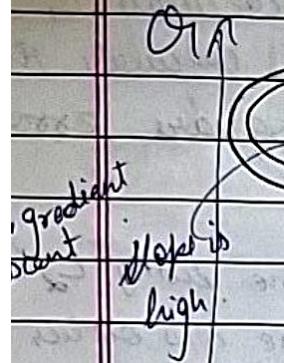
$$\begin{aligned} L(\theta_1) &\rightarrow \text{for } 2-D & \uparrow & \rightarrow \theta_1 \\ L(\theta_1, \theta_0) &\rightarrow \text{for } 3-D & \uparrow & \rightarrow \theta_0 \\ && \uparrow & \rightarrow \theta_0 \\ && L(\theta_0, \theta_1) & \end{aligned}$$

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

Our main motive is to find global minima

Contour Plots :- \rightarrow It is a 2-D plot to find.

global & minima by gradient descent on loss func.



Ellipses formed by cutting the graph (3-D)

$$\Rightarrow L = \frac{1}{2m} \sum_{i=1}^m (h_0(x^{(i)} - y^{(i)})^2$$

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

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

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

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

gradient for $j = 1 \text{ to } 2, 0$

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

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

}

$\alpha = \text{learning rate} = 0.1$

RANKA

DATE / /

PAGE

x	y	
1	2	
2	3	
3	5	

$h(x^{(i)}) = y_{\text{calculated}} = \theta_0 + \theta_1 x_1^{(i)} = \hat{y}^{(i)}$

↑ to 3

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

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

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

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

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

$$\theta_0^{\text{new}} = \theta_0 - \cancel{0.1} \frac{\partial L}{\partial \theta_0} \quad \theta_1^{\text{new}} = \theta_1 - \cancel{0.1} \frac{\partial L}{\partial \theta_1}$$

Iteration

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)}$$

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

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

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

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

$$S_0 \quad \theta_0^{\text{new}} = \theta_0^{\text{old}} - \alpha \frac{\partial L}{\partial \theta_0} = 0 - (0.1) \left(-\frac{2}{3} \right) = \frac{2}{3}$$

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

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

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

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

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

$$\begin{aligned} L &= \frac{1}{3} \left[(\hat{y}^{(1)} - y^{(1)})^2 + (\hat{y}^{(2)} - y^{(2)})^2 + (\hat{y}^{(3)} - y^{(3)})^2 \right] \\ &= \frac{1}{3} \left[(2.2 - 2)^2 + (3.733 - 3)^2 + (5.267 - 5)^2 \right] \\ &= \frac{1}{3} (0.04 + 0.5388 + 0.0712) = 0.216 \end{aligned}$$

$$\frac{\partial L}{\partial \theta_0} = \frac{2}{3} \left[(2.2 - 2) + (3.733 - 3) + (5.267 - 5) \right]$$

$$\frac{\partial L}{\partial \theta_0} = \frac{2}{3} [0.2 + 0.734 + 0.267] = 0.8$$

$$\begin{aligned} \frac{\partial L}{\partial \theta_1} &= \frac{2}{3} \left[(2.2 - 2)(1) + (3.733 - 3)(2) + (5.267 - 5)(3) \right] \\ &= \frac{2}{3} [0.2 + 1.468 + 0.801] \\ &= 1.646 \end{aligned}$$

$$\begin{aligned}\theta_0^{\text{new}} &= \theta_0^{\text{old}} - \alpha \frac{\partial L}{\partial \theta_0} \\ &= \frac{2}{3} - 0.1(0.8) \\ &= 0.586.\end{aligned}$$

$$\begin{aligned}\theta_1^{\text{new}} &= \theta_1^{\text{old}} - \alpha \frac{\partial L}{\partial \theta_1} \\ &= \frac{4.6}{3} - 0.1(1.646) \\ &= 1.368.\end{aligned}$$

Iteration 3

$\theta_0 = 0.586$

$\theta_1 = 1.368$

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

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

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

$$\begin{aligned}\frac{\partial L}{\partial \theta_0} &= \frac{2}{3} \left[(1.954 - 2) + (3.322 - 3) + (4.69 - 5) \right] \\ &= \frac{2}{3} \left[-0.046 + 0.322 + -0.309 \right] \\ &= -0.022\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial \theta_1} &= \frac{2}{3} \left[(1.954 - 2)1 + (3.322 - 3)2 + (4.69 - 5)3 \right] \\ &= \frac{2}{3} \left[(-0.046)1 + (3.22)2 + (-0.309)3 \right] \\ &= -0.2817\end{aligned}$$

$$\begin{aligned}\theta_0^{\text{new}} &= \theta_0^{\text{old}} - \alpha \frac{\partial L}{\partial \theta_0} \\ &= 0.586 - 0.1(-0.022) \\ &= 0.5891\end{aligned}$$

$$\begin{aligned}\theta_1^{\text{new}} &= \theta_1^{\text{old}} - \alpha \frac{\partial L}{\partial \theta_1} \\ &= 1.368 - 0.1(-0.2817) \\ &= 1.389\end{aligned}$$

hyperparameters \rightarrow These are variable which are in your hand, we can change its value on the basis of trial & error.

Eg θ_0 & θ_1

for "n" of m being large.

\Rightarrow Matrix Form of linear regression. (Vectorisation)

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

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

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

one way to calculate primal θ

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

$$L = \frac{1}{2m} \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)}))^2 = \frac{1}{2m} \|X\theta - y\|_2^2$$

$$= \frac{1}{2m} (x\theta - y)^T (x\theta - y)$$

$$= \frac{1}{2m} [(x\theta)^T - (y)^T] (x\theta - y)$$

$$= \frac{1}{2m} [\theta^T x^T - y^T] [x\theta - y]$$

$$= \frac{1}{2m} [\theta^T x^T x\theta - \theta^T x^T y - x^T x\theta + y^T y]$$

$$L = \frac{1}{2m} [\theta^T x^T x\theta - 2\theta^T x^T y + y^T y]$$

other way
calculate
primal θ
vectorizing &
gradient to zero

$$\frac{\partial L}{\partial \theta} = \frac{1}{2m} [x^T x\theta - 2x^T y] = 0$$

$$x^T x\theta - x^T y = 0$$

θ is a column vector.

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

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

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

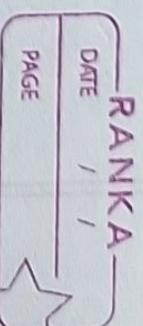
optimal θ

by normal eqn.

$$\text{Pseudo inverse} \Rightarrow \theta^* = (X^T X)^{-1} X^T Y$$

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

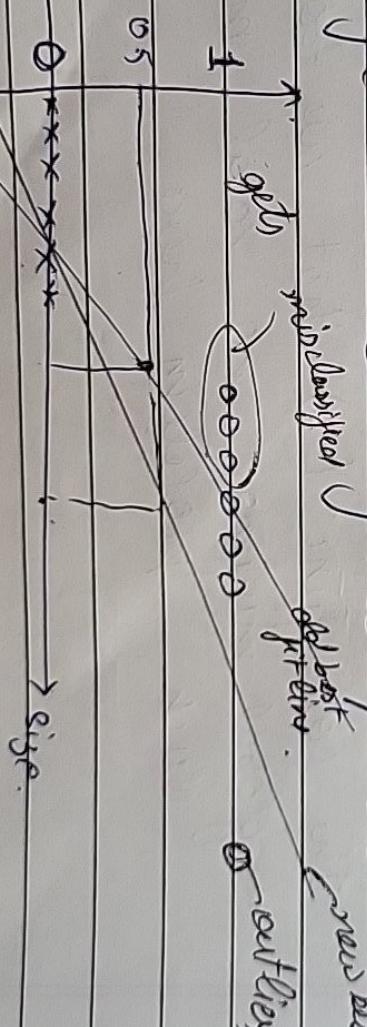
$(m+1) \times m$



#

Classification :-
⇒ Logistic Regression
 y (output) will have discrete values.

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



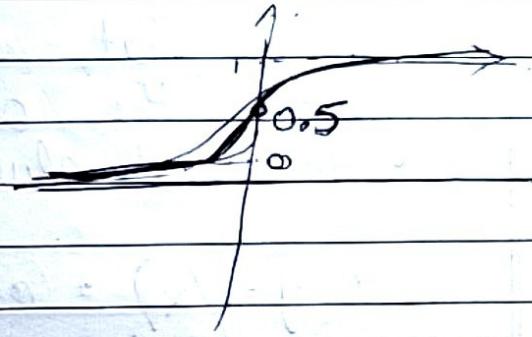
$h_0(x)$ - range in linear regression = $-\infty$ to ∞

$h_0(x)$ - range in classification problem = 0 & 1

(can be done using sigmoid funcⁿ)

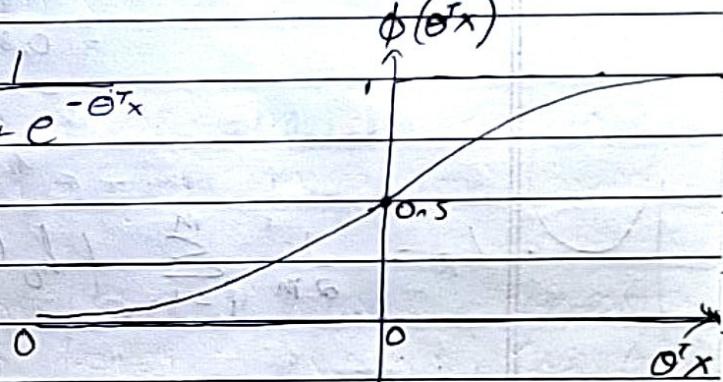
"of activation funcⁿ".

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



$$h_0(x) = \phi(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

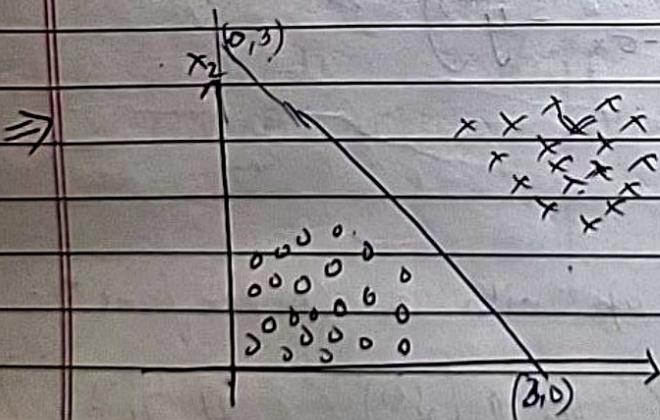
$$\phi(\theta^T x)$$



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

decision boundary

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



linear regression won't work here for classification.

$$\text{g}' h_0(x) \Rightarrow \phi' (\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

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

$$h_0(x) = \phi(-3 + 1x_1 + 1x_2)$$

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

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

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

$$\begin{aligned} P(y=1/x) &= 0.95 \Rightarrow 95\% \text{ chance of } y=1 \\ &= 0.88 \Rightarrow 88\% \text{ chance of } y=1 \\ &\quad 12\% \text{ chance of } y=0 \end{aligned}$$

$$L_R = \frac{1}{2m} \sum_{i=1}^m (h_0(x) - y^{(i)})^2. \quad \begin{array}{l} \text{loss func'' for} \\ \text{linear regression} \end{array}$$

$$\begin{aligned} L_C &= \frac{1}{2m} \sum_{i=1}^m (\phi(\theta^T x) - y^{(i)})^2. \quad \begin{array}{l} \text{loss func'' for} \\ \text{classification (logistic regression)} \end{array} \\ &= \frac{1}{2m} \sum_{i=1}^m \left(\frac{1}{1+e^{-\theta^T x}} - y^{(i)} \right)^2. \end{aligned}$$

many local minima.
not differentiable.
non-convex.

"Sol" \Rightarrow to come up with a convex func".

$$\text{cost}(h_0(x^{(i)}), y^{(i)}) = \begin{cases} -\log h_0(x^{(i)}) & \text{if } y^{(i)} = 1 \\ -\log (1 - h_0(x^{(i)})) & \text{if } y^{(i)} = 0 \end{cases}$$

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

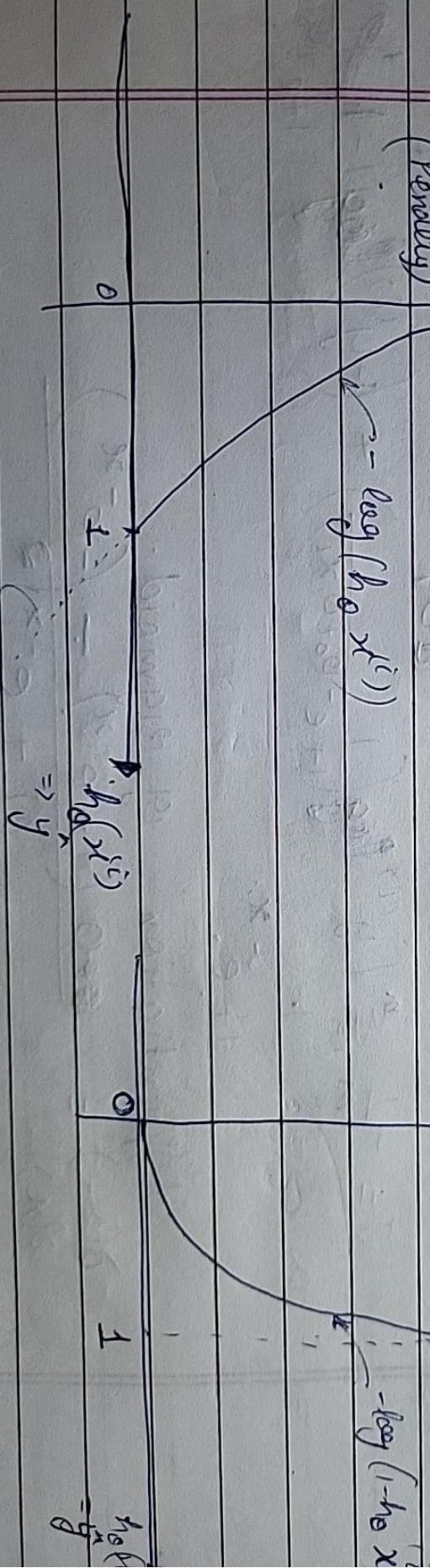
for $y = 1$

$$J \uparrow \text{(Penalty)}$$

$$-\log(h_\theta(x^{(i)}))$$

for $y = 0$

$$-\log(1-h_\theta(x^{(i)}))$$



Performing gradient descent on J
 i.e. $\theta_0^{\text{new}} = \theta_0^{\text{old}} - \alpha \frac{\partial J}{\partial \theta_0}$

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

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

$$\Rightarrow \alpha \theta_0^{(t+1)} = \theta_0^{(t)} - \frac{x_i}{1+e^{-x_i}}.$$

Differentiation of Sigmoid:

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

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

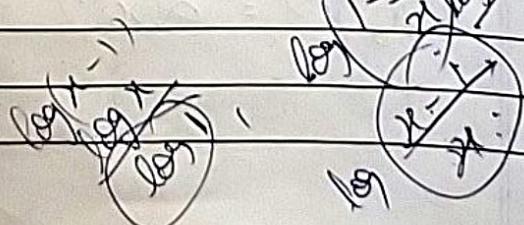
$\cancel{a(x) + 1}$
 $\cancel{to \quad see}$

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

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

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

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



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

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

$$(1-y^{(i)}) \frac{1 \cdot - (a(x^{(i)}) (1-a(x^{(i)})))}{1-a(x^{(i)})}$$

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

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

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

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

$$(1-y^{(i)}) \frac{1 \cdot - (a(x^{(i)}) (1-a(x^{(i)})) \cdot x_1)}{1-a(x^{(i)})}$$

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

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

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

\Rightarrow The eqⁿ of $\frac{\partial J}{\partial \theta_0}$ & $\frac{\partial J}{\partial \theta_1}$ are same as that of linear

regression, the only difference is the eqⁿ of $h_{\theta}(x)$ or $\alpha(x)$

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 \quad \text{for linear regression}$$

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

for logistic regression

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

with varying θ
the value of sigmoid func changes

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

$$x_0^{(i)} = 1$$

Vectorisation

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

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

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

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

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

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

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

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

$\hat{y}^{m \times 1}$ $(1-\hat{y})^{m \times 1}$ $m \times 1$

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

1×1

$$\frac{\partial L}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i) x_i^{(j)}$$

$$\frac{\partial L}{\partial \theta_j} = \frac{1}{m} \left[\underbrace{x^T}_{m \times 1} \underbrace{(y - \hat{y})}_{m \times 1} \right]$$

$(n+1) \times m$

$(n+1) \times m$

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

$$\theta^{new} = \theta^{old} - \alpha \frac{1}{m} x^T (y - \hat{y})$$

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

no. of hours Outcomes

1	0
2	0
3	0

4	1
5	1
6	1

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

no. of samples = $m = 6$

True label = $y = \text{outcome}$

$$h_0(x^{(i)}) = \frac{1}{1 + e^{-(\alpha_0 + \theta_j x_j^{(i)})}} = \hat{y}^{(i)}$$

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

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

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

initially $\theta_0 = \theta_1 = 0$

$\alpha = 0.1$

$$\hat{y} = h_0(x^{(i)}) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1^{(i)})}}$$

$$\Rightarrow \hat{y}_1 = \frac{1}{1 + e^0} = \frac{1}{2} = 0.5 \quad \begin{array}{l} \text{(predicted probability)} \\ \text{So using threshold } 0.5 \text{ we can classify it as 0.} \end{array}$$

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

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

$$= \frac{1}{6} \left[\frac{-0.5 + 1 + 1.5 + 2 + (-2.5) - 3}{-4} \right]$$

$$= \frac{1}{6} [-4.5] = -0.75$$

$$\hat{O}_0^{\text{new}} = \hat{O}_0^{\text{old}} - \frac{0.1}{6} (0) = 0$$

$$\hat{O}_1^{\text{new}} = \hat{O}_1^{\text{old}} - \frac{0.1}{6} (-0.75) = 0.075$$

Iteration 2.

$$\hat{O}_0 = 0$$

$$\hat{O}_1 = 0.075$$

$$\alpha = 0.1$$

$$\hat{y}_1 = \alpha(x_1) = \frac{1}{1 + e^{-(0.075)x_1}} \quad x_1 = 1 \text{ to } 6$$

$$\hat{y}_2 = \alpha(x_2) = \frac{1}{1 + e^{-(0.075)x_2}} = 0.518$$

$$\hat{y}_3 = \alpha(x_3) = \frac{1}{1 + e^{-(0.075)x_3}} = 0.537$$

$$\hat{y}_4 = \alpha(x_4) = \frac{1}{1 + e^{-(0.075)x_4}} = 0.556$$

$$\hat{y}_5 = \alpha(x_5) = \frac{1}{1 + e^{-(0.075)x_5}} = 0.574$$

$$\hat{y}_6 = \alpha(x_6) = \frac{1}{1 + e^{-(0.075)x_6}} = 0.592$$

$$\hat{y}_7 = \alpha(x_7) = \frac{1}{1 + e^{-(0.075)x_7}} = 0.610$$

$$\begin{aligned} \frac{\partial J}{\partial O_0} &= \frac{1}{6} \left[(0.518 - 0) + (0.537 - 0) + (0.556 - 0) + (0.574 - 0.426) \right. \\ &\quad \left. - (0.592 - 1) + (0.610 - 1) \right] \\ &= \frac{1}{6} [0.387] = 0.0645 \end{aligned}$$

$$\frac{\partial J}{\partial \theta_1} = \frac{1}{6} \left[(0.518 \times 1) + (0.537 \times 2) + (0.556 \times 3) \right. \\ \left. + (-0.426 \times 4) + (-0.408 \times 5) + (-0.39 \times 6) \right] \\ = -0.470$$

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

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

Gradient descent on regularised eqⁿ

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} - \frac{\alpha}{m} \left\{ \sum_{i=1}^m (\hat{y}_i x_i^{(i)} - y^{(i)}) x_j^{(i)} + d \theta_j \right\}$$

$$\theta_j^{\text{new}} = \left(\theta_j^{\text{old}} - \frac{\alpha d}{m} \theta_j^{\text{old}} \right) - \frac{\alpha}{m} \sum_{i=1}^m (\hat{y}_i x_i^{(i)} - y^{(i)}) x_j^{(i)}$$

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} \left(1 - \frac{\alpha d}{m} \right) - \frac{\alpha}{m} \sum_{i=1}^m (\hat{y}_i x_i^{(i)} - y^{(i)}) x_j^{(i)}$$

$$\theta_j^{\text{new}} = \underbrace{\beta}_{\beta < 1} \theta_j^{\text{old}} - \frac{\alpha}{m} \sum_{i=1}^m (\hat{y}_i x_i^{(i)} - y^{(i)}) x_j^{(i)}$$

extra weight that is added to the θ_j^{old} .
 $\because \alpha$ is very small.

→ Normal Eqⁿ modified (Regularised)

$$J = \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i x_i^{(i)} - y^{(i)})^2 + \frac{d}{2m} \theta^T I \theta$$

$$\frac{\partial J}{\partial \theta} = \left[\frac{1}{m} \sum_{i=1}^m x_i^{(i)} (\hat{y}_i x_i^{(i)} - y^{(i)}) + \frac{d}{m} I \theta \right] = 0$$

$$\frac{1}{m} \sum_{i=1}^m (x_i^T \theta - y^{(i)}) + \frac{d}{m} I \theta = 0$$

$$x^T \theta - y^T + d I \theta = 0$$

$$(x^T x + d I) \theta = x^T y$$

$$\theta = \frac{x^T y}{x^T x + d I}$$

$$\Rightarrow \theta = (x^T x + d I)^{-1} x^T y$$

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

for $\lambda = 0$

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

for $\lambda \neq 0$

$\hat{\theta}$ = Regularized θ

⇒ For Logistic Regression

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

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

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

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

logistic regression

Gradient descent on regularised eq:

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

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