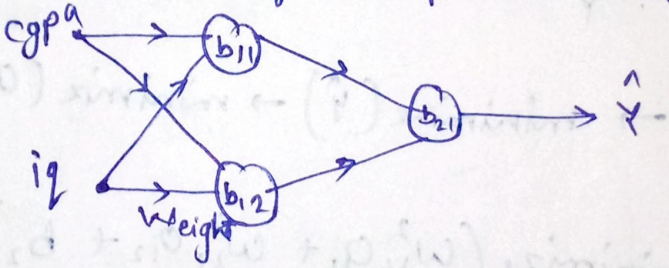
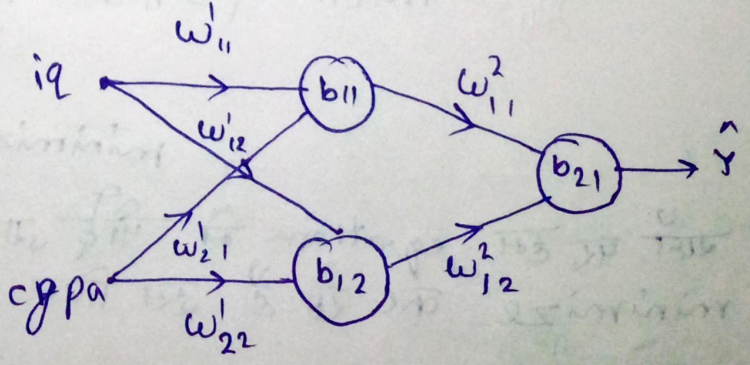
BACKPROPAGATION

* Backpropagation, short for "backpropagation of errors", is an algorithm for supervised learning of ANN using GD.
* Given an ANN and an error function, the method calculates the gradient of the error function with respect to the neural network's weights.
* In simple it is ~~used to~~ to algorithm used to train the Neural network.
* here the term training mean. find the optimal value of the weights and bias. where we got the best result of the prediction ($\hat{y}$)



. prerequisite: Gradient Descent, forward propagation.

iq	cgpa	Lpa
80	8	3
60	9	5
70	5	8
120	7	11



Here the activation function used is linear.

Total trainable parameter = 9

Backpropagation step:

1. initialize the weights (w) and bias (b), e.g. $w=1, b=0$ (we can also take random)
2. You select the point (row)
3. predict (Lpa) → forward propagation [Dot product]
4. for adjust the weight and bias choose a loss function example (MSE) using GD.

(2 to 4 repeat) → Till convergence

$$L = (Y-\hat{Y})^2$$

$$L_2 = (3-18)^2 = 225 \text{ (error)}$$

In the above we see that the error is high so that we have to minimize the error. it is only possible in MSE if we tes or tes the $\hat{Y}$

minimize(L) → minimize($\hat{Y}$) → minimize(O_{21})

$$O_{21} = \text{minimize}(w_{11}^2 O_{11} + w_{21}^2 O_{12} + b_{21})$$

$$O_{11} = \text{minimize}(w_{11}' iq + w_{21}' cgpa + b_{11})$$

$$O_{12} = \text{minimize}(w_{21}' iq + w_{22}' cgpa + b_{12})$$

यहाँ पर हम equation में पीछे जा कर values को minimize कर रहे है इस लिए back propagation. error. कहते है।

η = learning rate

for weight update

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w_{old}}$$

for bias update

$$b_{new} = b_{old} - \eta \frac{\partial L}{\partial b_{old}}$$

for calculating $\frac{\partial L}{\partial w_{old}}$

$$\frac{\partial L}{\partial w_{old}} = \frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial w_{old}} \longrightarrow \text{chain rule}$$

$$\frac{\partial L}{\partial \hat{Y}} = \frac{\partial}{\partial \hat{Y}} (Y - \hat{Y})^2 = -2(Y - \hat{Y})$$

$$\frac{\partial \hat{Y}}{\partial w^2_{old}} = \frac{\partial}{\partial w^2_{old}} \left[O_{11} w^2_{11} + O_{12} w^2_{21} + b_{21} \right]$$

$$= O_{11}$$

Here $\hat{Y} \approx O_{21}$

Hence

$$\frac{\partial L}{\partial w^2_{11}} = -2(Y - \hat{Y}) O_{11}$$

similarly we can calculate the other one. in the above example.

$$\frac{\partial L}{\partial w^2_{11}}, \frac{\partial L}{\partial w^2_{21}}, \frac{\partial L}{\partial b_{21}} \;\Big|\; \frac{\partial L}{\partial w^1_{11}}, \frac{\partial L}{\partial w^1_{21}}, \frac{\partial L}{\partial b_{11}} \;\Big|\; \frac{\partial L}{\partial w^1_{12}}, \frac{\partial L}{\partial w^1_{22}}, \frac{\partial L}{\partial b_{12}}$$

$$\frac{\partial L}{\partial w^2_{11}} = O_{11}, \quad \frac{\partial L}{\partial w^2_{21}} = O_{12}, \quad \frac{\partial \hat{Y}}{\partial b_{21}} = 2(Y - \hat{Y})$$

$$\frac{\partial L}{\partial w_{11}^2} = -2(Y-\hat{Y})\,O_{11}$$

$$\frac{\partial L}{\partial w_{21}^2} = -2(Y-\hat{Y})\,O_{12} \qquad \text{——— (i)}$$

$$\frac{\partial L}{\partial b_{21}} = -2(Y-\hat{Y})$$

$$\frac{\partial L}{\partial w_{11}'} = \frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial w_{11}'}$$

$$\frac{\partial L}{\partial w_{21}'} = \frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial w_{21}'} \qquad \text{——— (ii)}$$

$$\frac{\partial L}{\partial b_{11}} = \frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial b_{11}}$$

$$\frac{\partial L}{\partial w_{12}'} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial w_{12}'}$$

$$\frac{\partial L}{\partial w_{22}'} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial O_{12}} \cdot \frac{\partial O_{12}}{\partial w_{12}'} \qquad \text{——— (iii)}$$

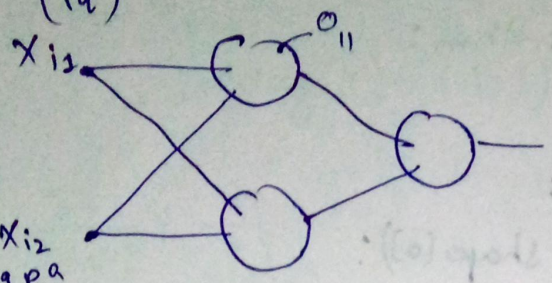
$$\frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial O_{12}} \cdot \frac{\partial O_{12}}{\partial b_{12}}$$

here $\frac{\partial \hat{Y}}{\partial O_{11}} = \frac{\partial}{\partial O_{11}}\left[w_{11}^2 O_{11} + w_{21}^2 O_{21} + b_{21}\right] = w_{11}^2$

$$\frac{\partial \hat{Y}}{\partial O_{12}} = \frac{\partial}{\partial O_{12}}\left[w_{11}^2 O_{11} + w_{21}^2 O_{21} + b_{21}\right] = w_{21}^2$$

i = row

(iq)



$$\begin{cases} \dfrac{\partial O_{11}}{\partial w'_{11}} = \dfrac{\partial [iq\, w'_{11} + cgpa\, w'_{21} + b_{11}]}{\partial w'_{11}} = iq = x_{i1} \\ \dfrac{\partial O_{11}}{\partial w'_{21}} = x_{i2} \qquad \dfrac{\partial O_{11}}{\partial b_{11}} = 1 \end{cases}$$

$$\begin{cases} \dfrac{\partial O_{12}}{\partial w'_{12}} = \dfrac{\partial}{\partial w'_{12}} [iq\, w'_{12} + cgpa\, w'_{22} + b_{12}] = iq = x_{i1} \\ \dfrac{\partial O_{12}}{\partial w'_{22}} = x_{i2} \qquad \dfrac{\partial O_{12}}{\partial b_{12}} = 1 \end{cases}$$

Now put all these values in equations (i), (ii) and (iii)
Hence our nine derivatives are

$$\frac{\partial L}{\partial w'_{11}} = -2(Y-\hat{Y})\, w^2_{11}\, x_{i1}$$

$$\frac{\partial L}{\partial w'_{21}} = -2(Y-\hat{Y})\, w^2_{11}\, x_{i2}$$

$$\frac{\partial L}{\partial b_{11}} = -2(Y-\hat{Y})\, w^2_{11}$$

$$\frac{\partial L}{\partial w'_{12}} = -2(Y-\hat{Y})\, w^2_{21}\, x_{i1}$$

$$\frac{\partial L}{\partial w'_{22}} = -2(Y-\hat{Y})\, w^2_{21}\, x_{i2}$$

$$\frac{\partial L}{\partial b_{12}} = -2(Y-\hat{Y})\, w^2_{21}$$

$$\frac{\partial L}{\partial w^2_{11}} = -2(Y-\hat{Y})\, O_{11}$$

$$\frac{\partial L}{\partial w^2_{21}} = -2(Y-\hat{Y})\, O_{12}$$

$$\frac{\partial L}{\partial b_{21}} = -2(Y-\hat{Y})$$

Back propagation Algorithm:

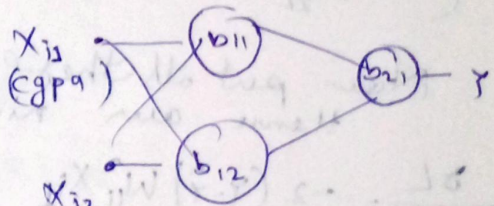
epoch=(5)

```
for i in range(epochs):
    for j in range(X.shape[0]):
        → select 1 row (random)
        → predict (using forward loop)
        → Calculate loss (using loss function → MSE)
        → update weight and bias using GD
```

$$W_n = W_o - \eta \frac{\partial L}{\partial W}$$

calculate avg loss for the epoch.

cgpa	profile score	placement
8	8	1
7	9	1
6	10	0
5	5	0



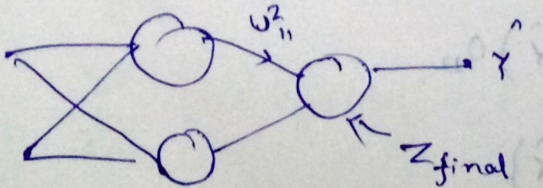
Activation function = sigmoid

trainable parameters = 9

$$Loss = -Y \log(\hat{Y}) - (1-Y) \log(1-\hat{Y})$$

$$z = 0.1 \times 8 + 0.1 \times 8 + 0$$

$$\sigma(z) = 0.11$$



$L \to \hat{y} \to z_{final} \to z = w_{11}^2 a_{11} + w_{21}^2 a_{12} + b_{21}$

$$L = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\left.\begin{aligned} \frac{\partial L}{\partial w_{11}^2} &= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w_{11}^2} = -(y-\hat{y})\, a_{11} \\ \frac{\partial L}{\partial w_{21}^2} &= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w_{21}^2} = -(y-\hat{y})\, a_{12} \\ \frac{\partial L}{\partial b_{21}} &= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial b_{21}} = -(y-\hat{y}) \end{aligned}\right\}$$

Now $\frac{\partial L}{\partial \hat{y}} = \frac{\partial [-y \log(\hat{y}) - (1-y)\log(1-\hat{y})]}{\partial \hat{y}}$

$$= \frac{-y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})} = \frac{-y(1-\hat{y}) + \hat{y}(1-y)}{\hat{y}(1-\hat{y})}$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \quad \text{—(i)}$$

$$\frac{\partial \hat{y}}{\partial z} = \frac{\partial}{\partial z}(\sigma(z)) = \sigma(z)[1-\sigma(z)] = \hat{y}(1-\hat{y}) \quad \text{—(ii)}$$

from (i) and (ii)

$$\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} = -(y-\hat{y})$$

$\frac{\partial z}{\partial (w_{11}^2)}$

similarly:

* $L \to \hat{Y} \to z_f \to o_{11} \to z_{prev} \to w'_{11}$

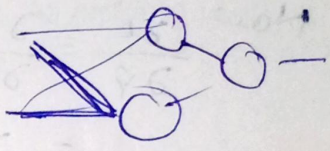
$$\frac{\partial L}{\partial w'_{11}} = \frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial z_f} \times \frac{\partial z_f}{\partial o_{11}} \times \frac{\partial o_{11}}{\partial z_{pre}} \times \frac{\partial z_{pre}}{\partial w'_{11}}$$

$\frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial z_f} = -(Y - \hat{Y})$; $\frac{\partial z_f}{\partial o_{11}} = w^2_{11}$; $\frac{\partial o_{11}}{\partial z_{pre}} = o_{11}(1 - o_{11})$; $\frac{\partial z_{pre}}{\partial w'_{11}} = x_{i1}$

$$\frac{\partial L}{\partial w'_{11}} = -(Y - \hat{Y}) \cdot w^2_{11} \cdot o_{11}(1 - o_{11}) x_{i1}$$

$$\frac{\partial L}{\partial w'_{21}} = -(Y - \hat{Y}) w^2_{11} o_{11}(1 - o_{11}) x_{i2}$$

$$\frac{\partial L}{\partial b_{11}} = -(Y - \hat{Y}) w^2_{11} o_{11}(1 - o_{11})$$



* $L \to \hat{Y} \to z_f \to o_{12} \to w'_{12} \to z_p$

$$\frac{\partial L}{\partial w'_{12}} = \frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial z_f} \times \frac{\partial z_f}{\partial o_{12}} \times \frac{\partial o_{12}}{\partial z_p} \times \frac{\partial z_p}{\partial w'_{12}}$$

$\frac{\partial L}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial z_f} = -(Y - \hat{Y})$; $\frac{\partial z_f}{\partial o_{12}} = w^2_{21}$ (since $z_f = w^2_{11} o_{11} + w^2_{21} o_{12} + b_{21}$) ; $\frac{\partial o_{12}}{\partial z_p} = o_{12}(1 - o_{12})$; $\frac{\partial z_p}{\partial w'_{12}} = x_{i1}$

$$\frac{\partial L}{\partial w'_{12}} = -(Y - \hat{Y}) w^2_{21} o_{12}(1 - o_{12}) x_{i1}$$

$$\frac{\partial L}{\partial w'_{22}} = -(Y - \hat{Y}) w^2_{21} o_{12}(1 - o_{12}) x_{i2}$$

$$\frac{\partial L}{\partial b_{12}} = -(Y - \hat{Y}) w^2_{21} o_{12}(1 - o_{12})$$

→ loss function is a function of all trainable parameters

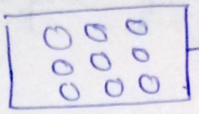
→ for example Backpropagation Regression.

Loss function = $(Y-\hat{Y})^2$ = MSE

OR $L(\hat{Y})$ because Y = const.

$$\hat{y} = w_{11}^2 o_{11} + w_{21}^2 o_{12} + b_{21}$$

$$\hat{Y} = w_{11}^2\left[w_{11}' x_{i1} + w_{21}' x_{i2} + b_{11}\right] + w_{21}^2\left[w_{12}' x_{i1} + w_{22}' x_{i2} + b_{12}\right] + b_{12}$$

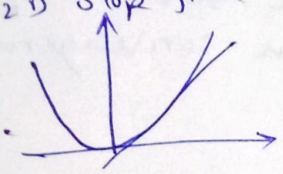


tune the value for find minimum loss

derivative function depends only on one variable

$y = f(x) = x^2 + x$

in 2D slope find

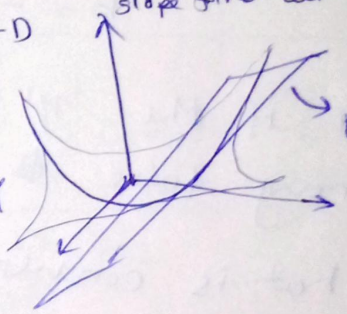


Gradient function depends on multiple variable

$z = f(x,y) = x^2 + y^2$

in 3-D

slope find with each dim



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

$\frac{\partial z}{\partial y} = 2y$

* concept of minima $\frac{dy}{dx} = 0$

at that point of x minima is lies

example $z = x^2 + y^2$

$(x, y) \rightarrow \min(z)$

$\frac{\partial z}{\partial x} = 0 \Rightarrow 2x = 0 \Rightarrow x = 0$

$\frac{\partial z}{\partial y} = 0 \Rightarrow 2y = 0 \Rightarrow y = 0$

If $(x, y) = (0, 0)$ then z is minimum.

→ effect of Learning Rate(η):

$$w_n = w_0 - \eta \frac{\partial L}{\partial w}$$

learning rate steps को smooth बनाता है।

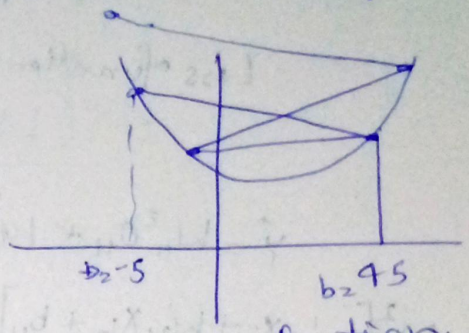
let $b = -5$

let $\frac{\partial L}{\partial b} = -50$

update 'b'

$b = -5 - (-50)$

$= 45$



- for control this type of zigzag motion.
use learning rate (η)

let $b = -5$ $\frac{\partial L}{\partial b} = -50$ $\eta = 0.01$

update 'b'

$b = -5 + (0.01 \times 50) = -4.5$

Here we take small steps.

- but if the 'η' is very small then convergence may take long time

→ what is convergence?

(i) if $w_{new} = w_{old}$

(ii) if run for loop in 100 or 1000 times

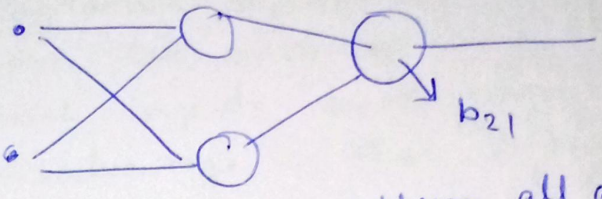
Backpropgation Intution:

- Now we understand

$$W_{new} = W_{old} - \eta \frac{\partial L}{\partial W}$$

- Here learning rate we assume $\eta = 1$

$$W_{new} = W_{old} - \frac{\partial L}{\partial W}$$

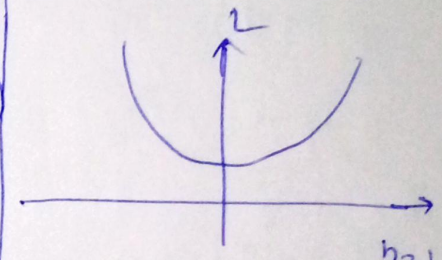


Now here less us assum all are constant except b_{21}

$$L = (Y - \hat{Y})^2 \Rightarrow L(b_{21})$$

Now update b_{21}

$$b_{21} = b_{21} - \frac{\partial L}{\partial b_{21}}$$



find the value of b_{21} where L is min

$\frac{\partial L}{\partial b_{21}}$: L derivative w.r.t b_{21}

if $\frac{\partial L}{\partial b_{21}} = +ve$ $\quad b_{21} \downarrow \quad L \downarrow$

if $\frac{\partial L}{\partial b_{21}} = -ve$ $\quad b_{21} \uparrow \quad L \downarrow$