Backpropagation in ANN (MLP)

---

* Chain Rule of Derivatives :-

The Chain Rule of Derivatives is used to find derivative of composite function (i.e function of function).

Suppose,

$y = f(u)$; y is a function of ~~y~~ u

& $u = g(x)$; u is function of ~~x~~ x then,

$y = f(g(x))$; y is function of ~~y~~ x

i.e
If y is function of u and u is function of x then chain rule of derivatives is,

$$\boxed{\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}}$$

* Backpropagation :-

Backpropagation is the algorithm used to train ANN models.

In Backpropagation we calculate the gradients of loss function w.r.t. weights and biases of the model. (i.e $\frac{\partial L}{\partial \omega}, \frac{\partial L}{\partial b}$) and update the values of Weights and Biases (W & b) until we get the values of weights and biases for which the loss function is minimum.

We use the weight updation formula as,

$$\boxed{\begin{aligned} W_{new} &= W_{old} - \eta * \left(\frac{\partial L}{\partial W_{old}}\right) \\ b_{new} &= b_{old} - \eta * \left(\frac{\partial L}{\partial b_{old}}\right) \end{aligned}}$$

And the method used to calculate gradients is the 'Gradient Descent'.

* Intution :-

Let's suppose we have dataset with two input columns iq & cgpa and our aim is to predict the salary of student (placed).

So this is regression problem.

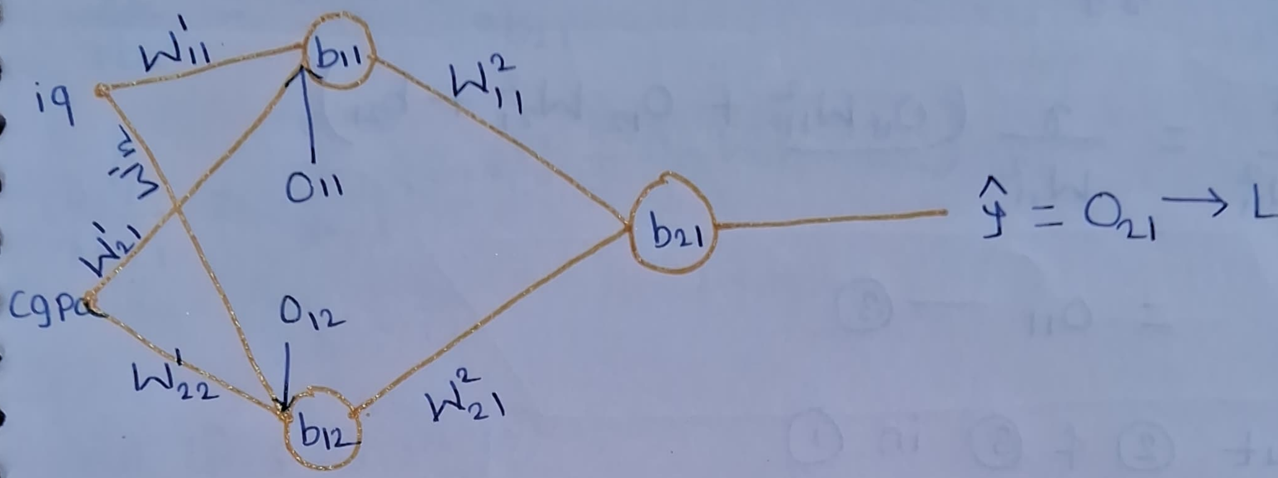
consider,

A.F (each layer) = linear

L.F (loss) = MSE

So the architecture is,

Two inputs, one hidden layer with two nodes and one output layer



Here, No. of trainable parameters are,

Layer 1 $= (2 \times 2) + 2 = 4 + 2 = 6$

Layer 2 $= (2 \times 1) + 1 = 2 + 1 = 3$

Total trainable parameters $= 6 + 3 = \underline{\underline{9}}$

In this problem we have to calculate 9 derivatives.

① $\frac{\partial L}{\partial W_{11}^2}$ ② $\frac{\partial L}{\partial W_{21}^2}$ ③ $\frac{\partial L}{\partial b_{21}}$ ④ $\frac{\partial L}{\partial W_{11}^1}$ ⑤ $\frac{\partial L}{\partial W_{12}^1}$

⑥ $\frac{\partial L}{\partial W_{21}^1}$ ⑦ $\frac{\partial L}{\partial W_{22}^1}$ ⑧ $\frac{\partial L}{\partial b_{11}}$ ⑨ $\frac{\partial L}{\partial b_{12}}$

Let's calculate one by one

① $\frac{\partial L}{\partial W_{11}^2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial W_{11}^2}$ —— ①

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}}(y-\hat{y})^2 = -2(y-\hat{y}) \quad \text{— (2)}$$

$$\frac{\partial \hat{y}}{\partial W_{11}^2} = \frac{\partial}{W_{11}^2}\left(\underline{O_{11} W_{11}^2} + O_{12} W_{21}^2 + b_{21}\right)$$

$$= O_{11} \quad \text{— (3)}$$

Put (2) & (3) in (1)

so the first derivative is,

$$\boxed{\frac{\partial L}{\partial W_{11}^2} = -2(y-\hat{y})\, O_{11}}$$

(2) $$\frac{\partial L}{\partial W_{21}^2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial W_{21}^2} \quad \text{— (4)}$$

$$\frac{\partial \hat{y}}{\partial W_{21}^2} = \frac{\partial}{\partial W_{21}^2}\left(O_{11} W_{11}^2 + \underline{O_{12} W_{21}^2} + b_{21}\right)$$

$$= O_{12} \quad \text{— (5)}$$

use (2) & (5) in (4)

$$\boxed{\frac{\partial L}{\partial W_{21}^2} = -2(y-\hat{y})\, O_{12}}$$

$$\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b_{21}} \quad \text{--- (6)}$$

$$\frac{\partial \hat{y}}{\partial b_{21}} = \frac{\partial}{\partial b_{21}}\left(O_{11} W_{11}^{2} + O_{12} W_{21}^{2} + \underline{b_{21}}\right)$$

$$= 1 \quad \text{--- (7)}$$

use (7) & (2) in (6)

$$\boxed{\frac{\partial \hat{y}}{\partial b_{21}} = -2(y-\hat{y})}$$

$$\frac{\partial L}{\partial W_{11}^{1}} \quad \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial W_{11}^{1}} \quad \text{--- (8)}$$

$$\frac{\partial \hat{y}}{\partial O_{11}} = \frac{\partial}{\partial O_{11}}\left(\underline{W_{11}^{2}} O_{11} + W_{21}^{2} O_{21} + b_{21}\right) = W_{11}^{2} \quad \text{--- (9)}$$

$$\frac{\partial O_{11}}{\partial W_{11}^{1}} = \frac{\partial}{\partial W_{11}^{1}}\left(iq\, W_{11}^{1} + cgpa\, W_{21}^{1} + b_{11}\right)$$

$$= iq\ (x_{i1}) \quad \text{--- (10)}$$

use (9) (10) & (2) in (8)

$$\boxed{\frac{\partial L}{\partial W_{11}^{1}} = -2(y-\hat{y})\, W_{11}^{2}\, x_{i1}}$$

similarly,

$$\frac{\partial L}{\partial W^1_{21}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial W^1_{21}}$$

$$\boxed{\frac{\partial L}{\partial W^1_{21}} = -2(y-\hat{y})W^2_{11}X_{i2}}$$

$$\frac{\partial L}{\partial b_{11}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial b_1}$$

$$\boxed{\frac{\partial L}{\cancel{\partial W^1_{21}}\ \partial b_{11}} = -2(y-\hat{y})W^2_{11}}$$

$$\frac{\partial L}{\partial W^1_{12}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial W^1_{12}}$$

$$\boxed{\frac{\partial L}{\partial W^1_{12}} = -2(y-\hat{y})W^2_{21}X_{i1}}$$

$$\frac{\partial L}{\partial W^1_{22}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial W^1_{22}}$$

$$\boxed{\frac{\partial L}{\partial W^1_{22}} = -2(y-\hat{y})W^2_{21}X_{i2}}$$

$$\frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial b_{12}}$$

$$\boxed{\frac{\partial L}{\partial b_{12}} = -2(y-\hat{y}) W_{21}^{2}}$$

After calculating all these weights, We update the weights using weight updation formula.

* Steps in Backpropagation :-

1. Initialise values of weights & Bias.

2. for j in (epochs): (eg: 100, 1000)

 For i in rang (n):

 2a) 1-stud. → Forward prop. → Predict

 2b) Loss calculation

 2c) Adjust all w. & b t

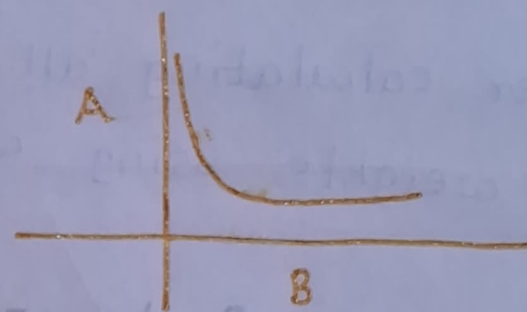
 n times

 $w_n = W_o - \eta\left(\frac{\partial L}{\partial W_o}\right)$ → one for each trainable para.

* key concepts :-

1 Derivative :-

Derivative is the method used to observe how 'A' changes when we make changes in 'B'.



suppose we write $\frac{dy}{dx}$

that means, if we twick x then how y changes according to x.

and we called it as derivative of y. w.r.t x.

and we can also say that,

y is a function of x. $\left(\frac{dy}{dx}\right)$

2 Gradient :-

As like derivative gradient is just an extension of derivative.

i.e derivative term is refered when ~~the funct~~ something is function of single variable.

i.e y is function of x

and derivative of y w.r.t x means,

How y changes according to x.

but,

if y is function of $x_1, x_2 \ldots x_n$

then, y is function of multiple variables, in this case if we want to know how y changes w.r.t. all these variables, we calculate partial derivatives and these partial derivative are called as gradients.

i.e $\dfrac{\partial y_i}{\partial x}$

So,

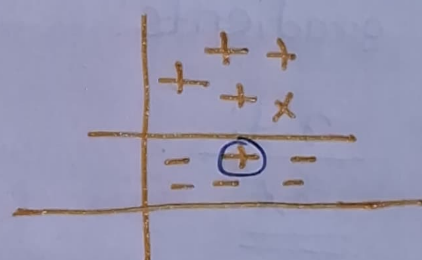
In Deep Learning our Loss Function is function of all trainable parameters in the network ie (all weights W & Biases b) therefore here we calculate partial derivatives i.e gradients using a method called Gradient Descent.

* Weight Updation Formula intution:-

We know the weight updation formula is,

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

Why (-) :-
We will take classification example,
suppose the actual value is + but model predicts -
i.e the positive point stuck in negative region
To bring back this point we have to add something in weights.
our prediction is -tive & the value of gradiet also become - (because the value of gradient is less than 0)
Therefore, it become positive and that exactly we want.



and it is same for when model predict - but actual value is +.
And for correct prediction gradient value is zero.

* What is convergence :-
We have converged when,

$$W_{new} \approx W_{old}$$

This happens because as we near the minimum, the slope (gradient) approches zero, making the update negligible.
We usually set a fixed number of epochs (eg.100) rather than waiting for perfect mathematical convergence.