Notes

Thursday

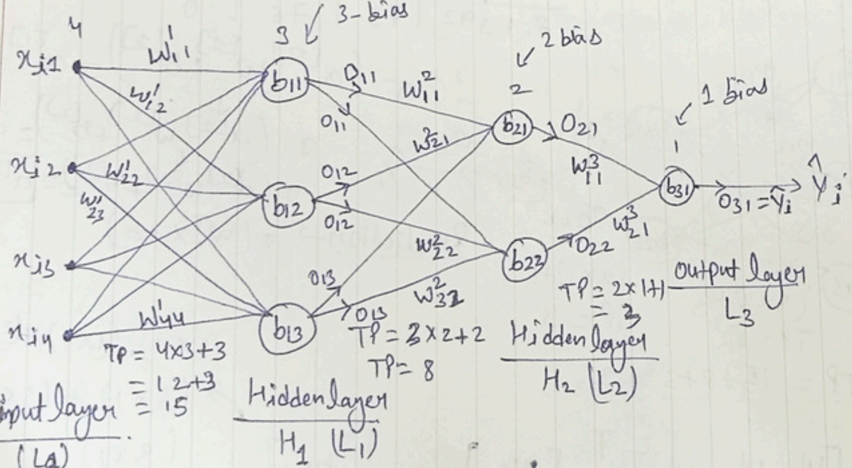
19 December

we have Data $\{m \times n\}$

$m \rightarrow$ row

$n \rightarrow$ colum. $\rightarrow 4$

	1	2	3	4	$\hat{y}$
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	



Notations

bias $\rightarrow b_{ij}$

i = layer No.

j = Node No.

output $\rightarrow O_{ij}$

weights $\rightarrow w^k_{ij}$

$k \rightarrow$ Entring in which layer

$i \rightarrow$ Coming from which Node

$j \rightarrow$ Entering in which Node

Total Trainable Parameters (TP) $= 15 + 8 + 3$

$= 26$

This means that when we train this NN, then our Backpropagation Algorithm try to find the values of 26 weights and bias.

Life is full of misery, loneliness, and suffering - and it's all over much too soon - Woody Allen

December 2024

wk	Mo	Tu	We	Th	Fr	Sa	Su
48	30	31					1
49	2	3	4	5	6	7	8
50	9	10	11	12	13	14	15
51	16	17	18	19	20	21	22
52	23	24	25	26	27	28	29

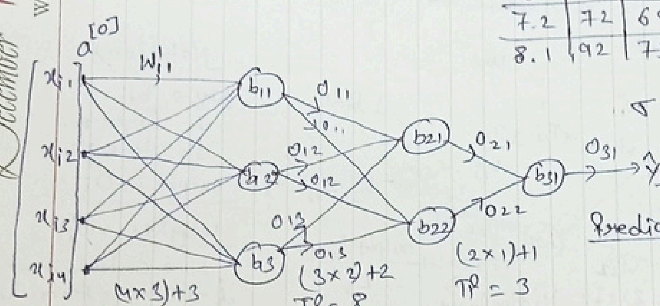
Week 51 Day (354-012)

2024

* Forward Propagation:-

Data

cgpa	iq	10th m	12th m	Placed (y)
7.2	72	69	81	1
8.1	92	75	76	0



$(4\times3)+3$ TP = 15

$(3\times2)+2$ TP = 8

$(2\times1)+1$ TP = 3

$\sigma(a^{[0]} w^{[1]} + b^{[1]}) = a^{[1]}$

$\sigma(a^{[1]} w^{[2]} + b^{[2]}) = a^{[2]}$

$\sigma(a^{[2]} w^{[3]} + b^{[3]}) = \boxed{a^{[3]}}$

Prediction → $\sigma(w^T x + b)$

TTP = 15+8+3 = 26

Layer #1

$$\begin{bmatrix} w^1_{11} & w^1_{12} & w^1_{13}\\ w^1_{21} & w^1_{22} & w^1_{23} \\ w^1_{31} & w^1_{32} & w^1_{33}\\ w^1_{41} & w^1_{42} & w^1_{43}\end{bmatrix}^T \begin{bmatrix} x_{i1}\\ x_{i2}\\ x_{i3}\\ x_{i4}\end{bmatrix} + \begin{bmatrix} b_{11}\\ b_{12}\\ b_{13}\end{bmatrix}$$

$(4\times3)^T \to (3\times4)$ (4×1) should same (3×1)

$(3\times4).(4\times1) + (3\times1)$
$\to (3\times1) + (3\times1)$
(3×1)

$$= \begin{bmatrix} w^1_{11} x_{i1} + w^1_{21} x_{i2} + w^1_{31} x_{i3} + w^1_{41} x_{i4} \\ w^1_{12} x_{i1} + w^1_{22} x_{i2} + w^1_{32} x_{i3} + w^1_{42} x_{i4} \\ w^1_{13} x_{i1} + w^1_{23} x_{i2} + w^1_{33} x_{i3} + w^1_{43} x_{i4} \end{bmatrix}_{(3\times1)} + \begin{bmatrix} b_{11}\\ b_{12}\\ b_{13}\end{bmatrix}_{(3\times1)}$$

$$= \sigma\left(\begin{bmatrix} +b_{11}\\ +b_{12}\\ +b_{13}\end{bmatrix}\right)$$

$$= \begin{bmatrix} O_{11}\\ O_{12}\\ O_{13}\end{bmatrix} \to a^{[1]}$$

Layer #2

$$\begin{bmatrix} w^2_{11} & w^2_{12}\\ w^2_{21} & w^2_{22}\\ w^2_{31} & w^2_{32}\end{bmatrix}^T_{3\times2} \begin{bmatrix} O_{11}\\ O_{12}\\ O_{13}\end{bmatrix}_{3\times1} + \begin{bmatrix} b_{21}\\ b_{22}\end{bmatrix}_{2\times1}$$

$(3\times2)^T.(3\times1) + (2\times1)$
$(2\times3)\ (3\times1) + (2\times1)$
$(2\times1) + (2\times1)$
(2×1)

$$= \sigma\left(\begin{bmatrix} w^2_{11} O_{11} + w^2_{21} O_{12} + w^2_{31} O_{13} + b_{21} \\ w^2_{12} O_{11} + w^2_{22} O_{12} + w^2_{32} O_{13} + b_{22}\end{bmatrix}_{2\times1}\right) \to a^{[2]}$$

$$= \begin{bmatrix} O_{21}\\ O_{22}\end{bmatrix}$$

Layer #3

$$\begin{bmatrix} w^3_{11} \\ w^3_{21} \end{bmatrix}^T_{2\times1} \begin{bmatrix} O_{21} \\ O_{22} \end{bmatrix}_{2\times1} + \begin{bmatrix} b_{31} \end{bmatrix}_{1\times1} \rightarrow \begin{array}{l} (2\times1)^T(2\times1)+(1\times1) \\ (1\times2)(2\times1)+(1\times1) \\ (1\times1)+(1\times1) \\ (1\times1) \end{array}$$

$$= \cancel{\begin{bmatrix} w^3_{11} O_{21} \\ w^3_{21} O_{22} \end{bmatrix}} + \cancel{(b_{31})}$$

$$= \sigma\left(\begin{bmatrix} w^3_{11} O_{21} + w^3_{21} O_{22} + b_{31} \end{bmatrix}\right)$$

$$= \hat{Y}_i = O_{31} \rightarrow a^{[3]}$$

$$\sigma\Bigg(\underbrace{\sigma\Big(\underbrace{\big(\sigma\big(\underbrace{a^{[0]} w^{[1]} + b^{[1]}}_{a^{[1]}}\big)\big) w^{[2]} + b^{[2]}}_{a^{[2]}}\Big)\Big) w^{[3]} + b^{[3]}}\Bigg)$$

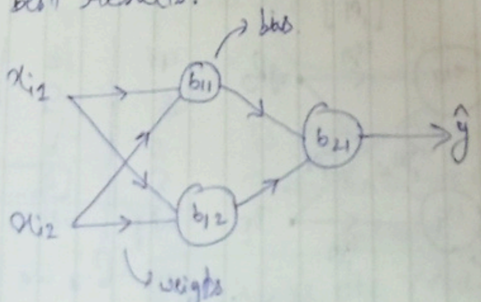
$$a^{[3]} = \hat{Y}_i$$

Backpropagation.

Backpropagation, short for backward propagation of errors, is an algorithm for supervised learning of artificial ~~ANN~~ Neural Networks using gradient descent.

Given an ANN and error function, the method calculates the gradient of error function with respect to the NN weights.

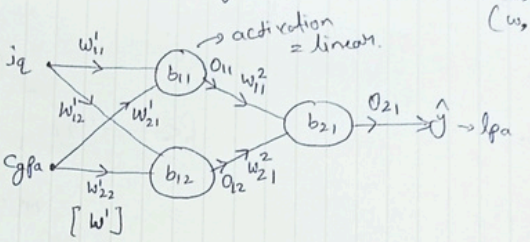
In Simple :- It is an algorithm to train NN.
means this algorithm try to find the best optimum values of weights and bais for given data, on which our model gives best results.



* Backpropagation follows the series of steps to find the values (w, b)

lets take an e.g. data

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



Steps → *) Initialization of weights and bias
(w, b) → we can initialize it with rand. values
but we take $w \to 1$, $b \to 0$

* We select a point (row) → e.g. $\underline{80, 8}$ (iq, cgpa)

* Predict (lpa) → forward propagation.
By using the w, b values we predict the first 'lpa' → $\hat{y}$
$(W^T x + b)$

lets assume our pred. is 17 lpa, but its is not correct means our w, b values are not correct.
So for this we have to find the error values and reduce it.

* Choose a loss function :-
Since we are using linear f^n here
So we take $\boxed{MSE}$ as a loss f^n

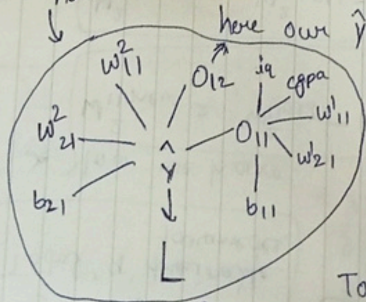
$$L = (y - \hat{y})^2 \quad y \to \text{actual value}, \ \hat{y} \to \text{pred. values}$$

$$L = 3 - 17$$

↳ error, means we have to reduce the error
So here we can only change the '$\hat{y}$'

hierarchy

here our $\hat{y}$ is depend of O_{21} which further depend on other values



$$\hat{y} = O_{21} = w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}$$

further $O_{11} = w'_{11} iq + w'_{21} cgpa + b_{11}$

$O_{12} = w'_{12} iq + w'_{22} cgpa + b_{12}$

To reduce the loss is depend upon that how we change the values of (w, b) because our input is fixed.

To change the value by moving backword that's why we call it backword prop.

* Updation of weights and bias.
  ↳ By using gradient Descent
  ↳ $\boxed{W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}}}$
  ↳ $\boxed{b_{new} = b_{old} - \eta \frac{\partial L}{\partial b_{old}}}$

Here we have
6 → weights
3 → bias
Total '9' trainable parameters.

* Now we have.

$\boxed{w^2_{11\,new} = w^2_{11\,old} - \eta \frac{\partial L}{\partial w^2_{11\,old}}}$ → here we have η, w^2_{11} old and we have to find $\boxed{\frac{\partial L}{\partial w^2_{11\,old}}}$

It is derivative of loss w.r.t. weight.

$\boxed{w^2_{21\,new} = w^2_{21\,old} - \eta \frac{\partial L}{\partial w^2_{21\,old}}}$

$\boxed{b_{21\,new} = b_{21\,old} - \eta \frac{\partial L}{b_{21\,old}}}$

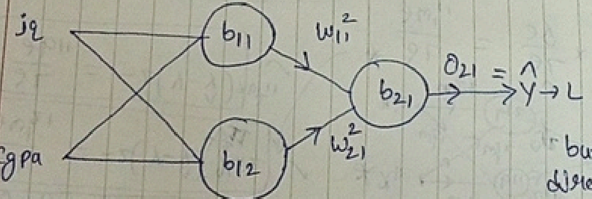
* So we have to Calculate 9 → derivatives.

January 2025

Wk	Mo	Tu	We	Th	Fr	Sa	Su
01			1	2	3	4	5
02	6	7	8	9	10	11	12
03	13	14	15	16	17	18	19
04	20	21	22	23	24	25	26
05	27	28	29	30	31		

By failing to prepare, you are preparing to fail. - Benjamin Franklin

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



$\frac{\partial L}{\partial w^2_{11}}$ ↘ but w^2_{11} is not directly related with (L)

So we have change w^2_{11} then $\hat{y}$ the L

$\hat{y} \to w^2_{11}$
L (min)
$\frac{\partial y}{\partial x}$ → change in y w.r.t. 'x'

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

$$\frac{\partial L}{\partial w^2_{11}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w^2_{11}}$$ → ~~change~~ Chain rule of diff.

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

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

$$\frac{\partial \hat{y}}{\partial w^2_{11}} = O_{11}$$

December 2024

Wk	Mo	Tu	We	Th	Fr	Sa	Su
48	30	31					1
49	2	3	4	5	6	7	8
50	9	10	11	12	13	14	15
51	16	17	18	19	20	21	22
52	23	24	25	26	27	28	29

... the gift of life; it is up to us to give ourselves the gift of living well - Voltaire

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

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

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

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

$$\boxed{\frac{\partial L}{\partial w^1_{21}} = -2(y-\hat{y})\,w^2_{11}\,x_{i2}} \checkmark$$

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

$$\boxed{\frac{\partial L}{\partial w^1_{12}} = -2(y-\hat{y})\,w^2_{21}\,x_{i1}} \checkmark$$

* $$\frac{\partial L}{\partial w^2_{21}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w^2_{21}}$$

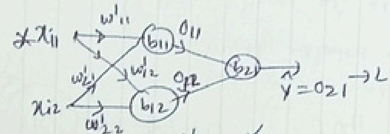
$\hat{y} \rightarrow w^2_{21}$

$L\,(\min)$

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

* $$\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b_{21}} = \frac{\partial}{\partial b_{21}}\left(w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}\right) = 1$$

$L \rightarrow \hat{y} \rightarrow b$



$L \rightarrow \hat{y} \rightarrow O_{11} \rightarrow w^1_{11}$

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

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

$$\boxed{\frac{\partial L}{\partial w^1_{22}} = -2(y-\hat{y})\,w^2_{21}\,x_{i2}} \checkmark$$

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

* $$\frac{\partial O_{11}}{\partial w^1_{11}} = \frac{\partial}{\partial w^1_{11}}\left[iq\,w^1_{11} + cgpa\,w^1_{21} + b_{11}\right] = iq \rightarrow x_{i1}$$

$$\frac{\partial O_{11}}{\partial w^1_{21}} = cgpa \rightarrow x_{i2}$$

$$\frac{\partial O_{11}}{\partial b_{11}} = 1$$

* $$\frac{\partial O_{12}}{\partial w^1_{12}} = \frac{\partial}{\partial w^1_{12}}\left[iq\,w^1_{12} + cgpa\,w^1_{22} + b_{12}\right] = iq \rightarrow x_{i1}$$

$$\frac{\partial O_{12}}{\partial w^1_{22}} = cgpa \rightarrow x_{i2}$$

$$\frac{\partial O_{12}}{\partial b_{12}} = 1$$

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

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

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

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

* we have to find $\frac{\partial L}{\partial \hat{y}} = -2(y-\hat{y})$, $\frac{\partial \hat{y}}{\partial O_{11}}$, $\frac{\partial \hat{y}}{\partial O_{12}}$ and all the other derivatives

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

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

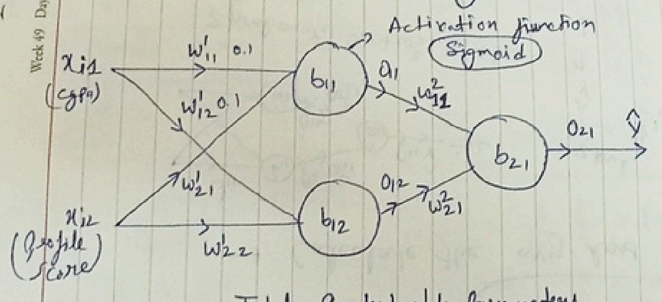
Backpropagation steps (once again)

epochs ↑ loop → 1000 ↓ till convergence.

* weight / bias → value initialize → random $w = 1$, $b = 0$
* for i in range(data):
  * 1 (row) → forward prop. – predict
  * loss Calculate (mse)
  * Adjust all weights and bias

$$\boxed{W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}}} \rightarrow 9 \text{ times}$$

Backpropagation for Classification



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

$Z = -$

$\sigma(z) = O_{11}$

Total 9 trainable parameters

Classification data

cgpa	p. score	lpa
8	8	1
7	9	1
6	10	0
5	5	0

Everything is same except the data and the loss is change (BCE) and also we will use Sigmoid activation f^n at the end of every node output

Activation f^n = Sigmoid

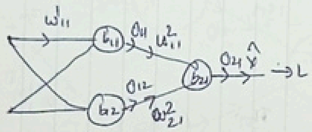
Binary cross entropy → $\boxed{Loss = -Y \log(\hat{Y}) - (1-Y)\log(1-\hat{Y})}$

Backpropagation algorithm

epoch = 5
for i in range(epoch):
  for j in range(X.shape[0]):
    → Select 1 row (random) or 1 by 1
    → Predict (using forward prop.)
    → Calculate loss (using loss f^n → BCE)
    → update weights and bias using GD

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

  → Calculate the avg. loss for the epoch.



$z_{final} \rightarrow w^2_{11}, w^2_{21}$
↑
$\hat{y}$
↑
L

$$z_{final} = w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}$$

*
$$\frac{\partial \hat{y}}{\partial z} = \frac{\partial}{\partial z}(\sigma(z))$$

$$\boxed{\frac{\partial(\sigma(z))}{\partial z} = \sigma(z)[1-\sigma(z)]}$$

$$\boxed{\frac{\partial \hat{y}}{\partial z} = \hat{y}(1-\hat{y})}$$

$$\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} = \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \times \cancel{\hat{y}(1-\hat{y})}$$

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

$$\frac{\partial z}{\partial w^2_{11}} = \frac{\partial}{\partial w^2_{11}}(w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21})$$

$$= O_{11}$$

$$\frac{\partial z}{\partial w^2_{11}} = O_{12}$$

$$\frac{\partial z}{\partial b_{21}} = 1$$

*
$$\frac{\partial L}{\partial w^2_{11}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w^2_{11}} = -(y-\hat{y}) O_{11}$$

$$\frac{\partial L}{\partial w^2_{21}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w^2_{21}} = -(y-\hat{y}) O_{12}$$

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

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

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

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

$$\boxed{\frac{\partial L}{\partial \hat{y}} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})}}$$

Now next weights

$L \to \hat{y} \to z_f \to O_{11} \to z_p \to w'_{11}$

$$\frac{\partial L}{\partial w'_{11}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_f} \times \frac{\partial z_f}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial z_p} \times \frac{\partial z_p}{\partial w'_{11}}$$

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

$$\frac{\partial z_f}{\partial O_{11}} = \frac{\partial}{\partial O_{11}} (w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21})$$

$$= w^2_{11}$$

$$\frac{\partial O_{11}}{\partial z_p} = \frac{\partial (\sigma(z_p))}{\partial z_p} \qquad \boxed{\sigma(z_p) = O_{11}}$$

$$= \sigma(z_p)(1-\sigma(z_p))$$

$$= O_{11}[1-O_{11}]$$

$$\frac{\partial z_p}{\partial w'_{11}} = \frac{\partial}{\partial w'_{11}} [w'_{11} x_{i1} + w'_{21} x_{i2} + b_{11}]$$

$$= x_{i1}$$

$$\boxed{\begin{aligned} \frac{\partial L}{\partial w^2_{11}} &= -(y-\hat{y}) O_{11} \\ \frac{\partial L}{\partial w^2_{21}} &= -(y-\hat{y}) O_{12} \\ \frac{\partial L}{\partial b_{21}} &= -(y-\hat{y}) \end{aligned}}$$

3

3

$$\boxed{\begin{aligned} \frac{\partial L}{\partial w'_{11}} &= -(y-\hat{y})\, w^2_{11}\, O_{11}(1-O_{11})\, x_{i1} \\ \frac{\partial L}{\partial w'_{21}} &= -(y-\hat{y})\, w^2_{11}\, O_{11}\, (1-O_{11})\, x_{i2} \\ \frac{\partial L}{\partial b_{11}} &= -(y-\hat{y})\, w^2_{11}\, O_{11}\, (1-O_{11}) \end{aligned}}$$

Now next weights

$L \to \hat{y} \to z_f \to O_{12} \to z_p \to w'_{12}$

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

$$\frac{\partial L}{\partial w'_{22}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_f} \times \frac{\partial z_f}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial z_p} \times \frac{\partial z_p}{\partial w'_{22}}$$

$$\frac{\partial L}{\partial b_{12}} = \underbrace{\boxed{\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_f}}}_{-(y-\hat{y})} \times \frac{\partial z_f}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial z_p} \times \frac{\partial z_p}{\partial b_{12}}$$

3

$$\boxed{\begin{aligned} \frac{\partial L}{\partial w'_{12}} &= -(y-\hat{y})\, w^2_{21}\, O_{12}(1-O_{12})\, x_{i1} \\ \frac{\partial L}{\partial w'_{22}} &= -(y-\hat{y})\, w^2_{21}\, O_{12}(1-O_{12})\, x_{i2} \\ \frac{\partial L}{\partial b_{12}} &= -(y-\hat{y})\, w^2_{21}\, O_{12}(1-O_{12}) \end{aligned}}$$

$$\frac{\partial z_f}{\partial O_{12}} = \frac{\partial}{\partial O_{12}} [w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}]$$

$$= w^2_{21}$$

$$\frac{\partial O_{12}}{\partial z_p} = \frac{\partial}{\partial z_p} (\sigma(z_p))$$

$$= \sigma(z_p)[1-\sigma(z_p)]$$

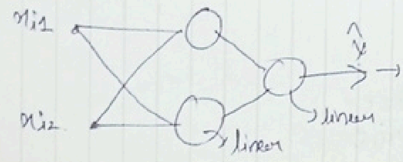
$$= O_{12}[1-O_{12}]$$

$$\frac{\partial z_p}{\partial w'_{12}} = \frac{\partial}{\partial w'_{12}} [w'_{12} x_{i1} + w'_{22} \times x_{i2} + b_{12}]$$

$$= x_{i1}$$

* Intution behind backpropagation algrithm.

* Loss f^n is a f^n of all trainable parameters.



mse $= (y-\hat{y})^2$

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

And $y \to$ constant

L is depend on $\hat{y}$

means $L(\hat{y})$

$$\hat{y} = w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}$$

$$\hat{y} = \underbrace{w^2_{11}\left[w^1_{11} x_{i1} + w^1_{21} x_{i2} + b_{11}\right] + w^2_{21}\left[w^1_{12} x_{i1} + w^1_{22} x_{i2} + b_{12}\right] + b_{12}}_{L(\hat{y})}$$

Means L (loss) is is f^n of 9 trainable paratmers

*Means if we change any paratmers the it will affect the L.

So we have to change the values of these parameters accordingly so that the loss (L) will be minimum.

* Concept of gradient :-

$\hookrightarrow$ gradient is a fancy ward for partial derivative

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

$\frac{dy}{dx} = \frac{dy}{dx}(fx) = 2x + 1$.

$y \to x$

y is depend only on x

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

$\frac{\partial z}{\partial x} = 2x$, $\frac{\partial z}{\partial y} = 2y$ here z, is depend on two variable (x, y)

gradient

when output is depend of ~~###~~ two or more than two variable. then we calculate the partial derivation and we call it gradiant.

and we have, a complex f^n

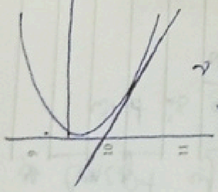
$$L(w'_{11}, w'_{12} \ldots w^2_{21}, b_{21}, b_{11}, b_{12})$$

$$\frac{\partial L}{\partial w}, \frac{\partial L}{\partial b}$$

So we have to calculate the loss w.r.t. to every 9 parameters (weight and bias)

in 2D $\to \frac{dy}{dx}$

but in 3D, 9D $\to$ there 3, or 9 diff. slopes



$\rightarrow$ deriavates is finding slope. of f^n!

November 27 Wednesday

* Concept of derivation:-

what is $\frac{dy}{dx}$ → it is rate of chance is 'y' w.r.t. 'x'

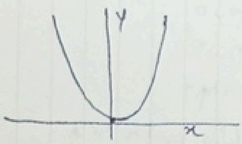
means if we change 'x' then how many units will change in 'y' in what direction

$\frac{dy}{dx} = 2$ (+ve)

mag. → 2, direction (sign) → (+ve)

$\frac{dy}{dx} = -2$ (-ve)

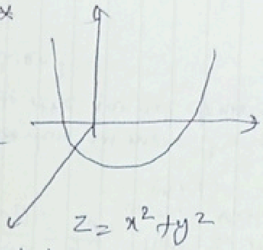
* The Concept of minima:-



$y = x^2$

$\frac{dy}{dx} = 2x = 0$

$x = 0$ → minima



$z = x^2 + y^2$

$(x, y) \rightarrow z_{min}$

$\frac{\partial z}{\partial x} = 0 \quad 2x = 0 \rightarrow x = 0$

$\frac{\partial z}{\partial y} = 0 \quad 2y = 0 \rightarrow y = 0$

Life isn't about finding yourself. Life is about creating yourself. - George Bernard Shaw

2024 Week 48 Day (332-034)

December 2024

November 26 Tuesday

2024 Week 48 Day (331-035)

November 2024

* $w_{new} = w_{old} - \eta \frac{\partial L}{\partial w}$ → this is the formula to update the parameters

* But why are we subtracting the gradient $\left(-\frac{\partial L}{\partial w}\right)$

Consider L is only depend on b_{21}

Means $L(b_{21})$, $\hat{y} \rightarrow b_{21}$

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

Change in L by one unit change in b_{21}

let $b_{21} = 5$ → old value and we have to reduce 'L' so we have to reduce something from b_{21} that's why we use '−' sign and $\frac{\partial L}{\partial b_{21}}$ is change in b_{21}

* role of learning rate (η):-

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

η → learning rate controls the value of change reaching to the minima.

generally it is $\eta \rightarrow 0.1/0.001$ so we can smoothly reach to minima.

if $\eta = 1$ → so it will take big jumps.

$\eta = 0.00001$ → the it will take long time to converge.

Be yourself; everyone else is already taken. - Oscar Wilde