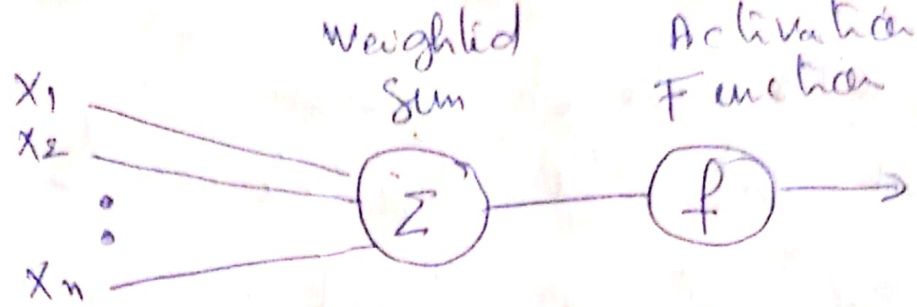


$$Z = W \times X + b$$

$$A = f(Z)$$

$\Rightarrow$ f is the Activation Function like sigmoid, Relu, tanh ...

Cost Function : $\mathcal{L}(y, A) = -y \log A - (1-y) \log(1-A)$

$$\nabla \mathcal{L} = \begin{Bmatrix} \frac{\partial \mathcal{L}}{\partial W} \\ \frac{\partial \mathcal{L}}{\partial b} \end{Bmatrix}$$

if we apply chain rule we will get :

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial A} \times \frac{\partial A}{\partial Z} \times \frac{\partial Z}{\partial W}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial A} \times \frac{\partial A}{\partial Z} \times \frac{\partial Z}{\partial b}$$

now let find those terms:

$$\frac{\partial \mathcal{L}}{\partial A}, \frac{\partial A}{\partial Z}, \frac{\partial Z}{\partial W}, \frac{\partial Z}{\partial b}$$

We already know that $Z = wx + b$ so:

$$\frac{\partial Z}{\partial w} = x \quad \text{and} \quad \frac{\partial Z}{\partial b} = 1$$

know let find $\frac{\partial \mathcal{L}}{\partial A}$ and $\frac{\partial A}{\partial Z}$:

the Activation function used in this example is sigmoid;

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$d\sigma(z) = \sigma(z) \times (1 - \sigma(z))$$

$$\frac{\partial A}{\partial Z} = A \times (1 - A)$$

$$\frac{\partial \mathcal{L}}{\partial A} = -\frac{y}{A} + \frac{1-y}{1-A}$$

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

$$= \frac{-y + \cancel{yA} + A - \cancel{Ay}}{A(1-A)}$$

$$\Rightarrow \boxed{\frac{\partial \mathcal{L}}{\partial A} = \frac{A - y}{A(1-A)}} \quad 2$$

Scanned with CamScanner

know let find $\frac{\partial \mathcal{L}}{\partial w}$ and $\frac{\partial \mathcal{L}}{\partial b}$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial A} \times \frac{\partial A}{\partial Z} \times \frac{\partial Z}{\partial w}$$

$$= \frac{A - Y}{\cancel{A(1-A)}} \times \cancel{A(1-A)} \times X$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial w} = (A - Y) \times X}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial A} \times \frac{\partial A}{\partial Z} \times \frac{\partial Z}{\partial b}$$

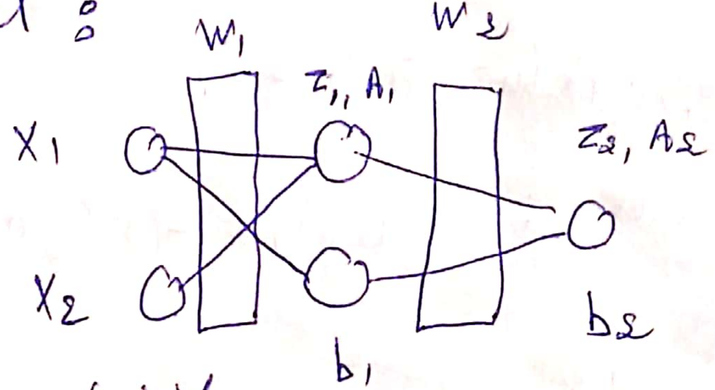
$$= \frac{A - Y}{\cancel{A(1-A)}} \times \cancel{A(1-A)} \times 1$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial b} = A - Y$$

result $\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = dw = X^T \times (A - Y) \\ \frac{\partial \mathcal{L}}{\partial b} = A - Y \end{cases}$

Know let's make it general and try to find out The update rules for the entire Model.

Suppose with me that we've this Model:



$$\nabla \mathcal{L} = \begin{pmatrix} \frac{\partial \mathcal{L}}{W_1} \\ \frac{\partial \mathcal{L}}{\partial W_2} \\ \frac{\partial \mathcal{L}}{\partial b_1} \\ \frac{\partial \mathcal{L}}{\partial b_2} \end{pmatrix}$$

$Z_1 = W_1 \times X + b_1$

$A_1 = f(Z_1)$

$Z_2 = W_2 \times A_1 + b_2$

$A_2 = \hat{Y} = f(Z_2)$

dZ_2

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial A_2} \times \frac{\partial A_2}{\partial Z_2} \times \frac{\partial Z_2}{\partial W_2} \quad \left(dZ_1 = dZ_2 \times \frac{\partial Z_2}{\partial A_1} \times \frac{\partial A_1}{\partial Z_1} \right)$$

$$\frac{\partial \mathcal{L}}{\partial b_2} = \frac{\partial \mathcal{L}}{\partial A_2} \times \frac{\partial A_2}{\partial Z_2} \times \frac{\partial Z^2}{\partial b_2} \qquad dZ_1$$

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial A_2} \times \frac{\partial A_2}{\partial Z_2} \times \frac{\partial Z_2}{\partial A_1} \times \frac{\partial A_1}{\partial Z_1} \times \frac{\partial Z_1}{\partial W_1}$$

$$\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial A_2} \times \frac{\partial A_2}{\partial Z_2} \times \frac{\partial Z_2}{\partial A_1} \times \frac{\partial A_1}{\partial Z_1} \times \frac{\partial Z_1}{\partial b_1}$$

The first thing we need to do is to find dZ_1 and dZ_2, then we can find the rest

$$dZ_2 = \frac{\partial \mathcal{L}}{\partial A_2} \times \frac{\partial A_2}{\partial W_2}$$

$$= \frac{A_2 - Y}{A_2 \cancel{(1 - A_2)}} \times A_2 \cancel{(1 - A_2)}$$

$$\Rightarrow \boxed{dZ_2 = A_2 - Y}$$

$$dZ_1 = dZ_2 \times \frac{\partial Z_2}{\partial A_1} \times \frac{\partial A_1}{\partial Z_1}$$

$$\boxed{dZ_1 = dZ_2 \times W_2 \times f'(A_1)}$$ (f is sigmoid Activation Function)

$$\frac{\partial \mathcal{L}}{\partial W_2} = dW_2 = dZ_2 \times \frac{\partial Z_2}{\partial W_2} = dZ_2 \times A_1$$

$$\frac{\partial \mathcal{L}}{\partial b_2} = db_2 = dZ_2 \times \frac{\partial Z_2}{\partial b_2} = dZ_2$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = dZ_1 \times \frac{dZ_1}{\partial w_1}$$

$$= dZ_1 \times X$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial w_1} = dw_1 = dZ_1 \times X$$

$$\frac{\partial \mathcal{L}}{\partial b_1} = dZ_1 \times \left(\frac{\partial Z_1}{\partial b_1}\right)$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial b_1} = db_1 = dZ_1$$

final result :

$$\begin{cases} dw_2 = dZ_2 \times A_1 \\ db_2 = dZ_2 \\ dw_1 = dZ_1 \times X \\ db_1 = dZ_1 \end{cases}$$

Now we need to find the relationship between (dw_1, db_1) and (dw_2, db_2) to find the update rules:

we've: $\begin{cases} dZ_2 = A_2 - Y \\ dZ_1 = W_2^T \times dZ_2 \times f'(A_1) \end{cases}$

The final update rules:

L = Number of layers

$dZ = A_L - Y$

for $i = L$ to 1

$W_i -= lr \times dZ \times A_{(i-1)}^T$ (lr is the learning rate)

$b_i -= lr \times dZ$

$dZ = W_i^T \times dZ \times f'(A_{(i-1)})$

end for