let, $X = n \times f$ ← input feat.

$y = n \times 1$ ← target labels.

1st layer → w_1 , b_1

2nd " → w_2 , b_2

3rd " → w_3 , b_3

output,
$$z_1 = w_1 x + b_1$$
$$a_1 = g(z_1)$$, g → activation function

$$z_2 = w_2 a_1 + b_2$$
$$a_2 = g(z_2)$$

$$z_3 = w_3 a_2 + b_3$$
$$a_3 = g(z_3)$$

final output,
$$\hat{y} = a_3$$

log loss / binary cross entropy loss,
$$L = -[(1-y)\log(1-\hat{y}) + y\log(\hat{y})]$$

steps:

1. initialize random w_1 , w_2 , w_3 and b_1, b_2, b_3.
2. gradient descent,
$$w_i = w_i - \alpha \frac{\delta L}{\delta w_i}$$

$$b_i = b_i - \alpha \frac{\delta L}{\delta b_i}$$

where, i refers to the ith layer.

3. Repeat until convergence.

$\frac{\delta L}{\delta W_3}$ to update the third layer,

$$\frac{\delta L}{\delta W_3} = \frac{\delta}{\delta W_3}\left[-\left[(1-y)\log(1-\hat{y}) + y\log(\hat{y})\right]\right]$$

$$= -\left[(1-y)\frac{\delta}{\delta W_3}\log(1-\hat{y}) + y\frac{\delta}{\delta W_3}\log(\hat{y})\right]$$

$$\therefore \frac{\delta L}{\delta W_3} = -\left[\frac{1-y}{1-\hat{y}} \cdot \frac{\delta}{\delta W_3}(1-\hat{y}) + \frac{y}{\hat{y}}\frac{\delta}{\delta W_3}(\hat{y})\right]$$

$$= -\left[\frac{1-y}{1-\hat{y}} \cdot \frac{\delta \hat{y}}{\delta W_3} + \frac{y}{\hat{y}} \cdot \frac{\delta \hat{y}}{\delta W_3}\right]$$

$$= \frac{1-y}{1-\hat{y}} \cdot \frac{\delta \hat{y}}{\delta W_3} - \frac{y}{\hat{y}} \cdot \frac{\delta \hat{y}}{\delta W_3}$$

$$= \left[\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right] \cdot \frac{\delta \hat{y}}{\delta W_3}$$

$$\frac{\delta L}{\delta W_3} = \frac{\hat{y}(1-y) - y(1-\hat{y})}{(1-\hat{y})\,\hat{y}} \cdot \frac{\delta \hat{y}}{\delta W_3}$$

$$= \frac{\hat{y} - y\hat{y} - y + y\hat{y}}{(1-\hat{y})\,\hat{y}} \cdot \frac{\delta \hat{y}}{\delta W_3}$$

$$\therefore \frac{\delta L}{\delta w_3} = \frac{\hat{y} - y}{(1-\hat{y})\,\hat{y}} \cdot \frac{\delta \hat{y}}{\delta w_3}$$

now, $\hat{y} = a_3 = g(z_3)$

if $g(z)$ is a sigmoid function,

$$g'(z) = g(z)\,(1 - g(z))$$

$$\boxed{\begin{aligned} \sigma(x) &= \frac{1}{1+e^{-x}} \\ \sigma'(x) &= \sigma(x)\,(1-\sigma(x)) \end{aligned}}$$

https://towardsdatascience.com/derivative-of-the-sigmoid-function-536880cf918e
Detailed derivation here

$$\therefore \frac{\delta \hat{y}}{\delta w_3} = \frac{\delta a_3}{\delta w_3} = \frac{\delta}{\delta w_3} g(z_3)$$

$$= g(z_3)\,(1 - g(z_3)) \cdot \frac{\delta z_3}{\delta w_3}$$

$$= a_3\,(1 - a_3)\,\frac{\delta}{\delta w_3}\,[w_3 a_2 + b_3]$$

$$= a_3\,(1 - a_3)\,a_2$$

$$= \hat{y}\,(1 - \hat{y}) \cdot a_2$$

$$\frac{\delta L}{\delta w_3} = \frac{\hat{y} - y}{\hat{y}\,(1-\hat{y})} \cdot [\hat{y}\,(1-\hat{y}) \cdot a_2]$$

$$\frac{\delta L}{\delta w_3} = (\hat{y} - y) \cdot a_2$$

$$\frac{\delta L}{\delta w_3} = (a_3 - y)\,a_2^T$$

similarly,

$$\frac{\delta L}{\delta b_3} = a_3 - y$$

④ Find $\frac{\delta L}{\delta w_2}$, use chain rule,

since, loss is dependant on $\hat{y} = a_3$

$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{?} \cdot \frac{?}{\delta w_2}$$

we know, $a_3 = g(z_3)$, a_3 depends on z_3

$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{\delta z_3} \cdot \frac{\delta z_3}{?} \cdot \frac{?}{\delta w_2}$$

now, $z_3 = w_3 a_2 + b_3$

z_3 depends on a_2.

$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{\delta z_3} \cdot \frac{\delta z_3}{\delta a_2} \cdot \frac{\delta a_2}{?} \cdot \frac{?}{\delta w_2}$$

But $a_2 = g(z_2)$

$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{\delta z_3} \cdot \frac{\delta z_3}{\delta a_2} \cdot \frac{\delta a_2}{\delta z_2} \cdot \frac{\delta z_2}{?} \cdot \frac{?}{\delta w_2}$$

But $z_2 = w_2 a_1 + b_2$

direct connection between z_2 and w_2.

$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{\delta z_3} \cdot \frac{\delta z_3}{\delta a_2} \cdot \frac{\delta a_2}{\delta z_2} \cdot \frac{\delta z_2}{\delta w_2}$$

rewrite $\frac{\delta L}{\delta w_3}$ using chainrule,

$$\frac{\delta L}{\delta w_3} = \underbrace{\frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{\delta z_3}}_{a_3 - y} \cdot \underbrace{\frac{\delta z_3}{\delta w_3}}_{a_2^T}$$

since,

$z_3 = w_3 a_2 + b_3$

$$\Rightarrow \frac{\delta L}{\delta w_2} = (a_3 - y) \frac{\delta z_3}{\delta a_2} \cdot \frac{\delta a_2}{\delta z_2} \cdot \frac{\delta z_2}{\delta w_2}$$

if $z_3 = w_3 a_2 + b_3$

$$\frac{\delta z_3}{\delta a_2} = w_3$$

if $a_2 = g(z_2)$

$$\frac{\delta a_2}{\delta z_2} = g'(z_2)$$

$z_1 = w_1 x + b$

$$\frac{\delta z_1}{\delta w_1} = x$$

if $z_2 = w_2 \cdot a_1 + b_2$

$$\frac{\delta z_2}{\delta w_2} = a_1 \; ; \; \frac{\delta z_2}{\delta a_1} = w_2$$

$$\therefore \frac{\delta L}{\delta w_2} = (a_3 - y) \, w_3 \, g'(z_2) \cdot a_1$$

Hence,

$$\frac{\delta L}{\delta w_2} = w_3^T g'(z_2) (a_3 - y) \cdot a_1^T .$$

Similarly, $\frac{\delta L}{\delta b_2} = w_3^T g'(z_2)(a_3 - y)$

For, $\frac{\delta L}{\delta w_1}$, the chain rule is,

$$\frac{\delta L}{\delta w_1} = \frac{\delta L}{\delta a_3} \cdot \frac{\delta a_3}{\delta z_3} \cdot \frac{\delta z_3}{\delta a_2} \cdot \frac{\delta a_2}{\delta z_2} \cdot \frac{\delta z_2}{\delta a_1} \cdot \frac{\delta a_1}{\delta z_1} \cdot \frac{\delta z_1}{\delta w_1}$$

$$\frac{\delta L}{\delta w_1} = (a_3 - y) \cdot w_3 \cdot g'(z_2) \cdot w_2 \cdot g'(z_1) \cdot x$$

similarly,

$$\frac{\delta L}{\delta b_1} = (a_3 - y) \cdot w_3 \cdot g'(z_2) \cdot w_2 \cdot g'(z_1)$$