$Z_1 = w_1 x + b_1$

$a_1 = g(z_1)$ → Sigmoid → for instance

$g(z_1) = Sig(z_1)$

$Z_2 = w_2 a_1 + b_2$

$a_2 = F(z_2)$ → linear ~~function~~ function

it does not matter

just a number

$i=2$

$MSE = \frac{1}{N} \sum_{i=1}^{n} (y - a_i)^2$ $\boxed{2 \text{ layer}}$

what is C: Loss function or Mean square Error

$$\frac{\partial L}{\partial w_2} = \frac{\partial}{\partial w_2}\left[\Sigma (y - a_2)^2\right] =$$

$$= \Sigma \frac{\partial}{\partial w_2} (y - a_2) \times (y - a_2)]$$

$$= \Sigma\left[(a_2 - y)\left(\frac{\partial a_2}{\partial w_2}\right)\right] \rightarrow \boxed{\frac{\partial L}{\partial w_2} = \Sigma (a_2 - y)(a_1)}$$

$$\frac{\partial a_2}{\partial z_2} \times \frac{\partial z_2}{\partial w_2}$$

a_1

$$\frac{\partial L}{\partial b_2} = \sum \frac{\partial}{\partial b_2}(y-a_2)^2 = \sum\left[2(y-a_2)\left(\frac{\partial y}{\partial b_2} - \frac{\partial a_2}{\partial b_2}\right)\right]$$

$$\frac{\partial a_2}{\partial z_2} \times \frac{\cancel{\partial z_2}}{\cancel{\partial b_2}} \quad 1$$

$$\frac{\partial L}{\partial b_2} = \sum\left[(a_2-y)\right] \checkmark$$

③

$$\frac{\partial L}{\partial b_1} = \frac{\partial}{\partial b_1}\left[\Sigma (y-a_2)^2\right] = \left[\frac{\partial}{\partial b_1}(y-a_2)(y-a_2)\right]$$

$$\frac{\partial L}{\partial b_1} = \Sigma\left[-(y-a_2)\left(\frac{\partial}{\partial b_1} a_2\right)\right]$$

$$\frac{\partial a_2}{\partial b_1} = \frac{\partial a_2}{\partial z_2} \times \underbrace{\frac{\partial z_2}{\partial a_1}}_{W_2} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z_1}{\partial b_1}$$

$$\frac{\partial L}{\partial b_1} = \Sigma (a_2-y)\, W_2\, a_1 (1-a_1)$$

In Practice, we don't normally use sigmoid as an activation function for our hidden layers. we use ReLu as an activation function for our hidden layers.

Log loss :

$$-\left(y \log(P) + (1-y) \log(1-p)\right)$$

It measures the performance of classification model whose output is a Probability value between 0 and 1.

for instance. update rule:

$$\frac{\partial L}{\partial w_2} = (a_1 - y)\, a_1^T$$

The results between two methods is Comparable.

$$\frac{\partial L}{b_2} = a_2 - y$$

in other words,

$$L = \frac{1}{2}\sum(\hat{y} - y)^2 \qquad \text{2layer}$$

$$\rightarrow g(z_2) = w_2 a_1 + b_1$$

$$T = w_2 a_1 + b_2 - y$$

$$\frac{\partial T}{\partial b_2} = 1$$

$$\frac{\partial L}{\partial T} = \hat{y} - y \quad \longrightarrow \quad \frac{\partial L}{\partial w_2} = a_1 \cdot (\hat{y} - y)$$

[redacted] $$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$\hat{y} = w_2 a_1 + b_2 \qquad \frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b_2} \qquad T = y - \hat{y}$$

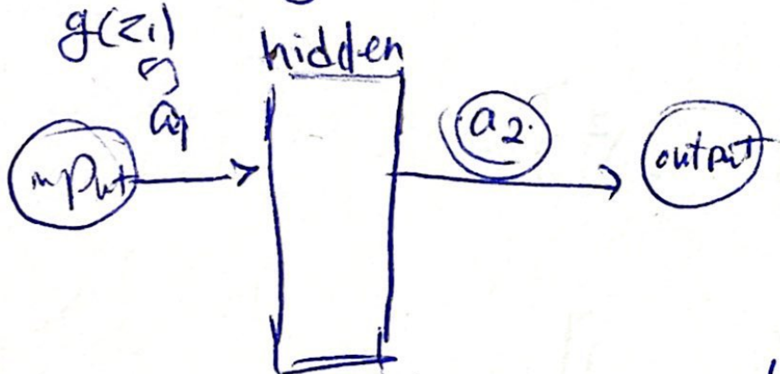
$$\frac{\partial L}{\partial y} = -2(y - \hat{y})$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{b_1} \rightarrow -2(y - \hat{y})\, w_2 \cdot g'(z_1) \qquad \frac{\partial L}{\partial \hat{y}} = -2(y - \hat{y})$$

$$\frac{\partial a_1}{\partial z_1} = g'(z_1)$$

$$\frac{\partial L}{\partial w_1} = -2(y-\hat{y}) \cdot w_2 \cdot g'(z_1) \cdot x$$

explain briefly the difference:
how to train a 2 layer with one hidden layer would be like this:



binary classification based on our lecture. sigmoid will be used for output. on the other hand, linear function will be used in the output for regression. ~~the practice, we normally do not use sigmoid as an ac~~