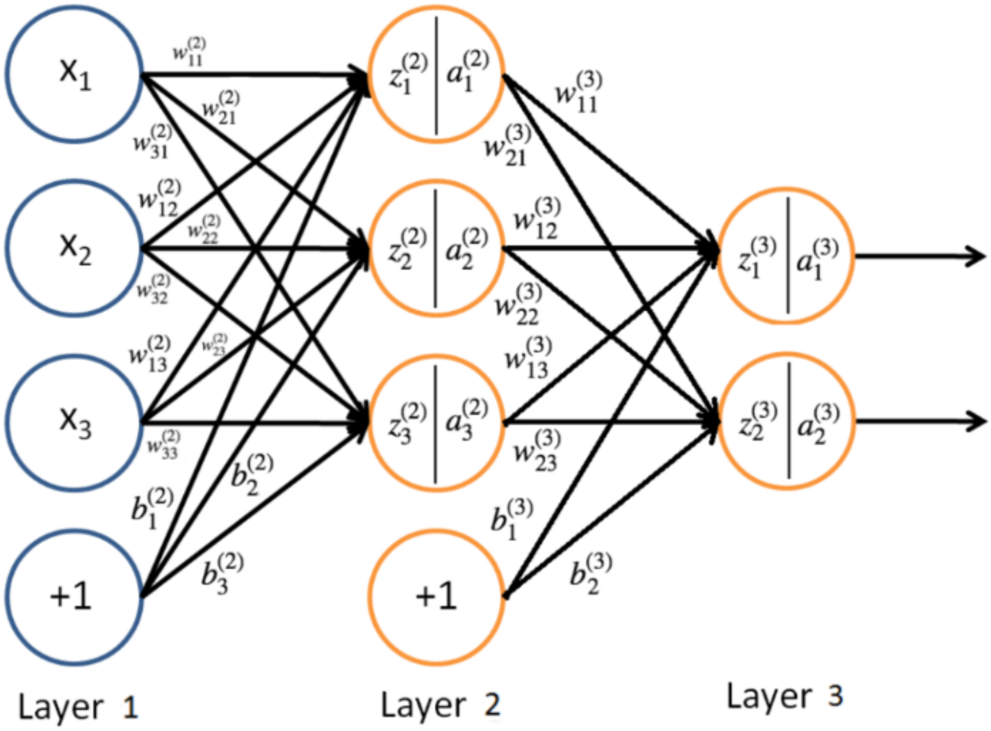

x_1
x_2
x_3
+1
$w_{11}^{(2)}$
$w_{21}^{(2)}$
$w_{31}^{(2)}$
$w_{12}^{(2)}$
$w_{22}^{(2)}$
$w_{32}^{(2)}$
$w_{13}^{(2)}$
$w_{23}^{(2)}$
$w_{33}^{(2)}$
$b_1^{(2)}$
$b_2^{(2)}$
$b_3^{(2)}$
$z_1^{(2)}$ $a_1^{(2)}$
$z_2^{(2)}$ $a_2^{(2)}$
$z_3^{(2)}$ $a_3^{(2)}$
+1
$w_{11}^{(3)}$
$w_{21}^{(3)}$
$w_{12}^{(3)}$
$w_{22}^{(3)}$
$w_{13}^{(3)}$
$w_{23}^{(3)}$
$b_1^{(3)}$
$b_2^{(3)}$
$z_1^{(3)}$ $a_1^{(3)}$
$z_2^{(3)}$ $a_2^{(3)}$
Layer 1
Layer 2
Layer 3

loss function: (Ordinary Least Squares) Back Propagation Algorithm

$$E(i) = \frac{1}{2}\|\mathbf{y}^{(i)} - \mathbf{o}^{(i)}\| = \frac{1}{2}\sum_{k=1}^{n_L}(y_k^{(i)} - o_k^{(i)})^2$$

1> where $\mathbf{y}^{(i)}$ is the ground truth, $\mathbf{o}^{(i)}$ is the output of neural network.

2> The coefficient $\frac{1}{2}$ is unnecessary, which is for calculating easily in the future.

3> From the given picture of a neural network, let $n_L = 2$, $\mathbf{y}^{(i)} = (y_1^{(i)}, y_2^{(i)})^T$.

Thus, $E(i) = \frac{1}{2}(y_1^{(i)} - a_1^{(3)})^2 + \frac{1}{2}(y_2^{(i)} - a_2^{(3)})^2$.

If we go further into hidden layer,

$$E(i) = \frac{1}{2}(y_1^{(i)} - f(\overbrace{w_{11}^{(3)}a_1^{(2)} + w_{12}^{(3)}a_2^{(2)} + w_{13}^{(3)}a_3^{(2)} + b_1^{(3)}}^{z_1^{(3)}}))^2 + \frac{1}{2}(y_2^{(i)} - f(\underbrace{w_{21}^{(3)}a_1^{(2)} + w_{22}^{(3)}a_2^{(2)} + w_{23}^{(3)}a_3^{(2)} + b_2^{(3)}}_{z_2^{(3)}}))^2.$$

If we go back further again into the input layer, $a_1^{(2)}, a_2^{(2)}, a_3^{(2)}$ can be replaced.

4> $$E_{total} = \frac{1}{N}\sum_{i=1}^{N} E(i)$$

The goal is to adjust the weights and [redacted] biases (W, b), decreasing the loss and finally aquiring the best weights and biases when the minimum of loss occurs.

5> Calculating the $\boxed{\text{derivatives of weights in output layer}}$ by using derived chain rules.

$$\frac{\partial E}{\partial w_{11}^{(3)}} = \partial\left[\frac{1}{2}(y_1^{(3)} - a_1^{(3)})^2 + \frac{1}{2}(y_2^{(3)} - a_2^{(3)})^2\right] / \partial w_{11}^{(3)}$$

$$= 2\times\frac{1}{2}(y_1^{(3)} - a_1^{(3)})\left(-\frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}}\right) + 0 \qquad \because a_1^{(3)} = f(z_1^{(3)})$$

$$= -(y_1^{(3)} - a_1^{(3)})\, f'(z_1^{(3)}) \cdot \frac{\partial z_1^{(3)}}{\partial w_{11}^{(3)}} \quad \Rightarrow \quad = f(w_{11}^{(3)}a_1^{(2)} + w_{12}^{(3)}a_2^{(2)} + w_{13}^{(3)}a_3^{(2)} + b_1^{(3)})$$

$$= -(y_1^{(3)} - a_1^{(3)})\, f'(z_1^{(3)})\, a_1^{(2)} \quad (1) \qquad \therefore \frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} = f'(z_1^{(3)})\frac{\partial z_1^{(3)}}{\partial w_{11}^{(3)}}$$

$$\because z_1^{(3)} = w_{11}^{(3)}a_1^{(2)} + w_{12}^{(3)}a_2^{(2)} + w_{13}^{(3)}a_3^{(2)} + b_1^{(3)}$$

$$\therefore \frac{\partial z_1^{(3)}}{\partial w_{11}^{(3)}} = a_1^{(2)}$$

$\boxed{\text{Let } \delta_i^{(l)} = \frac{\partial E}{\partial z_i^{(l)}},}$ we will have:

$$\frac{\partial E}{\partial w_{11}^{(3)}} = \frac{\partial E}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial w_{11}^{(3)}} = \delta_1^{(3)} a_1^{(2)} \quad (2)$$

$$(1)(2) \Rightarrow \delta_1^{(3)} = -(y_1^{(3)} - a_1^{(3)})\, f'(z_1^{(3)}), \qquad \delta_2^{(3)} = -(y_2^{(3)} - a_1^{(3)})\, f'(z_2^{(3)})$$

$$\frac{\partial E}{\partial w_{12}^{(3)}} = \delta_1^{(3)} a_2^{(2)}, \quad \frac{\partial E}{\partial w_{13}^{(3)}} = \delta_1^{(3)} a_3^{(2)}$$

$$\frac{\partial E}{\partial w_{21}^{(3)}} = \delta_2^{(3)} a_1^{(2)}, \quad \frac{\partial E}{\partial w_{22}^{(3)}} = \delta_2^{(3)} a_2^{(2)}, \quad \frac{\partial E}{\partial w_{23}^{(3)}} = \delta_2^{(3)} a_3^{(2)}$$

Generally, assuming the number of layers is L, thus,

$$\delta_i^{(L)} = -(y_i - a_i^{(L)}) f'(z_i^{(L)}) \quad (1 \le i \le n_L) \Leftrightarrow \boldsymbol{\delta}^{(L)} = -(\mathbf{y} - \mathbf{a}^{(L)}) \odot f'(\mathbf{z}^{(L)})$$

L: the output layer

i: i^{th} neure

n_L: the number of neure in the L layer

$$\frac{\partial E}{\partial w_{ij}^{(L)}} = \delta_i^{(L)} a_j^{(L-1)} \quad (1 \le i \le n_L,\ 1 \le j \le n_{L-1}) \Leftrightarrow \nabla_{W^{(L)}} E = \boldsymbol{\delta}^{(L)} (\mathbf{a}^{(L-1)})^T$$

6> calculating the $\boxed{\text{derivatives of weights in hidden layer}}$.

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = \frac{\partial E}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}} = \delta_i^{(l)} a_j^{(l-1)} \quad (2 \le l \le L-1)$$

$$\delta_i^{(l)} = \frac{\partial E}{\partial z_i^{(l)}} = \sum_{j=1}^{n_{l+1}} \frac{\partial E}{\partial z_j^{(l+1)}} \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}}$$

$$= \sum_{j=1}^{n_{l+1}} \delta_j^{(l+1)} \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}}$$

$$\because z_j^{(l+1)} = \sum_{i=1}^{n_l} w_{ji}^{(l+1)} a_i^{(l)} + b_j^{(l+1)} = \sum_{i=1}^{n_l} w_{ji}^{(l+1)} f(z_i^{(l)}) + b_j^{(l+1)}$$

$$\therefore \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}} = \frac{\partial z_j^{(l+1)}}{\partial a_i^{(l)}} \frac{\partial a_i^{(l)}}{\partial z_i^{(l)}} = W_{ji}^{(l+1)} f'(z_i^{(l)})$$

$$\therefore \delta_i^{(l)} = \left(\sum_{j=1}^{n_{l+1}} \delta_j^{(l+1)} w_{ji}^{(l+1)} \right) f'(z_i^{(l)}) \Leftrightarrow \boldsymbol{\delta}^{(l)} = ((\mathbf{W}^{(l+1)})^T \boldsymbol{\delta}^{(l+1)}) \odot f'(\mathbf{z}^{(l)})$$

Thus, the $\delta^{(l+1)}$ in $l+1$ layer is calculated by $\delta^{(l)}$ in l layer, that is the reason why it's called Back Propagation.

7> derivatives of biases in output layer and hidden layer.

$$\frac{\partial E}{\partial b_i^{(l)}} = \frac{\partial E}{\partial z_{(i)}^{(l)}} \frac{\partial z_{(i)}^{(l)}}{b_i^{(l)}} = \delta_i^{(l)} \Leftrightarrow \nabla_{b^{(l)}} E = \boldsymbol{\delta}^l$$

8> Batch Gradient Descent:

$$\mathbf{w}^{(l)} = \mathbf{w}^{(l)} - \frac{\mu}{N} \sum_{i=1}^{N} \frac{\partial E_{(i)}}{\partial \mathbf{w}^{(l)}}$$

$$\mathbf{b}^{(l)} = \mathbf{b}^{(l)} - \frac{\mu}{N} \sum_{i=1}^{N} \frac{\partial E_{(i)}}{\mathbf{b}^{(l)}}$$