Assignment 2

① Let x be the input features : $[n \times f]$
y be the output (target) labels
: $[n \times 1]$

Defining the weights [parameters] of the neural networks.

First layer → w_1 , bias → b_1
Second layer → w_2 , bias → b_2
Third layer → w_3 , bias → b_3

The output of layer1 is

First layer

$$z_1 = w_1 x + b_1$$
$$a_1 = \text{sigmoid}(z_1)$$

Second layer (output)

$$z_2 = w_2 a_1 + b_2$$
$$a_2 = \text{Linear}(z_2) \quad [\because \text{ we need regression}]$$

The output of the neural network is

$$\hat{y} = a_2$$

Loss function (MSE) $= \frac{1}{n} \sum_{i=1}^{n} (y-\hat{y})^2$

To learn the parameters.

(1) Provide random values for weights w_1, w_2 and biases b_1, b_2, b_3

(2) Update the weights using gradient descent by

$$w_i = w_i - \alpha \cdot \frac{\partial L}{\partial w_i}$$

$$b_i = b_i - \alpha \cdot \frac{\partial L}{\partial b_i}$$

$$i = 1, 2$$

(3) repeat until converge

To find $\frac{\partial L}{\partial w_2}$ to update the second layer.

$$\frac{\partial L}{\partial w_2} = \frac{\partial}{\partial w_2}\left[\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2\right]$$

$$= \frac{\partial}{\partial w_2}\left[y-\hat{y}\right]^2 = 2(y-\hat{y})\cdot\frac{\partial}{\partial w_2}(y-\hat{y})$$

$$= -2(y-\hat{y})\cdot\frac{\partial \hat{y}}{\partial w_2}$$

We know that $\hat{y} = a_2 = g(z_2)$

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

$$\frac{\partial \hat{y}}{\partial w_2} = \frac{\partial (a_2)}{\partial w_2} = \frac{\partial (g(z_2))}{\partial w_2} = \frac{\partial (w_2 a_1 + b_2)}{\partial w_2}$$

$$= a_1$$

$$\therefore \frac{\partial L}{\partial w_2} = -2(y-\hat{y})\cdot a_1 = 2(a_2-y)\cdot a_1$$

Similarly for $\frac{\partial L}{\partial b_2} = \frac{\partial}{\partial b_2}(y-y) = -2\cdot(y-y)\cdot\frac{\partial y}{\partial w_2}$

We know that $\hat{y} = a_2 = g(z_2)$

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

$$\frac{\partial(\hat{y})}{\partial b_2} = \frac{\partial(a_2)}{\partial b_2} = \frac{\partial(g(z_2))}{\partial b_2} = \frac{\partial(w_2 a_1 + b_2)}{\partial b_2}$$

$$= 1$$

$$\therefore \frac{\partial L}{\partial b_2} = -2(y-\hat{y})\cdot 1 = 2(a_2 - y).$$

To find $\frac{\partial L}{\partial w_1}$, we need chain rule of differentiation as there is no direct connection between loss 'L' and weight w_1.

→ We know that loss is dependent on $\hat{y} = a_2 = g(z_2)$

$$\Rightarrow \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$= 2(a_2 - y) \cdot 1 \cdot \frac{\partial}{\partial a_1}(w_2 a_1 + b_2) \cdot (1-a_1)a_1 \cdot w_2 \cdot x_1$$

$\because a_1$ is Sigmoid.

$$\Rightarrow \frac{\partial (L)}{\partial b_1} = \frac{\partial L}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1}.$$

$$\boxed{\frac{\partial L}{\partial b_1} = 2(a_2 - y)a_1(1-a_1) \cdot w_2}$$

→ Since only the output activation function is linear regression, the updates are twice as much as classification's update.