Let x be the input with f features.

We have two layers.

Hidden layer : W_1 weight, b_1 bias
Output layer : W_2 weight, b_2 bias

We have,

$$z_1 = W_1 x + b_1$$
$$a_1 = \sigma(z_1) \quad \text{[Sigmoid function]}$$
$$z_2 = W_2 a_1 + b_2$$
$$\hat{y} = z_2$$

We have linear function for the output because this is the appropriate activation function for continuous output (regression).

$$L = MSE = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

$$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} = (\hat{y} - y)$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial W_2}$$

$$= (\hat{y} - y) \cdot a_1$$

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial b_2}$$

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

$$= (\hat{y} - y)$$

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

$$= (\hat{y} - y) \cdot W_2 \cdot \sigma'(z_1)$$

where, $\sigma'(z_1)$ is the derivative of $\sigma(z_1)$ function.

$$\sigma'(z_1) = \sigma(z_1)\,(1 - \sigma(z_1))$$

$$\therefore \frac{\partial L}{\partial z_1} = (\hat{y} - y)\, W_2 \cdot \sigma(z_1)\,(1 - \sigma(z_1))$$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial W_1}$$

$$= (\hat{y} - y) \cdot W_2 \cdot X \cdot \sigma'(z_1)$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial W_1}$$

$$= (\hat{y} - y) \cdot W_2 \cdot \sigma'(z_1)$$

Update rules

$$W_1 = W_1 - \alpha \frac{\partial L}{\partial W_1}$$

$$= W - \alpha \cdot (\hat{y} - y) \cdot W_2 \cdot X \cdot \sigma'(z_1)$$

$$b_1 = b_1 - \alpha \frac{\partial L}{\partial b_1}$$

$$= b_1 - \alpha \cdot (\hat{y} - y) \cdot W_2 \cdot \sigma'(z_1)$$

$$W_2 = W_2 - \alpha \frac{\partial L}{\partial W_2}$$

$$= W - \alpha (\hat{y} - y) \cdot a_1$$

$$b_2 = b_2 - \alpha \frac{\partial L}{\partial b_2}$$

$$= b_2 - \alpha (\hat{y} - y)$$

Using there rules, we will perform backpropagation and train the neural network.

This is different from binary classification, because we are using identity function in the output to ensure regression instead of any other activation function that works for discrete classification.