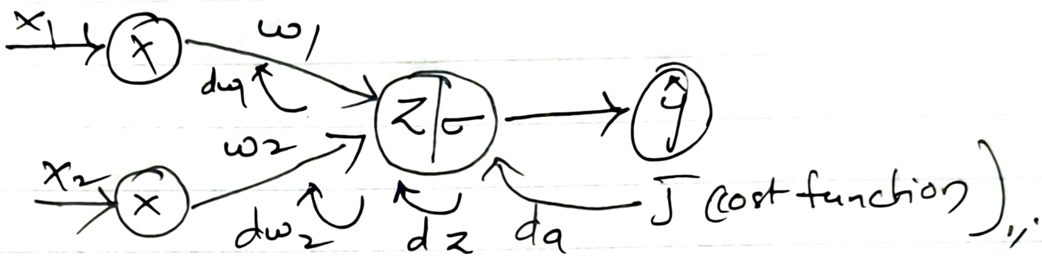
Question-1

For a 2 layer Neural Network with 1 hidden layer and 1 output layer with sigmoid as a activation function ~~but~~ in the hidden layer can be pictured as



Here is the steps for forward propagation

$$Z = W^T X + b.$$

we can derive forward propagation as

$$z^1 = W^{(1)} x + b^1$$
$$A^1 = \sigma(z^1) \quad \sigma \text{ is sigmoid.}$$
$$z^2 = W^{(2)} A^{(1)} + b^1$$
$$A^2 = \sigma(z^2).$$

Gradient descent to update the weights

$$W = W - \alpha \frac{dJ}{dW}.$$

Gradient descent can be calculated using backpropagation by applying 2nd order derivative. We need to pass output of the forward propagation to the backpropagation and calculate the error which is ~~∂f~~

$$z^2 = W^2 A^1 + b^2$$

derivative of z^2 w.r.t w^2

$$\frac{dj}{dw^2} = \left[-Y + \cancel{YA^2} + A^2 - \cancel{YA^2}\right]\left[A^1\right]$$

$$= \left[A^2 - Y\right]\left[A^1\right]$$

$$= dz^2 A^1$$

$$\frac{dj}{db^2} = \frac{dj}{dA^2} \cdot \frac{dA^2}{dz^2} \cdot \frac{dz^2}{db^2}$$

$$= dz^2$$

$$\frac{dj}{dw^1} = \frac{dj}{dA^2} \cdot \frac{dA^2}{dz^2} \cdot \frac{dz^2}{dA^1} \cdot \frac{dA^1}{dz^1} \cdot \frac{dz^1}{dW^1}$$

$$= \frac{dj}{dz^2} \cdot \frac{dz^2}{dA^1} \cdot \frac{dA^1}{dz^1} \cdot \frac{dz^1}{dw^1}$$

$$= [A^2 - Y][w^2][g'(z^1)][A^0]$$

$$= dz^2 w^2 g'(z^1) A^0$$

$$= dz^1 A^0$$

$$\frac{dz}{dA^1} = \frac{dj}{dA^2} \cdot \frac{dA^2}{dz^2} \cdot \frac{dz^2}{dA^1} \Rightarrow \frac{dj}{dz^2} w^2 \Rightarrow dz^2 \cdot w^2$$

$$\frac{dj}{db^1} = \frac{dj}{dA^2} \cdot \frac{dA^2}{dz^2} \cdot \frac{dz^2}{dA^1} \cdot \frac{dA^1}{dz^1} \cdot \frac{dz^1}{db^1}$$

$$= \frac{dj}{dz^2} \cdot \frac{dz^2}{dA^1} \cdot \frac{dA^1}{dz^1} \cdot \frac{dz^1}{db^1}$$

$$= [A^2 - Y][w^2][g'(z^1)][1]$$

$$= dz^2 w^2 g'(z^1) = dz^1.$$

Now, we need to add activation function for every derivative of backpropagation.

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} f'(x) \text{ where } f'(x) \text{ is derived}$$

from every neuron $\Rightarrow Y = [f(x^1) f(x^2) \cdots f(x^i)]$

$$= f(x)$$

Sigmoid activation function:-

$$\sigma(x) = 1/1+e^{-x}$$

after derivating this function

$$\sigma'(x) = (1+e^{-x})(1-(1+e^{-x})).$$

which can be termed as

$$\sigma'(x) = x(1-x)$$

Now by updating weights for every layer we need to calculate error using mean squared error.

$$MSE = 1/n \sum_{i}^{n} (y_i^* - y_i)^2$$

→ If we use sigmoid in the binary classification problem it will be hard to interpret the result as sigmoid will return log values and binary classification will always expect binary value of either '0' or '1'.

Sigmoid will always gives probability ranging from 0 to 1.