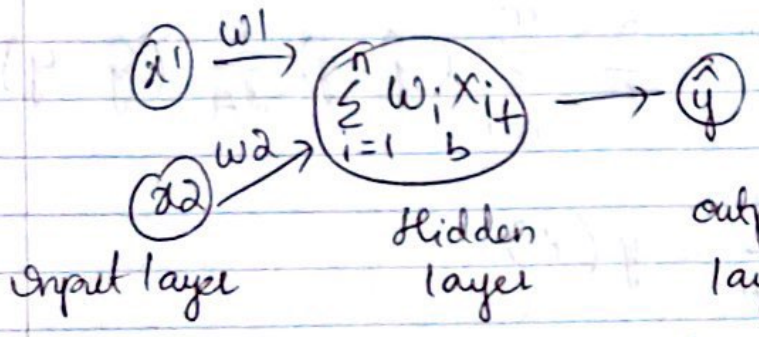
Deepthi Mikkilineni
CWID: A20349205.

Feed forward:

$\Rightarrow$



[$\because \hat{y}$ is the predicted value]

So as of the question and concepts

$$Z(i) = W(i)X + b$$ [$\because$ Here W = weight and b = bias]

$$A^{(i)} = g(Z(i))$$ [$\because$ A is the activation]

By saying that mean Square error $= \frac{1}{n}\sum_{i=1}^{n} (\hat{y}-y)^2$

So the sigmoid activation function being used for the hidden layers then as of the below function

The MSE is the Average of (predicted value − actual value)^2

Now the sigmoid activation will be

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

Scanned with CamScanner

Chain rule in order to get the output is

$$\rightarrow \frac{dl}{dw} = \frac{dL}{dA} \times \frac{dA}{dz}$$

$$\frac{dl}{dA} = \frac{1}{n} \frac{d}{dA} \sum_{i=1}^{m} (\hat{y} - y)^2$$

$$= \frac{2}{n} \sum_{i=1}^{m} (\hat{y} - y) \frac{d}{dA} (\hat{y} - y)$$

$$\frac{dz}{dw} = g'(z)$$

We know the Sigmoid function

$$g'(z) = g(z)(1 - g(z))$$

$$\frac{dL}{dw} = \frac{2}{n} (\hat{y} - y) \sum_{i=1}^{m} \frac{d}{dA} (\hat{y} - y) \left(\frac{dz}{dw}\right)$$

Assume $\hat{y} = a$

$$\frac{dl}{dw} = \frac{2}{n} (o_i - y) \frac{do_i}{dA} \frac{dz}{dw}$$

⇒ While coming to the concept of regression, the output layer uses the linear activation and the difference between both using log loss and MSE as the loss function, is that MSE cannot be used to measure the probability to find an outcome to the defined problem. Its evident from the above that updates for regression is twice as much as the classification.

Scanned with CamScanner