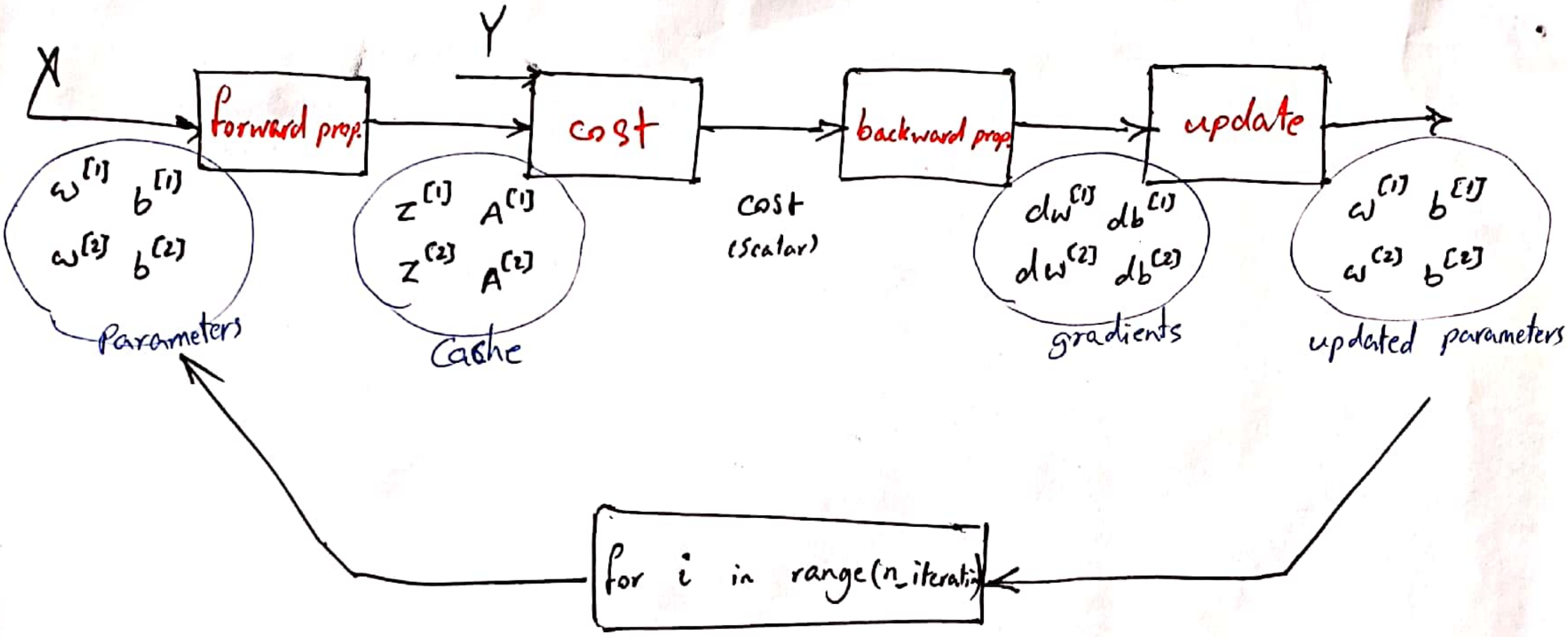
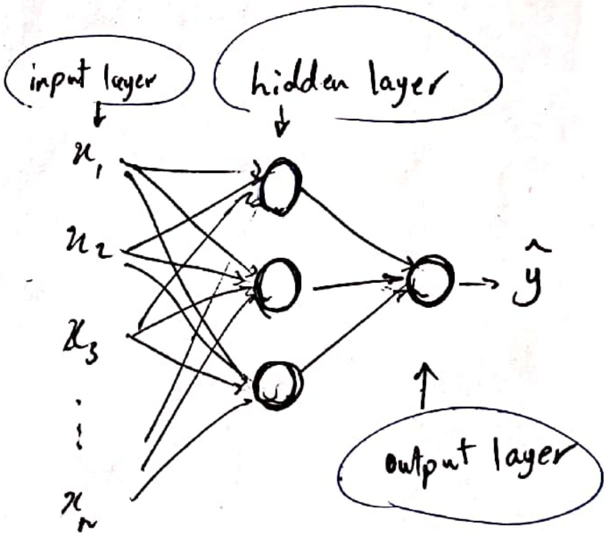


Forward propagation

$$Z_1^{[1]} = w_1^{[1]T} \cdot X + b_1^{[1]} \quad , \quad a_1^{[1]} = \sigma(Z_1^{[1]})$$

$$Z_2^{[1]} = w_2^{[1]T} \cdot X + b_2^{[1]} \quad , \quad a_2^{[2]} = \sigma(Z_2^{[1]})$$

$$Z^{[1]} = \begin{bmatrix} - w_1^{[1]T} - \\ - w_2^{[1]T} - \\ \vdots \\ - w_{n_1}^{[1]T} - \end{bmatrix}_{n_1 \times 2} \cdot \overset{X}{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{2\times 1} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ \vdots \\ b_{n_1}^{[1]} \end{bmatrix}$$



$$A^{[1]} = \sigma(Z^{[1]})$$

$$Z^{[2]} = \begin{bmatrix} - w_1^{[2]T} - \end{bmatrix}_{1\times n_1} \begin{bmatrix} A_1^{[1]} \\ A_2^{[1]} \\ \vdots \\ A_{n_1}^{[1]} \end{bmatrix}_{n_1 \times 1} + \begin{bmatrix} b_1^{[2]} \end{bmatrix}_{1\times 1}$$

$$\hat{Y}_{1\times 1} = Z^{[2]}$$

Back propagation

$$\hat{Y}$$

$$\mathcal{L} = \frac{1}{2m}\left(A^{[2]} - Y\right)^T\left(A^{[2]} - Y\right)$$

$$\boxed{\hat{Y} = A^{[2]} = Z^{[2]}}$$

the last layer (output layer) has linear activation function

$$dZ^{[2]} = \frac{\partial \mathcal{L}}{\partial Z^{[2]}} = \frac{\partial \mathcal{L}}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} = \frac{\partial \mathcal{L}}{\partial A^{[2]}} = A^{[2]} - Y$$

$$\Rightarrow \boxed{dZ^{[2]} = A^{[2]} - Y}$$

$$\boxed{Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]}}$$

$$dW^{[2]} = \frac{\partial \mathcal{L}}{\partial W^{[2]}} = \frac{\partial \mathcal{L}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} = dZ^{[2]} \cdot A^{[1]T}$$

$$\Rightarrow \boxed{dW^{[2]} = \frac{1}{m}\left(A^{[2]} - Y\right) A^{[1]T}}$$

$$db^{[2]} = \frac{\partial \mathcal{L}}{\partial b^{[2]}} = \frac{\partial \mathcal{L}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial b^{[2]}} = \frac{\partial \mathcal{L}}{\partial Z^{[2]}} \cdot 1 = A^{[2]} - Y$$

$$\Rightarrow \boxed{db^{[2]} = \frac{1}{m} \sum^{\text{columns}} \left(A^{[2]} - Y\right)}$$

$$X$$

$$\boxed{Z^{[1]} = W^{[1]} A^{[0]} + b^{[1]}}$$

$$dZ^{[1]} = \frac{\partial \mathcal{L}}{\partial Z^{[1]}} = \frac{\partial \mathcal{L}}{\partial A^{[1]}} \cdot \frac{\partial A^{[1]}}{\partial Z^{[1]}} = dZ^{[2]} \cdot W^{[2]} \cdot \sigma(Z^{[1]})(1 - \sigma(Z^{[1]}))$$

$$\Rightarrow \boxed{dZ^{[1]} = W^{[2]T}\left(A^{[2]} - Y\right) * \sigma(Z^{[1]})\left(1 - \sigma(Z^{[1]})\right)}$$

$$dA^{[1]} = \frac{\partial \mathcal{L}}{\partial A^{[1]}} = \frac{\partial \mathcal{L}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial A^{[1]}}$$

$$\frac{\partial \mathcal{L}}{\partial A^{[1]}} = \boxed{dA^{[1]} = dZ^{[2]}\, W^{[2]}}$$

$$dW^{[1]} = \frac{\partial \mathcal{L}}{\partial W^{[1]}} = \frac{\partial \mathcal{L}}{\partial Z^{[1]}} \cdot \frac{\partial Z^{[1]}}{\partial W^{[1]}} = dZ^{[1]} \cdot A^{[0]}$$

$$\Rightarrow \boxed{dW^{[1]} = \frac{1}{m} W^{[2]T}\left(A^{[2]} - Y\right) * \sigma(Z^{[1]})\left(1 - \sigma(Z^{[1]})\right) X^T}$$

$$A^{[1]} = \sigma(Z^{[1]})$$

$$\Rightarrow \boxed{\frac{\partial A^{[1]}}{\partial Z^{[1]}} = \sigma(Z^{[1]})\left(1 - \sigma(Z^{[1]})\right)}$$

$$db^{[1]} = \frac{\partial \mathcal{L}}{\partial b^{[1]}} = \frac{\partial \mathcal{L}}{\partial Z^{[1]}} \cdot \frac{\partial Z^{[1]}}{\partial b^{[1]}} = dZ^{[1]} \cdot 1$$

$$\Rightarrow \boxed{db^{[1]} = \frac{1}{m} \sum^{\text{columns}} W^{[2]T}\left(A^{[2]} - Y\right) * \sigma(Z^{[1]})\left(1 - \sigma(Z^{[1]})\right)}$$

In binary classification we have:

$$\mathcal{L} = -y \log \hat{y} - (1-y)\log(1-\hat{y}) \quad \text{or} \quad \mathcal{L} = -y\log(a) - (1-y)\log(1-a)$$

$$\frac{\partial \mathcal{L}}{\partial a} = \frac{-y}{a} + \frac{1-y}{1-a} \Rightarrow \boxed{da = \frac{-y}{a} + \frac{1-y}{1-a}}$$

$$a = \sigma(z) \Rightarrow \frac{\partial a}{\partial z} = \sigma(z)(1-\sigma(z)) = \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) = \frac{e^{-z}}{(1+e^{-z})^2} = \boxed{a(1-a)}$$

$$dz = \frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z} = da \frac{\partial a}{\partial z} = \left(\frac{-y}{a} + \frac{1-y}{1-a}\right)(a(1-a)) = \boxed{a - y}$$

Therefore, the derivatives of the parameters are the same as regression.

Also, Regression and binary classification have the same update rules.

The difference between these two algorithms is the output layer activation functions which is sigmoid in case of binary classification while it is linear for regression.

Another difference between two algorithms is their loss function which is cross antropy for binary classification and it is mean square error for regression. But we saw that the same parameter derivatives were derived for both methods.