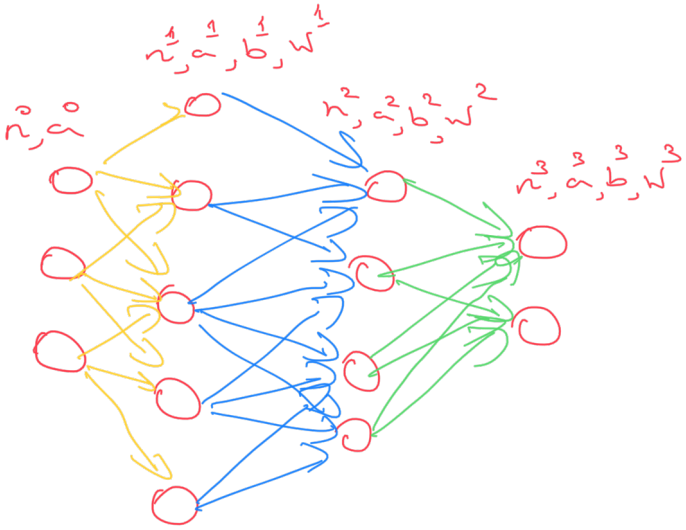
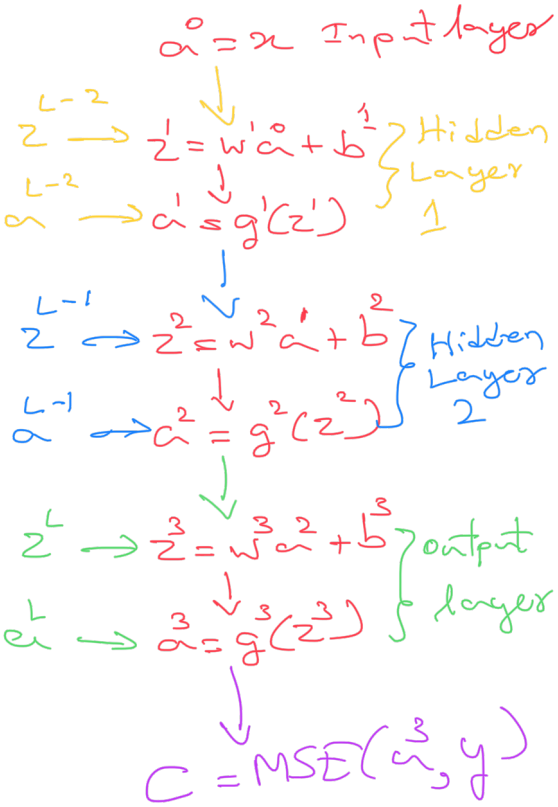
Gradient descent and back-propagation for deep learning models





n: number of neurons
W: layer weight matrix
b: layer bias vector
a: layer activation vector
z: layer input vector
σ: activation function

Layer "l" input before the activation

$$z^l = W^l * a^{l-1} + b$$

$$a^l = \sigma(z^l)$$

Cost function for a single sample: $mse = \frac{1}{2}(a^L - y)^2$

Where a^L is the output of the final layer (output layer) and y is the true values

** Goal of backpropagation: find the gradient of the cost w.r.t to the parameters W and b

** Once we have the partial derivatives we can compute the parameters for the next training iteration

Parameter update rule:

$$W^l_{new} = W^l_{old} - \alpha \frac{\partial C}{\partial W} \qquad -1$$

$$b^l_{new} = b^l_{old} - \alpha \frac{\partial C}{\partial b} \qquad -2$$

Neural network model as a set of composition functions

$$C(a^3(z^3(w^3, b^3, a^2(z^2, (w^2, b^2, a^1(z^1(w^1, b^1, a^0)))))))$$

By applying the chain rule to this composition function one can find the required gradients for the parameter update equations 1 and 2

The chain rule for three weight and bias parameters:

For W^3, b^3:

$$\frac{\partial C}{\partial W^3} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial W^3}$$

$$\frac{\partial C}{\partial b^3} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial b^3}$$

For W^2, b^2:

$$\frac{\partial C}{\partial W^2} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial a^2}\frac{\partial a^2}{\partial z^2}\frac{\partial z^2}{\partial w^2}$$

$$\frac{\partial C}{\partial b^2} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial a^2}\frac{\partial a^2}{\partial z^2}\frac{\partial z^2}{\partial b^2}$$

For W^1, b^1:

$$\frac{\partial C}{\partial W^1} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial a^2}\frac{\partial a^2}{\partial z^2}\frac{\partial z^2}{\partial a^1}\frac{\partial a^1}{\partial z^1}\frac{\partial z^1}{\partial W^1}$$

$$\frac{\partial C}{\partial b^1} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial a^2}\frac{\partial a^2}{\partial z^2}\frac{\partial z^2}{\partial a^1}\frac{\partial a^1}{\partial z^1}\frac{\partial z^1}{\partial b^1}$$

Four important partial derivatives that would help the calculation are: $\frac{\partial C}{\partial z^L}, \frac{\partial C}{\partial z^l}, \frac{\partial z^l}{\partial W^l}, \frac{\partial z^l}{\partial b^l}$

First lets find $\frac{\partial C}{\partial z^L}$

$$\frac{\partial C}{\partial z^L} = \frac{\partial C}{\partial a^L}\frac{\partial a^L}{\partial z^L} = \frac{\partial C}{\partial a^L}\frac{\partial \sigma(z^L)}{\partial z^L}$$

From the cost function for a single sample one can find

$$\frac{\partial C}{\partial a^L} = (a^L - y)$$

<u>The derivatives of the activation σ w.r.t z^L can be found as follows:</u>

$\sigma(z^L) = \frac{1}{1+e^{-z^L}}$ -> differentiation $\sigma' = \frac{\partial (1+e^{-z^L})^{-1}}{\partial z^L}$

From the reciprocal rule:

$$\sigma' = -\frac{1}{(1+e^{-z^L})^2}\frac{\partial(1+e^{-z^L})}{\partial z^L}$$

From the exponent rule:

$$\sigma' = -\frac{1}{(1+e^{-z^L})^2}e^{-z^L}\frac{\partial(-z^L)}{\partial z^L} = \frac{e^{-z^L}}{(1+e^{-z^L})^2}$$

$$\sigma' = \frac{e^{-z^L}+1-1}{(1+e^{-z^L})^2} = \frac{1}{(1+e^{-z^L})}[\frac{1+e^{-z^L}}{1+e^{-z^L}} - \frac{1}{1+e^{-z^L}}]$$

$$\sigma' = \sigma(z^L)[1-\sigma(z^L)]$$

$$\sigma' = a^L(1-a^L)$$

Finally:

$$\frac{\partial C}{\partial z^L} = (a^L - y)a^L(1-a^L)$$

The partial derivative of C w.r.t z^l

$$\frac{\partial C}{\partial z^l} = \frac{\partial C}{\partial z^{l+1}}\frac{\partial z^{l+1}}{\partial a^l}\frac{\partial a^l}{\partial z^l}$$

$$\frac{\partial z^{l+1}}{\partial a^l} = \frac{\partial[W^{l+1}a^l + b^{l+1}]}{\partial a^l}$$

$$\frac{\partial z^{l+1}}{\partial a^l} = W^{l+1}$$

Finally:

$$\frac{\partial C}{\partial z^l} = [W^{l+1}]^T\frac{\partial C}{\partial z^{l+1}} \circ \sigma'(z^L)$$

The partial derivatives of z^l w.r.t w^l

$$\frac{\partial z^l}{\partial W^l} = \frac{\partial[W^l a^{l-1} + b^l]}{\partial W^l}$$

$$\frac{\partial z^l}{\partial W^l} = a^{l-1}$$

The partial derivatives of z^l w.r.t b^l

$$\frac{\partial z^l}{\partial b^l} = \frac{\partial [W^l a^{l-1} + b^l]}{\partial b^l}$$

$$\frac{\partial z^l}{\partial b^l} = 1$$