ENGR 4350:Applied Deep Learning

Neural Network: Part 1
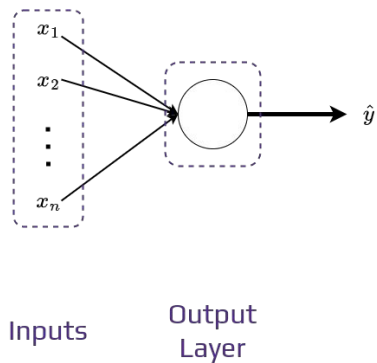
09/14/2022
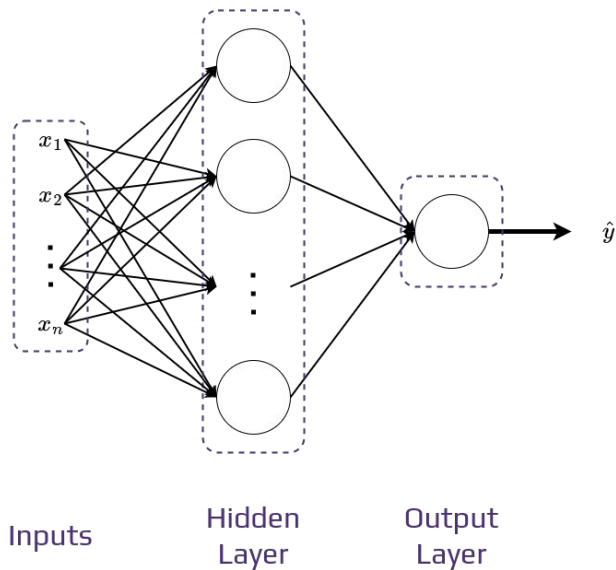
Outline

- Neural Network Representation
- Forward & Backward Propagation
- Activation Functions

Neural Network Representation

x_1
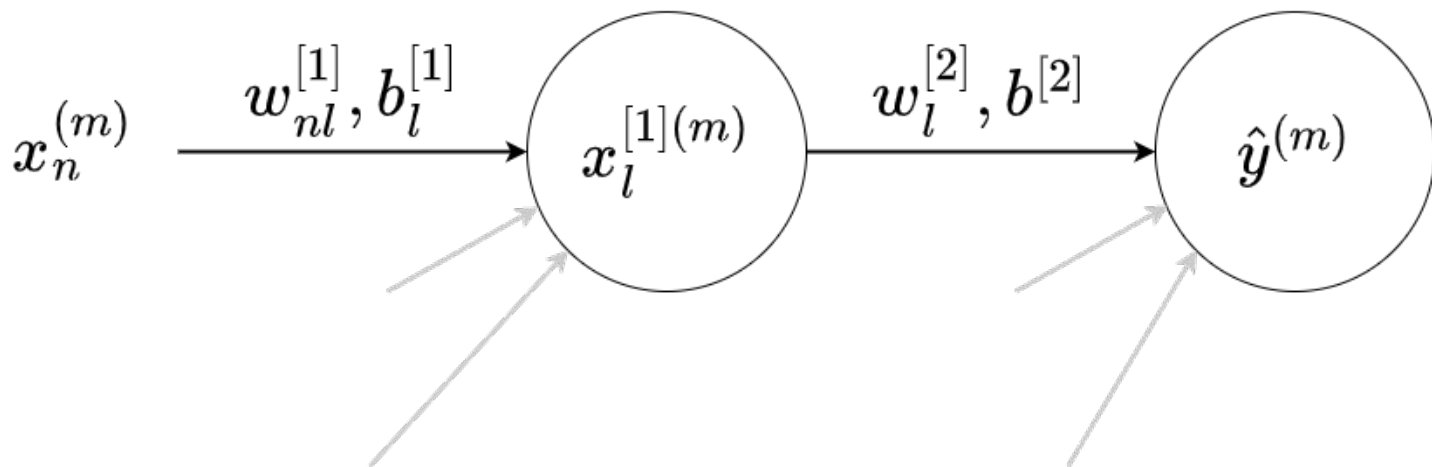x_2
$\vdots$
x_n

$\hat{y}$

Inputs

Output
Layer

1-Layer neural network, w/o hidden layer

(Logistic regression model)

x_1
x_2
$\vdots$
x_n

$\hat{y}$

Inputs

Hidden
Layer

Output
Layer

2-Layer neural network w/ 1 hidden layer

Neural Network Representation

$$x_n^{(m)} \quad \xrightarrow{\quad w_{nl}^{[1]}, b_l^{[1]} \quad} \quad x_l^{[1](m)} \quad \xrightarrow{\quad w_l^{[2]}, b^{[2]} \quad} \quad \hat{y}^{(m)}$$
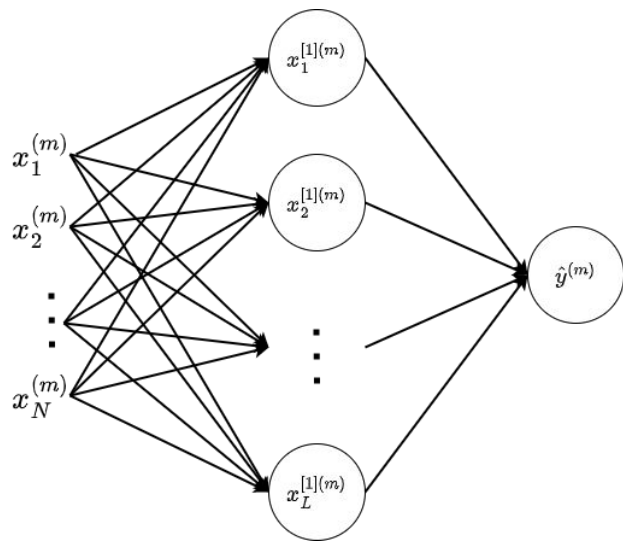
$l \in \{1, \ldots, L\}$: neuron index

$m \in \{1, \ldots, M\}$: example index

$n \in \{1, \ldots, N\}$: feature index

[1]: 1st layer

[2]: 2nd layer

Neural Network Representation

$$x_1^{[1](m)} = \sigma\left(w_{11}^{[1]}x_1^{(m)} + w_{21}^{[1]}x_2^{(m)} + \ldots + w_{N1}^{[1]}x_N^{(m)} + b_1^{[1]}\right)$$
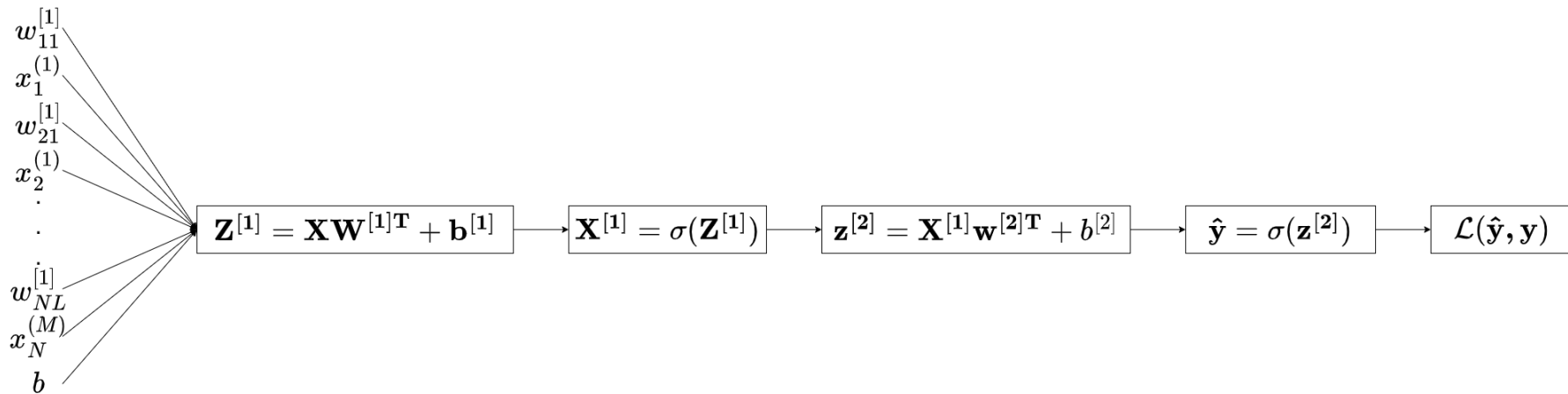
$$x_2^{[1](m)} = \sigma\left(w_{12}^{[1]}x_1^{(m)} + w_{22}^{[1]}x_2^{(m)} + \ldots + w_{N2}^{[1]}x_N^{(m)} + b_2^{[1]}\right)$$

$$\vdots$$

$$x_L^{[1](m)} = \sigma\left(w_{1L}^{[1]}x_1^{(m)} + w_{2L}^{[1]}x_2^{(m)} + \ldots + w_{NL}^{[1]}x_N^{(m)} + b_L^{[1]}\right)$$

$$\hat{y}^{(m)} = x^{[2](m)} = \sigma\left(w_1^{[2]}x_1^{[1](m)} + w_2^{[2]}x_2^{[1](m)} + \ldots + w_L^{[2]}x_L^{[1](m)} + b^{[2]}\right)$$

Neural Network Representation

$$\mathbf{Z}^{[1]} = \mathbf{X}\mathbf{W}^{[1]\mathbf{T}} + \mathbf{b}^{[1]} \longrightarrow \mathbf{X}^{[1]} = \sigma(\mathbf{Z}^{[1]}) \longrightarrow \mathbf{z}^{[2]} = \mathbf{X}^{[1]}\mathbf{w}^{[2]\mathbf{T}} + b^{[2]} \longrightarrow \hat{\mathbf{y}} = \sigma(\mathbf{z}^{[2]}) \longrightarrow \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$$

$\mathbf{X} : M \times N$ $\mathbf{W}^{[1]} : L \times N$ $\mathbf{b}^{[1]} : 1 \times L$ $\mathbf{Z}^{[1]} : M \times L$

$\mathbf{X}^{[1]} : M \times L$ $\mathbf{w}^{[2]} : 1 \times L$ $\mathbf{b}^{[2]} : 1 \times 1$ $\mathbf{z}^{[2]} : M \times 1$ $\hat{\mathbf{y}} : M \times 1$

$\mathbf{y} : M \times 1$

Forward Propagation

$$\mathbf{X}^{[1]} = \sigma\left(\mathbf{X}\mathbf{W}^{[1]} + \mathbf{b}^{[1]}\right)$$

$$\sigma\left(\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdot & \cdot & \cdot & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdot & \cdot & \cdot & x_N^{(2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_1^{(M)} & x_2^{(M)} & \cdot & \cdot & \cdot & x_N^{(M)} \end{bmatrix} \cdot \begin{bmatrix} w_{11}^{[1]} & w_{12}^{[1]} & \cdot & \cdot & \cdot & w_{1L}^{[1]} \\ w_{21}^{[1]} & w_{22}^{[1]} & \cdot & \cdot & \cdot & w_{2L}^{[1]} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ w_{N1}^{[1]} & w_{N2}^{[1]} & \cdot & \cdot & \cdot & w_{NL}^{[1]} \end{bmatrix} + \begin{bmatrix} b_1^{[1]} & b_2^{[1]} & \cdot & \cdot & \cdot & b_L^{[1]} \\ b_1^{[1]} & b_2^{[1]} & \cdot & \cdot & \cdot & b_L^{[1]} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ b_1^{[1]} & b_2^{[1]} & \cdot & \cdot & \cdot & b_L^{[1]} \end{bmatrix}\right) = \begin{bmatrix} x_1^{[1](1)} & x_2^{[1](1)} & \cdot & \cdot & \cdot & x_L^{[1](1)} \\ x_1^{[1](2)} & x_2^{[1](2)} & \cdot & \cdot & \cdot & x_L^{[1](2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_1^{[1](M)} & x_2^{[1](M)} & \cdot & \cdot & \cdot & x_L^{[1](M)} \end{bmatrix}$$

$$\hat{\mathbf{y}} = \mathbf{x}^{[2]} = \sigma\left(\mathbf{X}^{[1]}\mathbf{w}^{[2]\mathbf{T}} + b^{[2]}\right)$$

$$\sigma\left(\begin{bmatrix} x_1^{[1](1)} & x_2^{[1](1)} & \cdot & \cdot & \cdot & x_L^{[1](1)} \\ x_1^{[1](2)} & x_2^{[1](2)} & \cdot & \cdot & \cdot & x_L^{[1](2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_1^{[1](M)} & x_2^{[1](M)} & \cdot & \cdot & \cdot & x_L^{[1](M)} \end{bmatrix} \cdot \begin{bmatrix} w_1^{[2]} \\ w_2^{[2]} \\ \cdot \\ \cdot \\ \cdot \\ w_L^{[2]} \end{bmatrix} + \begin{bmatrix} b^{[2]} \\ b^{[2]} \\ \cdot \\ \cdot \\ \cdot \\ b^{[2]} \end{bmatrix}\right) = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}^{(M)} \end{bmatrix}$$

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = -(\mathbf{y}\log\hat{\mathbf{y}} + (1-\mathbf{y})\log(1-\hat{\mathbf{y}}))$$

$$J\left(\mathbf{W}^{[1]}, \mathbf{b}^{[1]}, \mathbf{w}^{[2]}, b^{[2]}\right) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$$

Backward Propagation

$$dz^{[2]} = \frac{\partial J}{\partial z^{[2]}} = \hat{y} - y$$

$$dw^{[2]} = dz^{[2]} \cdot \frac{\partial z^{[2]}}{\partial w^{[2]}} = \frac{1}{M} \left(\hat{y} - y \right)^T \cdot X^{[1]}$$

$$db^{[2]} = dz^{[2]} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{y} - y \right)^T$$

$$dX^{[1]} = dz^{[2]} \cdot \frac{\partial z^{[2]}}{\partial X^{[1]}} = \left(\hat{y} - y \right)^T \cdot w^{[2]}$$

$$dZ^{[1]} = dX^{[1]} \cdot \frac{\partial X^{[1]}}{\partial Z^{[1]}} = \left[\left(\hat{y} - y \right) \cdot w^{[2]} \right] * X^{[1]} * \left(1 - X^{[1]} \right)$$

$$dW^{[1]} = dZ^{[1]} \cdot \frac{\partial Z^{[1]}}{\partial W^{[1]}} = \frac{1}{M} \left[\left(\hat{y} - y \right) \cdot w^{[2]} * X^{[1]} * \left(1 - X^{[1]} \right) \right]^T \cdot X$$

$$db^{[1]} = dZ^{[1]} \cdot \frac{\partial Z^{[1]}}{\partial b^{[1]}} = \frac{1}{M} \sum_{m=1}^{M} \left[\left(\hat{y} - y \right) \cdot w^{[2]} * X^{[1]} * \left(1 - X^{[1]} \right) \right]$$
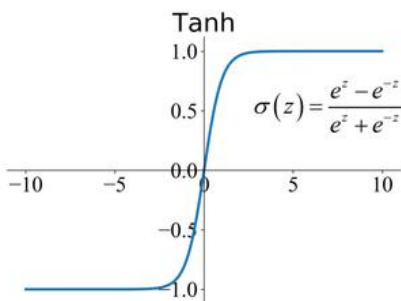
$*$: element-wise multiplication

Activation Functions

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

$$\sigma'(z) = 1 - \sigma^2(z)$$

Sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

(a)

Tanh

$$\sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

(b)

ReLU

$$\text{ReLU}(z) = \begin{cases} z, & z > 0 \\ 0, & otherwise \end{cases}$$

(c)

LeakyReLU(a=0.2)

$$\text{LeakyReLU}(z) = \begin{cases} z, & z > 0 \\ az, & otherwise \end{cases}$$
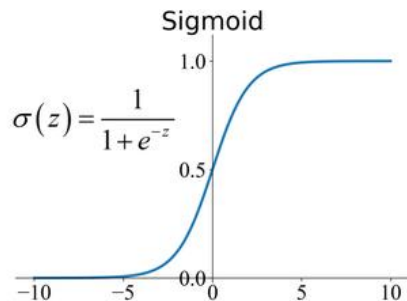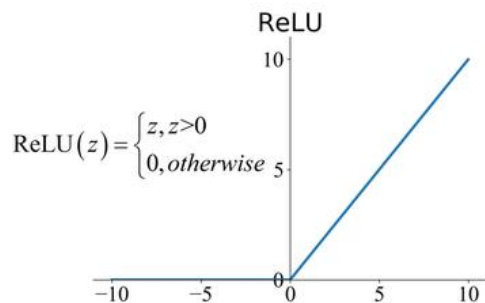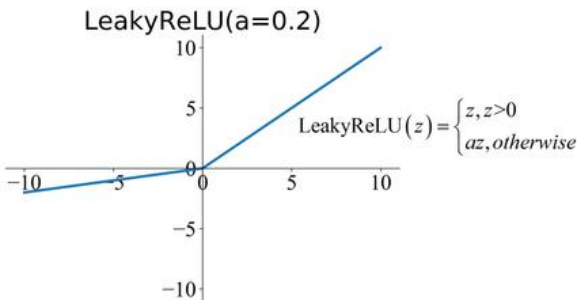
(d)

$$\text{ReLU}'(z) = \begin{cases} 1, & z > 0 \\ 0, & otherwise \end{cases}$$

$$\text{LeakyReLU}'(z) = \begin{cases} 1, & z > 0 \\ a, & otherwise \end{cases}$$

Backward Propagation
w/ any activation function

$$dz^{[2]} = \frac{\partial J}{\partial z^{[2]}} = \hat{y} - y$$

$$dw^{[2]} = dz^{[2]} \cdot \frac{\partial z^{[2]}}{\partial w^{[2]}} = \frac{1}{M}\left(\hat{y} - y\right)^T \cdot X^{[1]}$$

$$db^{[2]} = dz^{[2]} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} = \frac{1}{M}\sum_{m=1}^{M}(\hat{y} - y)$$

$$dX^{[1]} = dz^{[2]} \cdot \frac{\partial z^{[2]}}{\partial X^{[1]}} = \left(\hat{y} - y\right) \cdot w^{[2]}$$

$$dZ^{[1]} = dX^{[1]} \cdot \frac{\partial X^{[1]}}{\partial Z^{[1]}} = \left[(\hat{y} - y) \cdot w^{[2]}\right] * g'\left(Z^{[1]}\right)$$

$$dW^{[1]} = dZ^{[1]} \cdot \frac{\partial Z^{[1]}}{\partial W^{[1]}} = \frac{1}{M}\left[(\hat{y} - y) \cdot w^{[2]} * g'\left(Z^{[1]}\right)\right]^T \cdot X$$

$$db^{[1]} = dZ^{[1]} \cdot \frac{\partial Z^{[1]}}{\partial b^{[1]}} = \frac{1}{M}\sum_{m=1}^{M}\left[(\hat{y} - y) \cdot w^{[2]} * g'\left(Z^{[1]}\right)\right]$$