⊛ Building a neural network from scratch:

Problem statement: Digit Classification.

Dataset: MNIST dataset.

28X28 gray scale image.

784

28x28 m training images

$$X = \begin{bmatrix} -x^{(1)}- \\ -x^{(2)}- \\ \vdots \\ -x^{(m)}- \end{bmatrix}^T = \begin{bmatrix} | & | & & | \\ X^{(1)} & X^{(2)} & \cdots & X^{(m)} \\ | & | & & | \end{bmatrix}$$

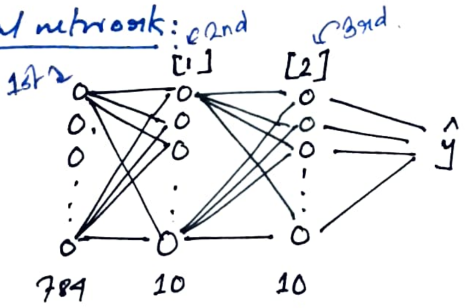
784xm

⇓

$\underbrace{0, 1, 2, \ldots 9}$

10 classes. (classification).

neural network:



input layer. hidden layer. output layer.

forward propagation:

$A^{[0]} = X\,(784 \times m) \longrightarrow$ Input layer.

$Z^{[1]} = W^{[1]} A^{[0]} + b^{[1]} \longrightarrow$ bias. $\longrightarrow$ Unactivated 1st layer.

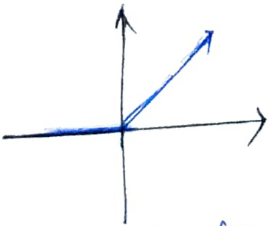
$10 \times m$ $\quad$ 10×784 $\quad$ $784 \times m$ $\quad$ $10 \times 1 \Rightarrow 10 \times m$.

weight

Apply activation function,

$A^{[1]} = g(Z^{[1]}) = ReLU(Z^{[1]})$

Activated 1st layer.



$$ReLU = \begin{cases} x & \text{for } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

Unactivated 2nd layer,

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

$Z^{[2]}$: $10 \times m$, $W^{[2]}$: 10×10, $A^{[1]}$: $10 \times m$, $b^{[2]}$: $10 \times 1 \Rightarrow 10 \times m$.

$$A^{[2]} = \text{softmax}(Z^{[2]})$$ → Activation function.

Eg:

Output layer. Softmax activation function. Probabilities.

$$\begin{bmatrix} 1.3 \\ 5.1 \\ 2.2 \\ 0.7 \\ 1.1 \end{bmatrix} \longrightarrow \boxed{\dfrac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}} \longrightarrow \begin{bmatrix} 0.02 \\ 0.90 \\ 0.05 \\ 0.01 \\ 0.02 \end{bmatrix}$$

Adds upto 1.

Backward propagation:

$$dZ^{[2]} = A^{[2]} - Y$$ → Error calculation.

$dZ^{[2]}$: $10 \times m$, $A^{[2]}$: $10 \times m$, Y: $10 \times m$.

How much the output value is deviated from actual value.

2nd layer

$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]T}$$

$dW^{[2]}$: 10×10, $dZ^{[2]}$: $10 \times m$, $A^{[1]T}$: $m \times 10$

$$db^{[2]} = \frac{1}{m} \sum dZ^{[2]}$$

$db^{[2]}$: 10×1, $\sum dZ^{[2]}$: 10×1

1st layer

$$dZ^{[1]} = W^{[2]T} dZ^{[2]} .* g'(Z^{[1]})$$

$dZ^{[1]}$: $10 \times m$, $W^{[2]T}$: 10×10, $dZ^{[2]}$: $10 \times m$, $g'(Z^{[1]})$: $10 \times m$.

$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^{T}$$

$dW^{[1]}$: 10×784, $dZ^{[1]}$: $10 \times m$, X^T: $m \times 784$.

$$db^{[1]} = \frac{1}{m} \sum dZ^{[1]}$$

$db^{[1]}$: 10×1, $\sum dZ^{[1]}$: 10×1

Now update parameters,

$$W^{[1]} := W^{[1]} - \alpha dW^{[1]}$$

$\alpha \rightarrow$ learning rate

$$b^{[1]} := b^{[1]} - \alpha db^{[1]}$$

$$W^{[2]} := W^{[2]} - \alpha dW^{[2]}$$

$$b^{[2]} := b^{[2]} - \alpha db^{[2]}$$