IT DOES NOT WORK!

Loss: $L = \frac{1}{2}(\hat{y}-y)^2 = \frac{1}{2}(0.751365-0.07)^2 \approx 0.274811$

2) Backward pass — compute gradients

Compute the error signals (deltas) and then gradients for weights and biases

- Output Layer delta

→ MSE : $L = \frac{1}{2}(\hat{y}-y)^2 \Rightarrow \frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}}\left(\frac{1}{2}(\hat{y}-y)^2\right)$ ← derivative

$\frac{\partial L}{\partial a^{(2)}} = \frac{\partial L}{\partial \hat{y}} = (\hat{y}-y) \cdot \sigma'$

derivative of sigmoid: $\sigma'(z) = \sigma(z)(1-\sigma(z))$

$\sigma(x) = \frac{1}{1+e^{-x}} \xrightarrow{\text{rewrite}} \sigma(x) = (1+e^{-x})^{-1}$

apply the chain rule to the outer function

~~$g(x) = (w^{(L+1)})^T \cdot g^{(L+1)} \odot \sigma'(z^{(L)})$~~

$= u^{-1}$ with $u = 1+e^{-x}$

$\frac{d}{dx}\left[(1+e^{-x})^{-1}\right] = -(1+e^{-x})^{-2} \cdot \frac{d}{dx}(1+e^{-x})$

$(x^n)' = n x^{n-1}$

$\frac{d}{dx}(1+e^{-x}) \rightarrow -e^{-x}$

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

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$

→ Solve for $1-\sigma(x)$

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

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

$$e^{-x}/1+e^{-x} = 1-\sigma(x)$$

now → multiply $\sigma(x)$ and $[1-\sigma(x)]$

$$\sigma(x)(1-\sigma(x)) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\sigma'(x) = \sigma(x)(1-\sigma(x)) = a(1-a)$$

if a function depend on another function, we use the chain rule. $\frac{dL}{dz} = \frac{dL}{da} \cdot \frac{da}{dz}$

Delta → $\delta = (\hat{y}-y) \cdot a(1-a)$

$a^{(2)}(1-a^{(2)}) = 0.751365(1-0.751365) \approx 0.186816$

$\delta^{(2)} = 0.7413565 \cdot 0.186816 \approx 0.153499$

Two-Layer Network

(1)

Input $x = \begin{bmatrix} 0.05 \\ 0.10 \end{bmatrix}$ $y = 0.01$ Learning Rate $= 0.5$

Hidden layer size: 2 neurons

Output layer: 1 neuron

First-Layer weight & Bias

$W^{(1)} = \begin{bmatrix} 0.15 & 0.20 \\ 0.25 & 0.30 \end{bmatrix}$ $b^{(1)} = \begin{bmatrix} 0.35 \\ 0.35 \end{bmatrix}$

Second-Layer weights & Bias

$W^{(2)} = [0.40 \ 0.45]$ $b^{(2)} = [0.60]$

Pre-

Hidden-Activation (2)

$z^{(1)} = W^{(1)} x + b^{(1)}$ (weighted sum, Bias)

neuron 1 → $= (0.15)(0.05) + (0.20)(0.10) + 0.35 = 0.3775$

neuron 2 → $= (0.25)(0.05) + (0.30)(0.10) + 0.35 = 0.3925$

$z^{(1)} = \begin{bmatrix} 0.3775 \\ 0.3925 \end{bmatrix} - z_1^{(1)} \\ - z_2^{(1)}$

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

Neuron 1 $\sigma(z_1^{(1)}) = 0.593269$

Neuron 2 $\sigma(z_2^{(1)}) = 0.596884$

$\sigma(z^{(1)}) = \begin{bmatrix} 0.593269 \\ 0.596884 \end{bmatrix} = a_2^{(1)}$

$W^{(2)} = [0.40 \ 0.45]$ $b^{(2)} = [0.60]$

Output pre-activation : $z^{(2)} = w^{(2)} a^{(1)} + b^{(2)}$

$(0.40)(0.593269) + (0.45)(0.596884) + 0.60$

$z^{(2)} = [1.1059054] +$.

Output - Activation $= \sigma(z^{(2)}) = \frac{1}{1+e^{-z^{(2)}}}$

$\hat{y} = \sigma(z^{(2)}) = 0.751365$ ← $a^{(2)}$

Loss Computation → $L = \frac{1}{2}(\hat{y} - y)^2$

$L = 0.274811$

(2) Backpropagation Deltas and Gradients

$\sigma'(z) = \sigma(z)(1 - \sigma(z))$ ← Sigmoid derivat

derivative of Loss → $\partial L / \partial a^{(2)} = \hat{y} - y$

$0.751365 - 0.01 = 0.741365$

Activation derivative wrt pre-activation

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

$0.751365(1 - 0.751365) =$

$a^{(2)}(1 - a^{(2)}) \approx 0.186816$

Chain rule multiplication: $\delta^{(2)} = \partial L / \partial a \cdot \partial a^{(2)} / \partial z^{(2)}$

$\delta^{(2)} = \partial L / \partial a^{(2)} \cdot \sigma'(z^{(2)}) = (\hat{y} - y) \cdot a^{(2)}(1 - a^{(2)})$

$\delta^{(2)} = 0.741365(0.186816) = 0.1384499$

Gradients for $W^{(2)}$ and $b^{(2)}$

$\frac{\partial L}{\partial W^{(2)}} = g^{(2)} \cdot (a^{(1)})^T$ (shape: 1×2)

$\frac{\partial L}{\partial W_1^{(2)}} = 0.138499(0.593269) = 0.082167$ (a_1^1)

$\partial L / \partial W_2^{(2)} = 0.138499(0.596884) = 0.082668$ (a_2^1)

Bias gradient

$\frac{\partial L}{\partial b^{(2)}} = g^{(2)} \approx 0.138499$

Hidden layers

$g^{(1)} = (W^{(2)})^T g^{(2)} \odot \sigma'(z^{(1)})$ $(W^{(2)})^T g^{(2)}$

For neuron 1: $0.4 \cdot 0.138499 = 0.055400$

For neuron 2: $0.45 \cdot 0.138499 = 0.062325$

$\sigma'(z^{(1)}) = a^{(1)}(1 - a^{(1)})$

$= 0.593269(1 - 593269) = 0.241300$

$= 0.596884(1 - 0.596884) = 0.240260$

$g_1^{(1)} = (0.055400)(0.241300) = 0.013368$

$g_2^{(1)} = (0.062325)(0.240260) = 0.014974$

Gradients for $W^{(1)}$ and $b^{(1)}$

$\partial L / \partial W^{(1)} = g_1^{(1)} (x)^T$ (shape 2×2)

neuron 1

$= 0.013368(0.05) = 0.0006684$

$0.013368(0.10) = 0.001336$

$g_2^{(1)}(x)$ → input

Neuron 2

$= 0.014974(0.05) = 0.0006684 \approx 0.0007487$

$= 0.014974(0.10) = 0.0013368 \approx 0.0014974$

- Bias gradient $\partial L / \partial b^{(1)} = g^{(1)}$

$$\partial L / \partial b^{(1)} = g^{(1)} = \begin{bmatrix} 0.013368 \\ 0.014974 \end{bmatrix}$$

3) Parameter Update (One gradient Step)

- Gradient descent update: $\theta \leftarrow \theta - \eta \cdot \nabla \theta$
- with $\eta = 0.5$
- Update $W^{(2)}$ and $b^{(2)}$
- $W^{(2)}_{new,1} = 0.40 - 0.5 \cdot 0.082167 \approx 0.358917$
- $W^{(2)}_{new,2} = 0.45 - 0.5 \cdot 0.082668 \approx 0.408666$
- $b^{(2)}_{new} = 0.60 - 0.5 \cdot 0.138499 \approx 0.530751$
- So $W^{(2)}_{new} \approx [0.358917, 0.408668]$
- $b^{(2)}_{new} \approx 0.530751$

- Update $W^{(1)}$ and $b^{(1)}$

$W^{(1)}_{new,11} = 0.15 - 0.5 * 0.0006689 \approx 0.14966$

$W^{(1)}_{new,12} = 0.20 - 0.5 * 0.0013368 \approx 0.199332$

$W^{(1)}_{new,21} = 0.25 - 0.5 * 0.0007487 \approx 0.249626$

$W^{(1)}_{new,22} = 0.30 - 0.5 * 0.0014974 \approx 0.299251$

$b^{(1)}_{new,1} = 0.35 - 0.5 * 0.013368 \approx 0.343316$

$b^{(1)}_{new,2} = 0.35 - 0.5 * 0.014974 \approx 0.342513$

$$W_{new} = \begin{bmatrix} 0.14966 & 0.199332 \\ 0.249626 & 0.299251 \end{bmatrix} \quad b_{new} = \begin{bmatrix} 0.343316 \\ 0.342513 \end{bmatrix}$$

1 Forward pass to confirm the loss decreased

$New z_1^{(1)} = 0.14966 \cdot 0.05 + 0.199332 \cdot 0.10 + 0.345316 \approx 0.370733$

$New z_2^{(1)} \approx 1.249826 \cdot 0.05 + 0.299251 \cdot 0.10 + 0.342513 \approx 0.384919$

$New a_1^{(1)} = \sigma(0.370733) \approx 0.591862$

$New a_2^{(1)} = \sigma(0.384919) \approx 0.595952$

$New z^{(2)} = 0.358917 + 0.591862 + 0.40866 \cdot 0.595952 + 0.536751$

≈ 0.98615

$New \hat{y} \approx \sigma(0.98615) \approx 0.727083$

new loss $\approx$

$L_{new} = 0.5(0.72703 - 0.01)^2 \approx 0.2570$

So the ~~gradient~~ Loss went down from ≈ 0.274811 to 0.2570