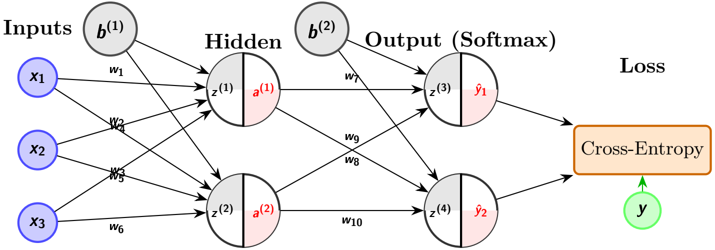
Understanding Neural Networks: Softmax, Cross-Entropy, and Backpropagation

Forward Pass, Loss Computation, and the Chain Rule

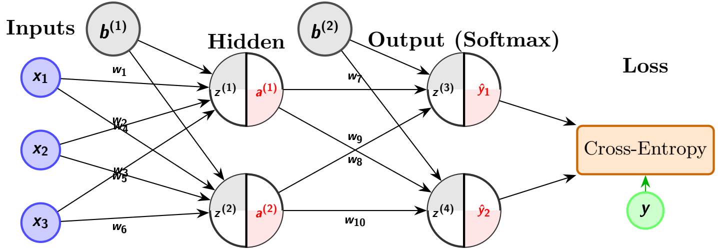
Bindeshwar Singh Kushwaha
PostNetwork Academy

Neural Network with Softmax + Cross-Entropy



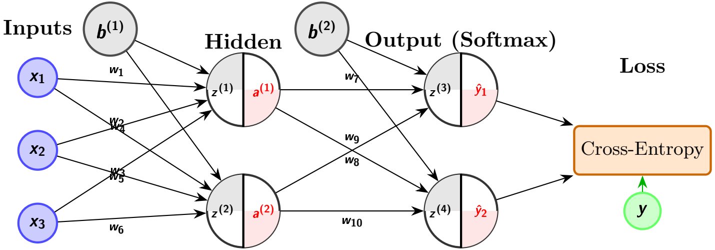
- **Input Layer**: The network receives 3 input features, denoted x_1, x_2, x_3.

Neural Network with Softmax + Cross-Entropy



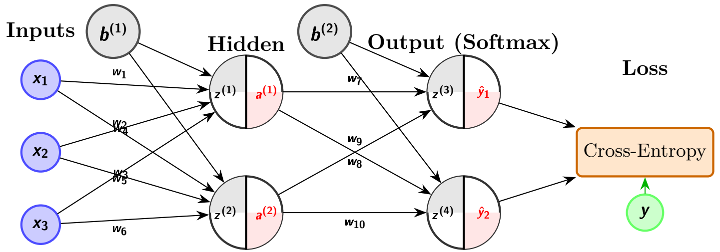
- **Input Layer**: The network receives 3 input features, denoted $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$.
- **Hidden Layer**: 2 neurons in the hidden layer with activations $\boldsymbol{a^{(1)}}$ and $\boldsymbol{a^{(2)}}$. Each neuron computes a weighted sum of inputs and applies an activation function.

Neural Network with Softmax + Cross-Entropy



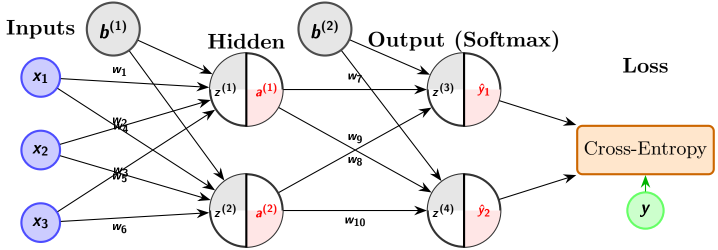
- **Input Layer**: The network receives 3 input features, denoted $\boldsymbol{x_1, x_2, x_3}$.
- **Hidden Layer**: 2 neurons in the hidden layer with activations $\boldsymbol{a^{(1)}}$ and $\boldsymbol{a^{(2)}}$. Each neuron computes a weighted sum of inputs and applies an activation function.
- **Output Layer**: 2 output neurons $\boldsymbol{z^{(3)}, z^{(4)}}$, which are then passed through softmax to produce predictions.

Neural Network with Softmax + Cross-Entropy



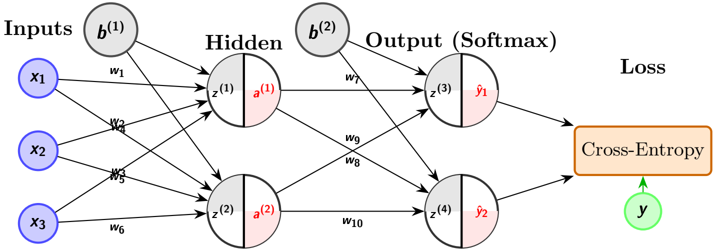
- **Input Layer**: The network receives 3 input features, denoted $\boldsymbol{x_1, x_2, x_3}$.
- **Hidden Layer**: 2 neurons in the hidden layer with activations $\boldsymbol{a^{(1)}}$ and $\boldsymbol{a^{(2)}}$. Each neuron computes a weighted sum of inputs and applies an activation function.
- **Output Layer**: 2 output neurons $\boldsymbol{z^{(3)}, z^{(4)}}$, which are then passed through softmax to produce predictions.
- **Softmax Activation**: Applied to the output layer to obtain probability predictions $\boldsymbol{\hat{y}_1, \hat{y}_2}$.

Neural Network with Softmax + Cross-Entropy



- **Input Layer**: The network receives 3 input features, denoted x_1, x_2, x_3.
- **Hidden Layer**: 2 neurons in the hidden layer with activations $a^{(1)}$ and $a^{(2)}$. Each neuron computes a weighted sum of inputs and applies an activation function.
- **Output Layer**: 2 output neurons $z^{(3)}, z^{(4)}$, which are then passed through softmax to produce predictions.
- **Softmax Activation**: Applied to the output layer to obtain probability predictions $\hat{y}_1, \hat{y}_2$.
- **Loss Function**: Cross-Entropy Loss is used to measure the difference between predicted outputs $\hat{y}$ and target labels y.

Step 1: Hidden Pre-Activation



$$x_1 = 1,\ x_2 = 2,\ x_3 = -1$$

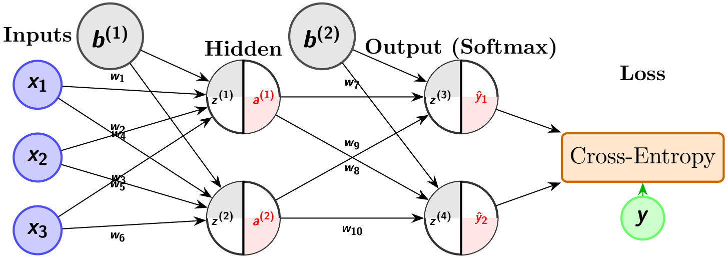
$$w_1 = 0.2,\ w_2 = -0.3,\ w_3 = 0.4,\ b^{(1)} = 0.5$$

$$w_4 = -0.5,\ w_5 = 0.1,\ w_6 = 0.2$$

$$z^{(1)} = w_1x_1 + w_2x_2 + w_3x_3 + b^{(1)} = 0.2 * 1 - 0.3 * 2 + 0.4 * (-1) + 0.5 = -0.3$$

$$z^{(2)} = w_4x_1 + w_5x_2 + w_6x_3 + b^{(1)} = -0.5 * 1 + 0.1 * 2 + 0.2 * (-1) + 0.5 = 0$$

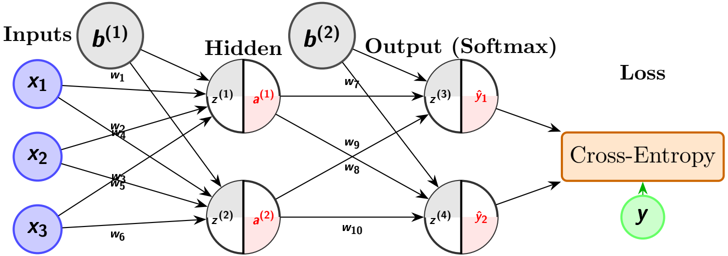
Step 2: Hidden Activation (Sigmoid)



$$a^{(1)} = \sigma(z^{(1)}) = \frac{1}{1 + e^{0.3}} \approx 0.426$$

$$a^{(2)} = \sigma(z^{(2)}) = \frac{1}{1 + e^{0}} = 0.5$$

Step 3: Output Pre-Activation

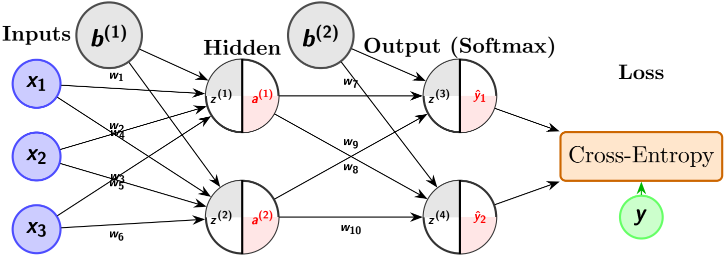


$$w_7 = 0.3, w_8 = -0.1, w_9 = 0.4, w_{10} = 0.2, b^{(2)} = 0.1$$

$$z^{(3)} = w_7 a_1 + w_9 a_2 + b^{(2)} = 0.3 * 0.426 + 0.4 * 0.5 + 0.1 \approx 0.428$$

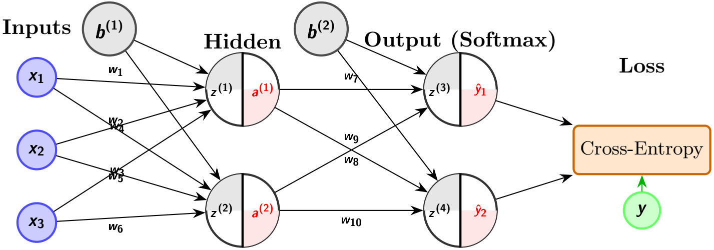
$$z^{(4)} = w_8 a_1 + w_{10} a_2 + b^{(2)} = -0.1 * 0.426 + 0.2 * 0.5 + 0.1 \approx 0.185$$

Step 4: Output Activation (Softmax)



$$\hat{y}_1 = \frac{e^{z^{(3)}}}{e^{z^{(3)}} + e^{z^{(4)}}} = \frac{e^{0.428}}{e^{0.428} + e^{0.185}} \approx 0.561$$

$$\hat{y}_2 = \frac{e^{z^{(4)}}}{e^{z^{(3)}} + e^{z^{(4)}}} = \frac{e^{0.185}}{e^{0.428} + e^{0.185}} \approx 0.439$$



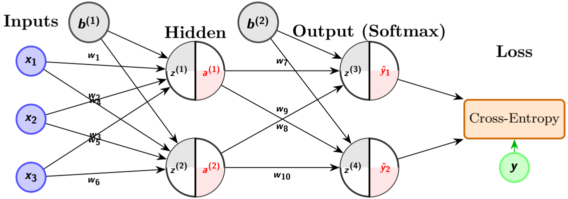
$$L = -\sum_{i=1}^{2} y_i \ln(\hat{y}_i)$$

For target vector $y = [1, 0]$:

$$L = -\Big(y_1 \ln(\hat{y}_1) + y_2 \ln(\hat{y}_2)\Big)$$

$$L = -(1 \cdot \ln(0.561) + 0 \cdot \ln(0.439)) = -\ln(0.561) \approx 0.579$$

Step 6: Chain Rule for w_7 (Expansion)



Gradient of the loss with respect to weight w_7:

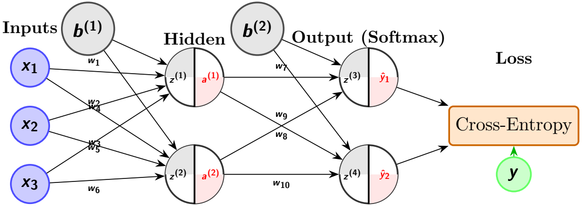
$$\frac{\partial L}{\partial w_7} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial w_7}$$

- $\frac{\partial L}{\partial \hat{y}_1} = -\frac{y_1}{\hat{y}_1}$
- $\frac{\partial \hat{y}_1}{\partial z^{(3)}} = \hat{y}_1(1 - \hat{y}_1)$
- $\frac{\partial z^{(3)}}{\partial w_7} = a^{(1)}$

Substituting:

$$\frac{\partial L}{\partial w_7} = \left(-\frac{y_1}{\hat{y}_1}\right) \hat{y}_1(1 - \hat{y}_1)\, a^{(1)}$$

Step 7: Simplified Gradient for w_7



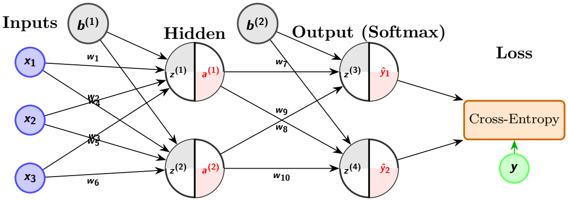
Using the softmax + cross-entropy simplification:

$$\frac{\partial L}{\partial z^{(3)}} = \hat{y}_1 - y_1$$

Hence,

$$\frac{\partial L}{\partial w_7} = (\hat{y}_1 - y_1)\, a^{(1)}$$

Step 8: Chain Rule for w_8 (Expansion)



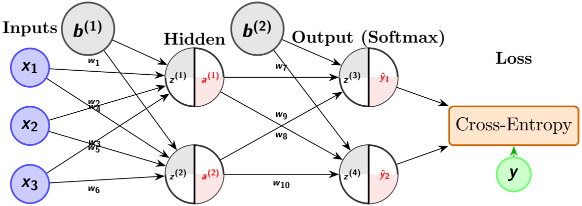
$$\frac{\partial L}{\partial w_8} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial w_8}$$

- $\frac{\partial z^{(3)}}{\partial w_8} = a^{(2)}$

Substituting:

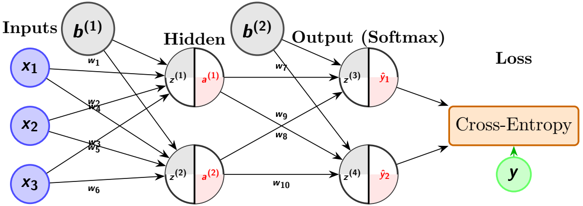
$$\frac{\partial L}{\partial w_8} = \left(-\frac{y_1}{\hat{y}_1}\right) \hat{y}_1(1 - \hat{y}_1)\, a^{(2)}$$

Step 9: Simplified Gradient for w_8



$$\frac{\partial L}{\partial w_8} = (\hat{y}_1 - y_1)\, a^{(2)}$$

Step 10: Chain Rule for w_9 (Expansion)



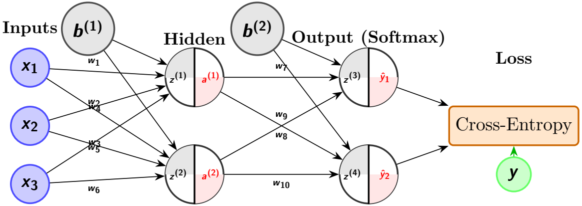
$$\frac{\partial L}{\partial w_9} = \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial z^{(4)}} \cdot \frac{\partial z^{(4)}}{\partial w_9}$$

- $\frac{\partial L}{\partial \hat{y}_2} = -\frac{y_2}{\hat{y}_2}$
- $\frac{\partial \hat{y}_2}{\partial z^{(4)}} = \hat{y}_2(1 - \hat{y}_2)$
- $\frac{\partial z^{(4)}}{\partial w_9} = a^{(1)}$

Substituting:

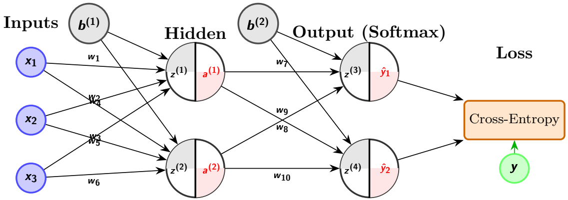
$$\frac{\partial L}{\partial w_9} = \left(-\frac{y_2}{\hat{y}_2}\right) \hat{y}_2(1 - \hat{y}_2)\, a^{(1)}$$

Step 11: Simplified Gradient for w_9



$$\frac{\partial L}{\partial w_9} = (\hat{y}_2 - y_2)\, a^{(1)}$$

Step 12: Chain Rule for w_{10} (Expansion)



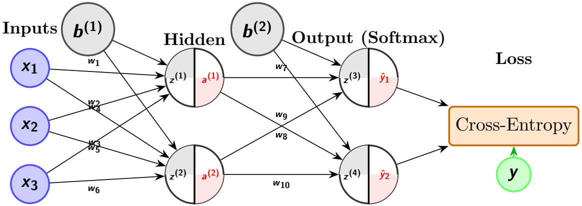
$$\frac{\partial L}{\partial w_{10}} = \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial z^{(4)}} \cdot \frac{\partial z^{(4)}}{\partial w_{10}}$$

- $\frac{\partial z^{(4)}}{\partial w_{10}} = a^{(2)}$

Substituting:

$$\frac{\partial L}{\partial w_{10}} = \left(-\frac{y_2}{\hat{y}_2}\right) \hat{y}_2(1-\hat{y}_2)\, a^{(2)}$$

Step 13: Simplified Gradient for w_{10}



$$\frac{\partial L}{\partial w_{10}} = (\hat{y}_2 - y_2)\, a^{(2)}$$

Update Rules for Output Weights

Using gradient descent with learning rate $\eta > 0$:

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

From previous derivations:

$$\frac{\partial L}{\partial w_7} = (\hat{y}_1 - y_1)\, a^{(1)}, \qquad \frac{\partial L}{\partial w_8} = (\hat{y}_1 - y_1)\, a^{(2)}$$
$$\frac{\partial L}{\partial w_9} = (\hat{y}_2 - y_2)\, a^{(1)}, \qquad \frac{\partial L}{\partial w_{10}} = (\hat{y}_2 - y_2)\, a^{(2)}$$

Therefore the update rules are:

$$\boxed{w_7 \leftarrow w_7 - \eta\,(\hat{y}_1 - y_1)\, a^{(1)}}$$

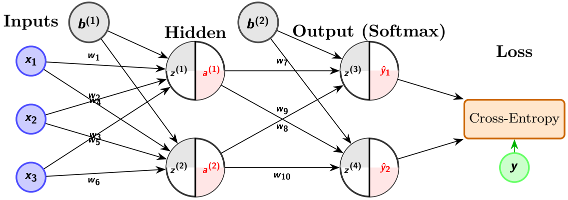
$$\boxed{w_8 \leftarrow w_8 - \eta\,(\hat{y}_1 - y_1)\, a^{(2)}}$$

$$\boxed{w_9 \leftarrow w_9 - \eta\,(\hat{y}_2 - y_2)\, a^{(1)}}$$

$$\boxed{w_{10} \leftarrow w_{10} - \eta\,(\hat{y}_2 - y_2)\, a^{(2)}}$$

Output-bias updates (often included):

$$b_1^{(out)} \leftarrow b_1^{(out)} - \eta\,(\hat{y}_1 - y_1), \qquad b_2^{(out)} \leftarrow b_2^{(out)} - \eta\,(\hat{y}_2 - y_2)$$



Gradient of the loss with respect to weight w_1:

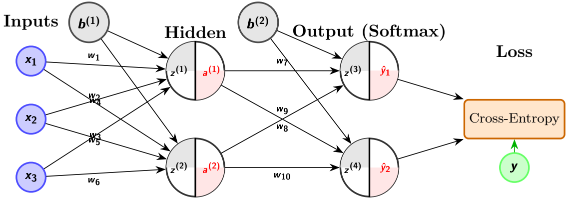
$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a^{(1)}} \cdot \frac{\partial a^{(1)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial w_1}$$

- $\frac{\partial a^{(1)}}{\partial z^{(1)}} = a^{(1)}(1 - a^{(1)})$
- $\frac{\partial z^{(1)}}{\partial w_1} = x_1$
- $\frac{\partial L}{\partial a^{(1)}} = (\hat{y}_1 - y_1)w_7 + (\hat{y}_2 - y_2)w_9$

Substituting:

$$\frac{\partial L}{\partial w_1} = \left[(\hat{y}_1 - y_1)w_7 + (\hat{y}_2 - y_2)w_9\right] a^{(1)}(1 - a^{(1)})\, x_1$$

Step 15: Chain Rule for w_2 (Expansion)



Gradient of the loss with respect to weight w_2:

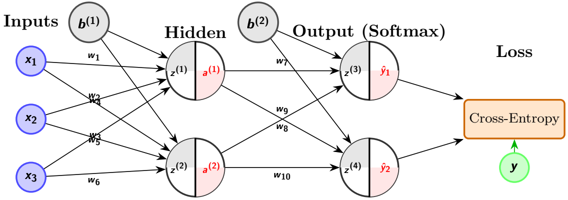
$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial a^{(1)}} \cdot \frac{\partial a^{(1)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial w_2}$$

- $\dfrac{\partial a^{(1)}}{\partial z^{(1)}} = a^{(1)}(1 - a^{(1)})$
- $\dfrac{\partial z^{(1)}}{\partial w_2} = x_2$
- $\dfrac{\partial L}{\partial a^{(1)}} = (\hat{y}_1 - y_1)w_7 + (\hat{y}_2 - y_2)w_9$

Substituting:

$$\frac{\partial L}{\partial w_2} = \left[(\hat{y}_1 - y_1)w_7 + (\hat{y}_2 - y_2)w_9\right] a^{(1)}(1 - a^{(1)})\, x_2$$

Step 16: Chain Rule for w_3 (Expansion)



Gradient of the loss with respect to weight w_3:

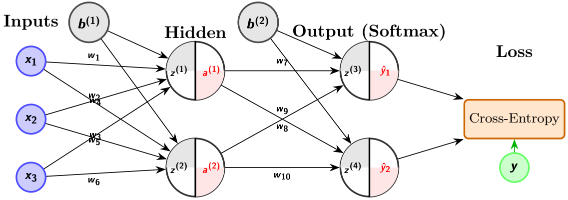
$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial a^{(1)}} \cdot \frac{\partial a^{(1)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial w_3}$$

- $\frac{\partial a^{(1)}}{\partial z^{(1)}} = a^{(1)}(1 - a^{(1)})$
- $\frac{\partial z^{(1)}}{\partial w_3} = x_3$
- $\frac{\partial L}{\partial a^{(1)}} = (\hat{y}_1 - y_1)w_7 + (\hat{y}_2 - y_2)w_9$

Substituting:

$$\frac{\partial L}{\partial w_3} = \left[(\hat{y}_1 - y_1)w_7 + (\hat{y}_2 - y_2)w_9\right] a^{(1)}(1 - a^{(1)})\, x_3$$

Step 17: Chain Rule for w_4 (Expansion)



Gradient of the loss with respect to weight w_4:

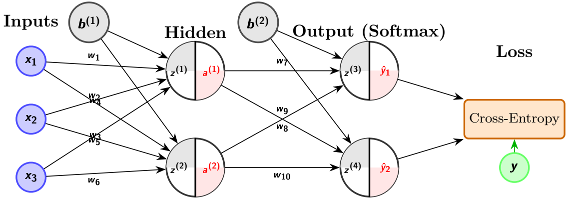
$$\frac{\partial L}{\partial w_4} = \frac{\partial L}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial w_4}$$

- $\frac{\partial a^{(2)}}{\partial z^{(2)}} = a^{(2)}(1 - a^{(2)})$
- $\frac{\partial z^{(2)}}{\partial w_4} = x_1$
- $\frac{\partial L}{\partial a^{(2)}} = (\hat{y}_1 - y_1)w_8 + (\hat{y}_2 - y_2)w_{10}$

Substituting:

$$\frac{\partial L}{\partial w_4} = \left[(\hat{y}_1 - y_1)w_8 + (\hat{y}_2 - y_2)w_{10}\right] a^{(2)}(1 - a^{(2)})\, x_1$$

Step 18: Chain Rule for w_5 (Expansion)



Gradient of the loss with respect to weight w_5:

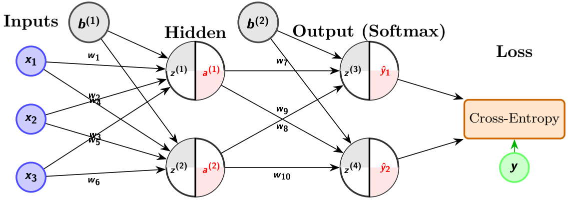
$$\frac{\partial L}{\partial w_5} = \frac{\partial L}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial w_5}$$

- $\frac{\partial a^{(2)}}{\partial z^{(2)}} = a^{(2)}(1 - a^{(2)})$
- $\frac{\partial z^{(2)}}{\partial w_5} = x_2$
- $\frac{\partial L}{\partial a^{(2)}} = (\hat{y}_1 - y_1)w_8 + (\hat{y}_2 - y_2)w_{10}$

Substituting:

$$\frac{\partial L}{\partial w_5} = \left[(\hat{y}_1 - y_1)w_8 + (\hat{y}_2 - y_2)w_{10}\right] a^{(2)}(1 - a^{(2)})\, x_2$$

Step 19: Chain Rule for w_6 (Expansion)



Gradient of the loss with respect to weight w_6:

$$\frac{\partial L}{\partial w_6} = \frac{\partial L}{\partial a^{(2)}} \cdot \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial w_6}$$

- $\frac{\partial a^{(2)}}{\partial z^{(2)}} = a^{(2)}(1 - a^{(2)})$
- $\frac{\partial z^{(2)}}{\partial w_6} = x_3$
- $\frac{\partial L}{\partial a^{(2)}} = (\hat{y}_1 - y_1)w_8 + (\hat{y}_2 - y_2)w_{10}$

Substituting:

$$\frac{\partial L}{\partial w_6} = \left[(\hat{y}_1 - y_1)w_8 + (\hat{y}_2 - y_2)w_{10}\right] a^{(2)}(1 - a^{(2)})\, x_3$$

Reach PostNetwork Academy

Website

www.postnetwork.co

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

GitHub Repositories

www.github.com/postnetworkacademy

Thank You!