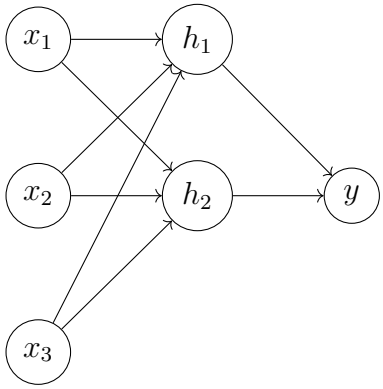
Class 3: Training by Hand

Slide 5: Forward Pass

We consider a simple neural network with two hidden units. Hidden unit 1 uses ReLU activation, hidden unit 2 uses ReLU, and the output uses a sigmoid activation.



We have input

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad y = 1.$$

Hidden layer:

$$z^{(1)} = W^{(1)}x + b^{(1)}, \quad h^{(1)} = \text{ReLU}(z^{(1)}).$$

Output layer:

$$z^{(2)} = W^{(2)}h^{(1)} + b^{(2)}, \quad \hat{y} = \sigma(z^{(2)}).$$

For training we use binary cross-entropy loss:

$$L = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})].$$

Suppose

$$W^{(1)} = \begin{bmatrix} 1 & -1 & 1 \\ -2 & 0 & 1 \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$W^{(2)} = \begin{bmatrix} 1 & -1 \end{bmatrix}, \quad b^{(2)} = 0.$$

Slide 6: Forward Pass Example

$$z^{(1)} = W^{(1)}x + b^{(1)} = \begin{bmatrix} 1 & -1 & 1 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$$h^{(1)} = \mathrm{ReLU}(z^{(1)}) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$z^{(2)} = W^{(2)}h^{(1)} + b^{(2)} = [1 \ -1] \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2$$

$$\hat{y} = \sigma(z^{(2)}) = \frac{1}{1+e^{-2}} \approx 0.881.$$

$$L = -\log(0.881) = 0.127.$$

Slide 7: Gradients

We compute gradients by backpropagation.

$$\frac{\partial L}{\partial W^{(2)}} = (\hat{y} - y)h^{(1)T}, \quad \frac{\partial L}{\partial b^{(2)}} = \hat{y} - y.$$

$$\delta^{(1)} = (\hat{y} - y)W^{(2)T} \odot \mathbf{1}(z^{(1)} > 0),$$

where $\odot$ is element-wise multiplication.

Then

$$\frac{\partial L}{\partial W^{(1)}} = \delta^{(1)}x^T, \quad \frac{\partial L}{\partial b^{(1)}} = \delta^{(1)}.$$

Epoch 1

Forward pass gave $\hat{y} = 0.881$, $L_1 = 0.127$.

Gradients:

$$\frac{\partial L}{\partial W^{(2)}} = -0.119 \begin{bmatrix} 2 & 0 \end{bmatrix} = \begin{bmatrix} -0.238 & 0 \end{bmatrix}, \quad \frac{\partial L}{\partial b^{(2)}} = -0.119.$$

$$\delta^{(1)} = -0.119 \begin{bmatrix} -1 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.119 \\ 0 \end{bmatrix}.$$

$$\frac{\partial L}{\partial W^{(1)}} = \delta^{(1)}x^T = \begin{bmatrix} -0.119 & 0 & -0.119 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial L}{\partial b^{(1)}} = \begin{bmatrix} -0.119 \\ 0 \end{bmatrix}.$$

Update parameters with $\eta = 0.1$:

$$W^{(2)} \leftarrow W^{(2)} - \eta \frac{\partial L}{\partial W^{(2)}} = [1 \ -1] - 0.1[-0.238 \ 0] = [1.024 \ -1],$$

$$b^{(2)} \leftarrow b^{(2)} - \eta \frac{\partial L}{\partial b^{(2)}} = 0 - 0.1(-0.119) = 0.012.$$

Epoch 2

$$W^{(1)} \leftarrow W^{(1)} - 0.1 \frac{\partial L}{\partial W^{(1)}} = \begin{bmatrix} 1.0119 & -1 & 1.0119 \\ -2 & 0 & 1 \end{bmatrix},$$

$$b^{(1)} \leftarrow b^{(1)} - 0.1 \frac{\partial L}{\partial b^{(1)}} = \begin{bmatrix} 0.0119 \\ 1 \end{bmatrix}.$$

Forward pass:

$$z^{(1)} = W^{(1)}x + b^{(1)} = \begin{bmatrix} 2.036 \\ -1 \end{bmatrix}, \quad h = \mathrm{ReLU}(z^{(1)}) = \begin{bmatrix} 2.036 \\ 0 \end{bmatrix}.$$

$$z^{(2)} = W^{(2)}h + b^{(2)} = 2.096, \quad \hat{y} = \sigma(2.096) = 0.891.$$

$$L_2 = -\log(0.891) = 0.116.$$

Loss has decreased: $L_2 < L_1$.

Epoch 3: Updates

$$\frac{\partial L}{\partial W^{(2)}} = \begin{bmatrix} -0.223 & 0 \end{bmatrix}, \quad \frac{\partial L}{\partial b^{(2)}} = -0.109.$$

$$\frac{\partial L}{\partial W^{(1)}} = \begin{bmatrix} -0.112 & 0 & -0.112 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial L}{\partial b^{(1)}} = \begin{bmatrix} -0.112 \\ 0 \end{bmatrix}.$$

Update:

$$W^{(2)} = [1.046 \ -1], \quad b^{(2)} = 0.023,$$

$$W^{(1)} = \begin{bmatrix} 1.023 & -1 & 1.023 \\ -2 & 0 & 1 \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} 0.023 \\ 1 \end{bmatrix}.$$

Next: new forward pass and loss calculation.