Exercise 7.1: Backpropagation

(a)

Given activation functions,

$$h = \text{ReLU}(W_1 x + b_1)$$

$$y = \text{Sigmoid}(W_2 h + b_2)$$

Computing $W_1 x + b_1 = \begin{bmatrix} 0.1 & 0.4 \\ 0.2 & 0.5 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.2 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}$

$$= \begin{bmatrix} 0.17 \\ 0.2 \\ 0.27 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.27 \\ 0.4 \\ 0.57 \end{bmatrix}$$

Applying ReLU,

$$h = \text{ReLU}\left(\begin{bmatrix} 0.27 \\ 0.4 \\ 0.57 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 0.27 \\ 0.4 \\ 0.57 \end{bmatrix}$$

Computing $W_2 h + b_2 = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.4 \end{bmatrix} \begin{bmatrix} 0.27 \\ 0.4 \\ 0.57 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$

$$= \begin{bmatrix} 0.33 \\ 0.48 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.43 \\ 0.68 \end{bmatrix}$$

Applying Sigmoid,

$$y = \text{Sigmoid}\left(\begin{bmatrix} 0.43 \\ 0.68 \end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{1}{1+e^{-0.43}} \\ \frac{1}{1+e^{-0.68}} \end{bmatrix}$$

$$= \begin{bmatrix} 0.60 \\ 0.66 \end{bmatrix}$$

(b)

The given loss function,

$$L = -\frac{1}{2} \sum_{i=1}^{2} \left[t_i \log(y_i) + (1-t_i) \log(1-y_i) \right]$$

Given, $t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $y = \begin{bmatrix} 0.60 \\ 0.66 \end{bmatrix}$

$$L = -\frac{1}{2}\left[1 \cdot \log(0.60) + (1-1) \cdot \log(1-0.60) + 0 \cdot \log(0.66) + (1-0) \cdot \log(1-0.66)\right]$$

$$= 0.3451$$

(c)

$$z_1 = W_1 x + b_1$$

$$z_2 = W_2 h + b_2$$

Applying the chain rule,

$$\frac{dL}{dW_2(1,1)} = \frac{dL}{dy_1} \cdot \frac{dy_1}{dz_2(1)} \cdot \frac{dz_2(1)}{dW_2(1,1)}$$

Finding the derivative of L with respect to y_1,

$$\frac{dL}{dy_1} = -\frac{1}{2}\left(\frac{t_1}{y_1} - \frac{1-t_1}{1-y_1}\right)$$

Substituting $t_1 = 1$ and $y_1 = 0.60$

$$\frac{dL}{dy_1} = -0.83$$

The derivative of the Sigmoid,

$$\frac{dy_1}{dz_2(1)} = y_1(1-y_1)$$

$$= 0.60(1-0.60)$$

$$= 0.24$$

For the final layer,

$$z_2(1) = W_2(1,1)h_1 + W_2(1,2)h_2 + W_3(1,3)h_3 + b_2(1)$$

Taking its derivative with respect to $W_2(1,1)$,

$$\frac{\partial z_2(1)}{\partial W_2(1,1)} = h_1$$

$$= 0.27$$

Combining everything,

$$\frac{\partial L}{\partial W_2(1,1)} = (-0.83)\cdot(0.24)\cdot(0.27) = 0.053$$

Computing gradient with respect to $b_2(1)$,

$$\frac{\partial L}{\partial b_2(1)} = \frac{\partial L}{\partial y_1}\cdot\frac{\partial y_1}{\partial z_2(1)}\cdot\frac{\partial z_2(1)}{\partial b_2(1)}$$

As b_2 is added directly, $\frac{\partial z_2(1)}{\partial b_2(1)} = 1$,

$$\frac{\partial L}{\partial b_2(1)} = (-0.83)\cdot(0.24)\cdot(1) = 0.19$$

Computing gradients with respect to $W_1(1,1)$,

$$\frac{\partial L}{\partial W_1(1,1)} = \frac{\partial L}{\partial y_1}\cdot\frac{\partial y_1}{\partial z_2(1)}\cdot\frac{\partial z_2(1)}{\partial h_1}\cdot\frac{\partial h_1}{\partial z_1(1)}\cdot\frac{\partial z_1(1)}{\partial W_1(1,1)}$$

In the above equation $\frac{\partial z_2(1)}{\partial h_1} = W_2(1,1) = 0.7$

Since $z_1(1) > 0$, $\frac{\partial h_1}{\partial z_1(1)} = 1$ [ReLU]

$$\frac{\partial z_1(1)}{\partial w_1(1,1)} = x_1 = 0.5$$

Finally,

$$\frac{\partial L}{\partial w_1(1,1)} = (-0.83)\cdot(0.24)\cdot(0.7)\cdot(1)\cdot(0.5) = -0.06$$

Computing gradients with respect to $b_1(2)$.

$$\frac{\partial L}{\partial b_1(2)} = \frac{\partial L}{\partial y_1}\cdot\frac{\partial y_1}{\partial z_2(1)}\cdot\frac{\partial z_2(1)}{\partial h_2}\cdot\frac{\partial h_2}{\partial z_1(2)}\cdot\frac{\partial z_1(2)}{\partial b_1(2)}$$

$$\frac{\partial z_2(1)}{\partial h_2} = w_2(1,2) = 0.2$$

$$\frac{\partial h_2}{\partial z_1(2)} = 1$$ because $z_1(2) > 0$ [ReLU]

$$\frac{\partial z_1(2)}{\partial b_1(2)} = 1$$ because $b_1(2)$ is added directly.

Combining all,

$$\frac{\partial L}{\partial b_1(2)} = (-0.83)\cdot(0.24)\cdot(0.2)\cdot(1)\cdot(1) = -0.03$$

Exercise 7.3: Backpropagation through sort

(a)

$F_1(x)$ rearranges the elements in x,

$$F_1(x) = (x_2, x_4, x_3, x_0, x_1)$$

The mapping of the gradients is direct, x_2 contributes to $F_1(x)_0$, x_4 contributes to $F_1(x)_1$ and goes on. The gradient at each position in x is the gradient of its corresponding position in $F_1(x)$.

So,

$$\frac{dL}{dx} = \begin{bmatrix} \frac{dL}{dF_1(x)_3} \\ \frac{dL}{dF_1(x)_4} \\ \frac{dL}{dF_1(x)_0} \\ \frac{dL}{dF_1(x)_2} \\ \frac{dL}{dF_1(x)_1} \end{bmatrix} = \begin{bmatrix} d_3 \\ d_4 \\ d_0 \\ d_2 \\ d_1 \end{bmatrix}$$

(b)

$$F_2(x) = (x_0 \cdot x, x_0 \cdot x_1, x_0 \cdot x_2, x_0 \cdot x_3, x_0 \cdot x_4)$$

For $F_2(x)_i = x_0 . x_i$,

when $i = 0$,

$$F_2(x)_0 = x_0 . x_0$$

$$\frac{dF_2(x)_0}{dx_0} = 2x_0$$

when $i \,!= 0$,

$$\frac{dF_2(x)_i}{dx_0} = x_i,$$

$$\frac{dF_2(x)_i}{dx_i} = x_0$$

The total gradients for each x_i combines all the terms,

$$\frac{dL}{dx_0} = \sum_{i=0}^{4} \frac{dL}{dF_2(x)_i} \cdot \frac{dF_2(x)_i}{dx_0}$$

$$= d_0 . 2x_0 + d_1 . x_1 + d_2 x_2 + d_3 x_3 + d_4 x_4$$

(c)

$$F_3(x) = (x_0 . x_2, x_0 . x_4, x_0 . x_3, x_0 . x_0, x_0 . x_1)$$

For x_0,

Given x_0 appears in every term,

$$\frac{dL}{dx_0} = \sum_{i=0}^{4} \frac{dL}{dF_3(x)_i} \cdot \frac{dF_3(x)_i}{dx_0}$$

So,

$$\frac{dL}{dx_0} = d_0 x_2 + d_1 x_4 + d_2 x_3 + d_3 \, 2x_0 + d_4 x_1$$

For x_2, x_4, x_3 and x_1,

$$\frac{dL}{dx_2} = d_0 \cdot x_0, \quad \frac{dL}{dx_4} = d_1 \cdot x_0, \quad \frac{dL}{dx_3} = d_2 \cdot x_0, \quad \frac{dL}{dx_1} = d_4 \cdot x_0$$