Rui Song 5576534 ; Hanzhi Zhuang 5578347 ; Tengyunhao Yang 5587382

Group Makabaka - Exercise 3.

2. 1) Forward Pass:

$L = |y - \hat{y}|$

$\hat{y} = g_2(z_2) = z_2$

$z_2 = w_2 h_1 + w_3 h_0$ $\Rightarrow$

Backward Pass:

$\frac{\partial L}{\partial \hat{y}} = \begin{cases} 1, & \hat{y} > y \\ -1, & \hat{y} \le y \end{cases}$

$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \cdot 1$

$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2} = \frac{\partial L}{\partial z_2} \cdot h_1$

$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_3} = \frac{\partial L}{\partial z_2} h_0$

$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_1} = \frac{\partial L}{\partial z_2} \cdot w_2.$

$h_1 = g_1(z_1) = ReLU(z_1) = \begin{cases} 0, & z_1 < 0 \\ z_1, & \text{else.} \end{cases}$ $\Rightarrow$

$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial h_1} \frac{\partial h_1}{\partial z_1} = \begin{cases} 0, & z_1 < 0 \\ \frac{\partial L}{\partial h_1} \cdot 1, & \text{else.} \end{cases}$

$z_1 = w_1 h_0$

$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial w_1} = \frac{\partial L}{\partial z_1} \cdot h_0$

$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_0} + \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial h_0} = \frac{\partial L}{\partial z_2} \cdot w_3 + \frac{\partial L}{\partial z_1} \cdot w_1$

$h_0 = g_0(z_0) = ReLU(z_0) = \begin{cases} 0, & z_0 < 0 \\ z_0, & \text{else} \end{cases}$

$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial h_0} \cdot \frac{\partial h_0}{\partial z_0} = \begin{cases} 0, & z_0 < 0 \\ \frac{\partial L}{\partial h_0} \cdot 1, & \text{else.} \end{cases}$

$z_0 = w_0 x$

$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial z_0} \frac{\partial z_0}{\partial w_0} = \frac{\partial L}{\partial z_0} \cdot x$

2) The third layer receives two sources of inputs, which can help avoiding the vanishing gradient problem. The skip connection of w_3 will help bypass the layer 1 and make the gradient uninterrupted.

3) $(x_1, y_1) = (1, -3)$

$w_0 = w_1 = w_2 = w_3 = 0.5$

LearningRate = 1

Forward Pass:

$z_0 = w_0 x = 0.5$

$h_0 = g_0(z_0) = ReLU(z_0) = 0.5$

$z_1 = w_1 h_0 = 0.5 \times 0.5 = 0.25$

$h_1 = g_1(z_1) = ReLU(z_1) = 0.25$

Backward Pass:

$\frac{\partial L}{\partial \hat{y}} = 1$

$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2} = 1$

$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_1} = 0.5$

$z_2 = w_2 h_1 + w_3 h_0 = 0.5 \times 0.25 + 0.5 \times 0.5 = 0.375$

$\hat{y}_1 = g_2(z_2) = z_2 = 0.375$

$L = |y_1 - \hat{y}_1| = 3.375$

$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} = 0.5$$

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial h_0} + \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial h_0} = 0.75$$

$$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial h_0}\frac{\partial h_0}{\partial z_0} = 0.75$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial w_2} = 1 \times 0.25 = 0.25$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_1} = 0.5 \times 0.5 = 0.25$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_3} = 1 \times 0.5 = 0.5$$

$$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial z_0} \cdot \frac{\partial z_0}{\partial w_0} = 0.5 \times 1 = 0.75$$

$\alpha = 1$

$\Rightarrow$ Gradient Descent: $w_0 = w_0 - \alpha\frac{\partial L}{\partial w_0} = 0.5 - 0.75 = -0.25$

$w_1 = w_1 - \alpha\frac{\partial L}{\partial w_1} = 0.5 - 0.25 = 0.25$

$w_2 = w_2 - \alpha\frac{\partial L}{\partial w_2} = 0.5 - 0.25 = 0.25$

$w_3 = w_3 - \alpha\frac{\partial L}{\partial w_3} = 0.5 - 0.5 = 0$

Forward Pass:

$z_0 = w_0 x = -0.25$

$h_0 = g_0(z_0) = ReLU(z_0) = 0$

$z_1 = w_1 h_0 = 0$

$h_1 = g_1(z_1) = ReLU(z_1) = 0$

$z_2 = w_2 h_1 + w_3 h_0 = 0.25 \times 0 + 0 \times 0 = 0$

$\hat{y} = g(z_2) = z_2 = 0$

$\Rightarrow L = |y - \hat{y}| = 3.$

4.1) The loss remains the same.

2). The loss goes down after multiple steps of updating parameters (from 0.390529 to 0.000922). The loss does not always decrease as fluctuations exist. Loss may increase a little bit (for example, when reaching another local minima) before decrease again.

3). No. We also meet a situation where the loss for each epoch are always 0.693, which is $-\log\frac{1}{2}$, this means the model doesn't learn at all, the weights are not updated. As the weights are initialized randomly, the start point for each gradient descent are different, which for some of them may stuck in local minima.

4) The 'lr' is learning rate, which is a hyperparameter determines the size of the step taken during optimization (gradient descent). A relative larger value can speed up convergence, but it may miss the minima or oscillate around it. We can set a larger value at the very beginning of the training. When we want the optimize process to be stable, and lower convergence speed can be accepted, we can choose small value.