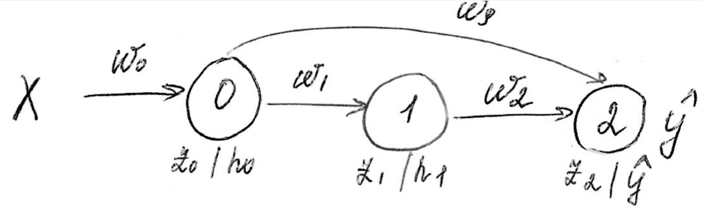
Sheet 03

Montag, 13. November 2023 20:49

Team MTE

Marta Gulida - 5585808 - mg776
Erik Bode - 4505199 - kb301
Tillman Heisner - 4517815 - th273



L1-loss: $L(\hat{y}, y) = |y - \hat{y}|$

$$g_0(z) = g_1(z) = \begin{cases} 0, & z < 0 \\ z, & \text{else} \end{cases}$$

$$g_2(z) = z_2 \qquad \text{no biases}$$

$z = Wx + b$ z - the value before applying the activation function.

1) Backpropagation

$$h(z) = \begin{cases} z, & z \geq 0 \\ 0, & \text{else} \end{cases} \Rightarrow$$

$$\Rightarrow h'(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{else} \end{cases}$$

$$h(z) = z \Rightarrow h'(z) = 1$$

Forward pass:

$$z_0 = w_0 x$$

$$h_0 = g_0(z_0)$$

$$z_1 = w_1 h_0$$

$$h_1 = g_1(z_1)$$

$$z_2 = w_2 h_1 + w_3 h_0$$

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

Backward pass:

$$\frac{\partial L}{\partial \hat{y}} = \begin{cases} 1 & \text{if } \hat{y} > 0 \\ -1 & \text{if } \hat{y} < 0 \end{cases} \qquad L(\hat{y}, y) = |y - \hat{y}|$$

$$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_1} = \begin{cases} 1 \cdot g_2'(z) \text{ if } & \hat{y} > 0 \\ -1 \cdot g_2'(z) & \text{if } \hat{y} < 0 \end{cases} = \begin{cases} 1 & \text{if } \hat{y} > 0 \\ -1 & \text{if } \hat{y} < 0 \end{cases}$$

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

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

$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_1} = \cancel{\begin{cases} \frac{\partial L}{\partial z_2} & \text{if } z_1 > 0 \\ 0 & \text{if } z_1 \leq 0 \end{cases}} \; \frac{\partial L}{\partial z_2} \cdot w_2$$

$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} = \begin{cases} \cancel{\frac{\partial L}{\partial z_2}} \; \frac{\partial L}{\partial h_1} & \text{if } z_1 > 0 \\ 0 & \text{if } z_1 \leq 0 \end{cases}$$

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

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

$$\frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial h_0} \cdot \frac{\partial h_0}{\partial z_0} = \begin{cases} \frac{\partial L}{\partial h_0} & \text{if } z_0 \geq 0 \\ 0 & \text{if } z_0 \leq 0 \end{cases}$$

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

2) Skip connection method can make difference because it combines simple features with more complex features from deeper layers.

3) $(x_1, y_1) = (1, -3)$ $\quad w_0 = w_1 = w_2 = w_3 = \frac{1}{2}$ $\quad$ Learning Rate $= 1$

$w = w - \alpha \frac{\partial L}{\partial w}$ ← updating the weights

1st Forward pass:

$z_0 = \frac{1}{2} \cdot 1 = \frac{1}{2}$

$h_0 = \frac{1}{2}$

$z_1 = w_1 \cdot h_0 = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$h_1 = \frac{1}{4}$

$z_2 = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} + \frac{1}{4} = \frac{1}{8} + \frac{2}{8} = \frac{3}{8}$

$\hat{y} = \frac{3}{8}$

First iteration of updating the weights:

$w_2^1 = w_2 - \frac{\partial L}{\partial w_2}$

$w_2^1 = \frac{1}{2} - h_1 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

$w_3^1 = w_3 - \frac{\partial L}{\partial w_3}$

$w_3^1 = \frac{1}{2} - h_0 = \frac{1}{2} - \frac{1}{2} = 0$

$w_1^1 = w_1 - \frac{\partial L}{\partial w_1}$

$w_1^1 = \frac{1}{2} - h_0 \cdot w_2 = \frac{1}{2} - \frac{1}{2} \cdot \frac{1}{2} =$

$= \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

$w_0^1 = w_0 - \frac{\partial L}{\partial w_0} =$

$w_0^1 = \frac{1}{2} - x \cdot (w_3 + w_1 \cdot w_2) =$

$= \frac{1}{2} - 1\left(\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}\right) = \frac{1}{2} -$

$-\left(\frac{1}{2} + \frac{1}{4}\right) = \frac{1}{2} - \frac{3}{4} = \frac{2-3}{4} = -\frac{1}{4}$

The weights and the loss after one gradient descent step:

$w_0 = -\frac{1}{4} \quad w_1 = \frac{1}{4} \quad w_2 = \frac{1}{4} \quad w_3 = 0$

$z_0 = -\frac{1}{4} \cdot 1 = -\frac{1}{4}$

$h_0 = 0$

$z_1 = w_1 \cdot h_0 = \frac{1}{4} \cdot 0 = 0$

$h_1 = 0$

$z_2 = w_2 \cdot h_1 + w_3 \cdot h_0 = 0 \qquad \hat{y} = z_2 = 0 \qquad L(y, \hat{y}) = |y - \hat{y}| = |3 - 0| = 3.$

Exercise 4

Montag, 13. November 2023 20:20

1) What do you observe with regards to the loss after 1 step of updating parameters?
   We observe, that the weights are updating and the gradients are not 0 anymore, but the loss does not change, even though the prediction sometimes gets better.

2) What do you observe after multiple steps of updating parameters? Does the loss always decrease? Explain why it may not.
   In some cases, the loss keeps decreasing until the function is learned. In other cases, the loss is not decreasing, indicating that the target function is not learned. A possible explaination could be a poorly choosen hyperparameter.

3) Run the experiment multiple times. Do you always end up with the correct final predictions? Explain why or why not?
   We have mutiple possible end-states, some have two wrongly predicted results, others have one or even none.
   Sometimes, the weights get negative and as such are set to zero, as in the hand-written example.
   This stops the learning process.

4) What is the role of the variable lr above? When would you set it to a relatively larger valueand when would you set it to a relatively smaller value? Explain.
   The value lr indicates the step size, how much the weights get updated in the direction of the gradient.
   When no value is learned or the results are unstable, it should be decreased, since this indicates overshooting the minimum.
   When we consistently learn a bit, so the loss is decreasing slowly, the learning rate should be increased.