Exercise 5

Pattern Recognition, Fall 2021

Nico Aebischer

Max Jappert

Theory/Coding questions

1. In linear neural networks, the output is predicted as a linear function of the inputs. Every node in the network just sums up its inputs, multiplied by the respective weights. This sum is then passed to the next node.
   Non-linear neural networks can predict non-linear outputs as well. This is achieved with the help of non-linear activation functions, which "preprocess" the content of a node, before it is passed onto the next one. This can be done for some or all the layers.
   An example of such activation functions is the Sigmoid function (it is actually a family of functions). One specific function would be the one we use in the first exercise, namely:

   $$a(t) = \frac{1}{1 + e^{-t}}$$

   Another family of such functions is the ReLU (Rectified Linear Unit) family. One specific function is the following:

   $$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

2. The "vanishing gradients" problem is very common in neural networks and appears as follows: The more layers are added to a network (all using certain activation functions), the harder it gets to train. The reason behind this phenomenon is that the gradient of the loss function gets close to zero. This is caused by certain activation functions such as a sigmoid function. It squashes a large input space to a small space between 0 and 1. Thus, large changes in the input of the sigmoid function only cause small changes of its output. That is why the derivative becomes very small. In networks with only a small number of layers, this is not a big problem. For deep neural nets however, this can cause the training to be completely ineffective. The gradients are found using backpropagation. Following the chain rule, the derivatives of every layer are multiplied down the network, from the output to the input layer. When the number of hidden layers is large and an activation function such as a sigmoid function is used, a large number of small derivatives are multiplied together. Therefore, the gradient gets exponentially smaller, resulting in the inability to effectively update the weights and biases.
   The most obvious approach to solve this problem is using other activation functions, which do not result in such a small derivative. An example would be a ReLU function.

3. - We calculated the network loss as follows:

$h_1 = a(h_{1in}) = a(x_1 w_1 + x_2 w_3 + b_1 w_5) = a(0.7 + 1.5) = \frac{1}{1+e^{-2.2}} \approx 0.9$

$h_2 = a(h_{2in}) = a(x_1 w_2 + x_2 w_4 + b_1 w_6) = a(1.3 + 0.1) = \frac{1}{1+e^{-1.4}} \approx 0.8$

$\hat{y} \approx h_1 w_7 + h_2 w_8 + b_2 w_9 = 0.9 \cdot 0.7 + 0.8 \cdot 0.8 \approx 1.27$

$\mathcal{L} = \frac{1}{2}(\hat{y} - y)^2 \approx \frac{1}{2}(0.27)^2 = 0.04$

- $\frac{\partial \mathcal{L}}{\partial w_1} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}_{in}}}_{=\hat{y}-y} \underbrace{\frac{\partial \hat{y}_{in}}{\partial \hat{y}}}_{=1} \underbrace{\frac{\partial \hat{y}}{\partial h_1}}_{=w_7} \underbrace{\frac{\partial h_1}{\partial h_{1in}}}_{=a(h_1)(1-a(h_1))} \underbrace{\frac{\partial h_{1in}}{\partial w_1}}_{=x_1} \approx 0.27 \cdot 1 \cdot 0.7 \cdot 0.09 \approx 0.02$

$\frac{\partial \mathcal{L}}{\partial w_3} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}_{in}}}_{=\hat{y}-y} \underbrace{\frac{\partial \hat{y}_{in}}{\partial \hat{y}}}_{=1} \underbrace{\frac{\partial \hat{y}}{\partial h_1}}_{=w_7} \underbrace{\frac{\partial h_1}{\partial h_{1in}}}_{=a(h_1)(1-a(h_1))} \underbrace{\frac{\partial h_{1in}}{\partial w_3}}_{=x_2} \approx 0.27 \cdot 1 \cdot 0.7 \cdot 0.09 \approx 0.02$

$\frac{\partial \mathcal{L}}{\partial w_2} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}_{in}}}_{=\hat{y}-y} \underbrace{\frac{\partial \hat{y}_{in}}{\partial \hat{y}}}_{=1} \underbrace{\frac{\partial \hat{y}}{\partial h_2}}_{=w_8} \underbrace{\frac{\partial h_2}{\partial h_{2in}}}_{=a(h_2)(1-a(h_2))} \underbrace{\frac{\partial h_{2in}}{\partial w_2}}_{=x_1} \approx 0.27 \cdot 1 \cdot 0.8 \cdot 0.16 \approx 0.03$

$\frac{\partial \mathcal{L}}{\partial w_4} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}_{in}}}_{=\hat{y}-y} \underbrace{\frac{\partial \hat{y}_{in}}{\partial \hat{y}}}_{=1} \underbrace{\frac{\partial \hat{y}}{\partial h_2}}_{=w_8} \underbrace{\frac{\partial h_2}{\partial h_{2in}}}_{=a(h_2)(1-a(h_2))} \underbrace{\frac{\partial h_{2in}}{\partial w_4}}_{=x_2} \approx 0.27 \cdot 1 \cdot 0.8 \cdot 0.16 \approx 0.03$

$\frac{\partial \mathcal{L}}{\partial w_5} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}_{in}}}_{=\hat{y}-y} \underbrace{\frac{\partial \hat{y}_{in}}{\partial \hat{y}}}_{=1} \underbrace{\frac{\partial \hat{y}}{\partial h_2}}_{=w_8} \underbrace{\frac{\partial h_2}{\partial h_{2in}}}_{=a(h_2)(1-a(h_2))} \underbrace{\frac{\partial h_{2in}}{\partial w_5}}_{=1} \approx 0.27 \cdot 1 \cdot 0.7 \cdot 0.09 \approx 0.02$

$\frac{\partial \mathcal{L}}{\partial w_6} = \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{y}_{in}}}_{=\hat{y}-y} \underbrace{\frac{\partial \hat{y}_{in}}{\partial \hat{y}}}_{=1} \underbrace{\frac{\partial \hat{y}}{\partial h_2}}_{=w_8} \underbrace{\frac{\partial h_2}{\partial h_{2in}}}_{=a(h_2)(1-a(h_2))} \underbrace{\frac{\partial h_{2in}}{\partial w_6}}_{=1} \approx 0.27 \cdot 1 \cdot 0.8 \cdot 0.16 \approx 0.03$

$\frac{\partial \mathcal{L}}{\partial w_7} = \frac{\partial \hat{y}}{\partial w_7} \frac{\partial \mathcal{L}}{\partial \hat{y}} = h_1(\hat{y} - y) \approx 0.24$

$\frac{\partial \mathcal{L}}{\partial w_8} = \frac{\partial \hat{y}}{\partial w_8} \frac{\partial \mathcal{L}}{\partial \hat{y}} = h_2(\hat{y} - y) \approx 0.22$

$\frac{\partial \mathcal{L}}{\partial w_9} = \frac{\partial \hat{y}}{\partial w_9} \frac{\partial \mathcal{L}}{\partial \hat{y}} = 1(\hat{y} - y) \approx 0.27$

- $w_1^* = w_1 - \eta \frac{\partial \mathcal{L}}{\partial w_1} = 0.7 - 0.2 \cdot 0.02 \approx 0.7$
  $w_2^* = w_2 - \eta \frac{\partial \mathcal{L}}{\partial w_2} = 1.3 - 0.2 \cdot 0.03 \approx 1.3$
  $w_3^* = w_3 - \eta \frac{\partial \mathcal{L}}{\partial w_3} = 1.5 - 0.2 \cdot 0.02 \approx 1.5$
  $w_4^* = w_4 - \eta \frac{\partial \mathcal{L}}{\partial w_4} = 0.1 - 0.2 \cdot 0.03 \approx 0.1$
  $w_5^* = w_5 - \eta \frac{\partial \mathcal{L}}{\partial w_5} = 0 - 0.2 \cdot 0.02 \approx -0.004$
  $w_6^* = w_4 - \eta \frac{\partial \mathcal{L}}{\partial w_6} = 0 - 0.2 \cdot 0.03 \approx -0.006$

$$w_7^* = w_7 - \eta \frac{\partial \mathcal{L}}{\partial w_7} = 0.7 - 0.2 \cdot 0.24 \approx 0.65$$
$$w_8^* = w_8 - \eta \frac{\partial \mathcal{L}}{\partial w_8} = 0.8 - 0.2 \cdot 0.22 \approx 0.7$$
$$w_9^* = w_9 - \eta \frac{\partial \mathcal{L}}{\partial w_9} = 0 - 0.2 \cdot 0.27 \approx -0.054$$

- With the new weights the loss function is calculated as such:

  $$h_1' = a(0.7 + 1.5 - 0.004) \approx 0.9$$

  $$h_2' = a(1.3 + 0.1 - 0.006) \approx 0.8$$

  $$\hat{y}' = h_1' w_7^* + h_2' w_8^* + b_2 w_9^* = 0.9 \cdot 0.65 + 0.8 \cdot 0.7 - 1 \cdot 0.054 = 1.092$$

  $$\mathcal{L}' = \tfrac{1}{2}(0.092)^2 = 0.0042$$

  And we can indeed observe that the prediction error decreased after the weights were updated with gradient descent.

4. We get the following losses for $\eta = 0.2$:

```
Error: 0.0370
Error: 0.0093
Error: 0.0024
Error: 0.0006
Error: 0.0002
Error: 0.0000
Error: 0.0000
Error: 0.0000
Error: 0.0000
Error: 0.0000
```

5. We get the following losses for $\eta = 1$:

```
Error: 0.0370
Error: 0.0813
Error: 0.1745
Error: 0.3822
Error: 0.7823
Error: 1.6351
Error: 2.6870
Error: 4.0413
Error: 1.7092
Error: 0.6663
```