Large Scale Machine Learning and Deep Learning
Review Questions 4

1. Why adding more neurons to a single layer neural network cannot solve the XOR problem, but adding more layers can?

   Answer: page 170 of the machine learning book

---

2. Assume a feedforward neural network with one hidden layer, in which the output of the hidden units and output units are computed by functions $\mathbf{h} = \mathtt{f}(\mathbf{x})$ and $\mathbf{out} = \mathtt{g}(\mathbf{h})$, respectively. Show that if we use linear functions in `f` and `g`, e.g., $\mathbf{h} = \mathtt{f}(\mathbf{x}) = \mathbf{w}_1^\mathsf{T}\mathbf{x}$ and $\mathbf{out} = \mathtt{g}(\mathbf{h}) = \mathbf{w}_2^\mathsf{T}\mathbf{h}$, then the feedforward network as a whole would remain a linear function of its input.

   Answer: page 171 of the machine learning book

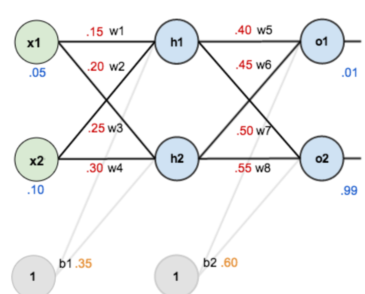
---

3. What's the problem of using `step` function as an activation function in deep feedforward neural networks?

   Answer: page 262 of the handson book

---

4. Compute the value of $\mathtt{w}_2$ and $\mathtt{w}_8$ after the first iteration of the backpropagation in the following figure. Assume all the neurons use the ReLU activation function and we use squared error function as the cost function. In this figure, red and orange colors indicate the initial values of the weights and biases, while the numbers in blue show the input and true output values.



Answer:

▷ **Forward pass**

$$
\begin{aligned}
\mathtt{net_{h1}} &= \mathtt{w_1 x_1 + w_2 x_2 + b_1} = 0.3775 \\
\mathtt{out_{h1}} &= \mathtt{max}(0, \mathtt{net_{h1}}) = 0.3775
\end{aligned}
$$

$$
\begin{aligned}
\mathtt{net_{h2}} &= \mathtt{w_3 x_1 + w_4 x_2 + b_1} = 0.3925 \\
\mathtt{out_{h2}} &= \mathtt{max}(0, \mathtt{net_{h2}}) = 0.3925
\end{aligned}
$$

$$
\begin{aligned}
\mathtt{net_{o1}} &= \mathtt{w_5 out_{h1} + w_6 out_{h2} + b_2} = 0.9276 \\
\mathtt{out_{o1}} &= \mathtt{max}(0, \mathtt{net_{o1}}) = 0.9276
\end{aligned}
$$

$$
\begin{aligned}
\mathtt{net_{o2}} &= \mathtt{w_7 out_{h1} + w_8 out_{h2} + b_2} = 1.0046 \\
\mathtt{out_{o2}} &= \mathtt{max}(0, \mathtt{net_{o2}}) = 1.0046
\end{aligned}
$$

Computing the total error:

$$
\begin{aligned}
\mathtt{E_{o1}} &= \frac{1}{2}(\mathtt{y_{o1}} - \mathtt{out_{o1}})^2 = 0.4210 \\
\mathtt{E_{o2}} &= \frac{1}{2}(\mathtt{y_{o2}} - \mathtt{out_{o2}})^2 = 0.0001 \\
\mathtt{E_{total}} &= \mathtt{E_{o1}} + \mathtt{E_{o2}} = 0.4211
\end{aligned}
$$

▷ **Backward pass**

$$
\begin{aligned}
\mathtt{ReLU(z)} &= \max(0, \mathtt{z}) \\
\frac{\partial \mathtt{ReLU(z)}}{\partial \mathtt{z}} &= \begin{cases} 0 & \text{if } \mathtt{z} \le 0 \\ 1 & \text{if } \mathtt{z} > 0 \end{cases}
\end{aligned}
$$

1. Computing the partial derivative with respect to $\mathtt{w_8}$:

$$
\begin{aligned}
\frac{\partial \mathtt{E_{total}}}{\partial \mathtt{w_8}} &= \frac{\partial \mathtt{E_{o1}}}{\partial \mathtt{w_8}} + \frac{\partial \mathtt{E_{o2}}}{\partial \mathtt{w_8}} \\
&= 0 + \frac{\partial \mathtt{E_{o2}}}{\partial \mathtt{w_8}} \\
&= \frac{\partial \mathtt{E_{o2}}}{\partial \mathtt{out_{o2}}} \times \frac{\partial \mathtt{out_{o2}}}{\partial \mathtt{net_{o2}}} \times \frac{\partial \mathtt{net_{o2}}}{\partial \mathtt{w_8}} \\
&= \frac{\partial(\frac{1}{2}(\mathtt{y_{o2}} - \mathtt{out_{o2}})^2)}{\partial \mathtt{out_{o2}}} \times \frac{\partial \mathtt{ReLU(net_{o2})}}{\partial \mathtt{net_{o2}}} \times \frac{\partial \mathtt{w_7 out_{h1} + w_8 out_{h2}}}{\partial \mathtt{w_8}} \\
&= (\mathtt{out_{o2}} - \mathtt{y_{o2}}) \times 1 \times \mathtt{out_{h2}} = 0.0057
\end{aligned}
$$

Update the weight $\mathtt{w_8}$:

$$
\begin{aligned}
\mathtt{w_8^{(next)}} &= \mathtt{w_8} - \eta \frac{\partial \mathtt{E_{total}}}{\partial \mathtt{w_8}} \\
&= 0.55 - \eta 0.0057
\end{aligned}
$$

2. Computing the partial derivative with respect to $\mathtt{w_2}$:

$$
\begin{aligned}
\frac{\partial \mathtt{E_{total}}}{\partial \mathtt{w_2}} &= \frac{\partial \mathtt{E_{o1}}}{\partial \mathtt{w_2}} + \frac{\partial \mathtt{E_{o2}}}{\partial \mathtt{w_2}} \\
&= \frac{\partial \mathtt{E_{o1}}}{\partial \mathtt{out_{o1}}} \times \frac{\partial \mathtt{out_{o1}}}{\partial \mathtt{net_{o1}}} \times \frac{\partial \mathtt{net_{o1}}}{\partial \mathtt{out_{h1}}} \times \frac{\partial \mathtt{out_{h1}}}{\partial \mathtt{net_{h1}}} \times \frac{\partial \mathtt{net_{h1}}}{\partial \mathtt{w_2}} \\
&+ \frac{\partial \mathtt{E_{o2}}}{\partial \mathtt{out_{o2}}} \times \frac{\partial \mathtt{out_{o2}}}{\partial \mathtt{net_{o2}}} \times \frac{\partial \mathtt{net_{o2}}}{\partial \mathtt{out_{h1}}} \times \frac{\partial \mathtt{out_{h1}}}{\partial \mathtt{net_{h1}}} \times \frac{\partial \mathtt{net_{h1}}}{\partial \mathtt{w_2}} \\
&= \frac{\partial (\frac{1}{2}(\mathtt{y_{o1}} - \mathtt{out_{o1}})^2)}{\partial \mathtt{out_{o1}}} \times \frac{\partial \mathtt{ReLU}(\mathtt{net_{o1}})}{\partial \mathtt{net_{o1}}} \times \frac{\partial \mathtt{w_5 out_{h1}} + \mathtt{w_6 out_{h2}}}{\partial \mathtt{out_{h1}}} \times \frac{\partial \mathtt{ReLU}(\mathtt{net_{h1}})}{\partial \mathtt{net_{h1}}} \times \frac{\partial \mathtt{w_1 x1} + \mathtt{w_2 x2}}{\partial \mathtt{w_2}} \\
&+ \frac{\partial (\frac{1}{2}(\mathtt{y_{o2}} - \mathtt{out_{o2}})^2)}{\partial \mathtt{out_{o2}}} \times \frac{\partial \mathtt{ReLU}(\mathtt{net_{o2}})}{\partial \mathtt{net_{o2}}} \times \frac{\partial \mathtt{w_7 out_{h1}} + \mathtt{w_8 out_{h2}}}{\partial \mathtt{out_{h1}}} \times \frac{\partial \mathtt{ReLU}(\mathtt{net_{h1}})}{\partial \mathtt{net_{h1}}} \times \frac{\partial \mathtt{w_1 x1} + \mathtt{w_2 x2}}{\partial \mathtt{w_2}} \\
&= [(\mathtt{out_{o1}} - \mathtt{y_{o1}}) \times 1 \times \mathtt{w_5} \times 1 \times \mathtt{x2}] + [(\mathtt{out_{o2}} - \mathtt{y_{o2}}) \times 1 \times \mathtt{w_7} \times 1 \times \mathtt{x2}] = 0.0374
\end{aligned}
$$

Update the weight $\mathtt{w_2}$:

$$
\begin{aligned}
\mathtt{w_2^{(next)}} &= \mathtt{w_2} - \eta \frac{\partial \mathtt{E_{total}}}{\partial \mathtt{w_2}} \\
&= 0.2 - \eta 0.0374
\end{aligned}
$$