Exercise 7.8 from "Learning from Data"

Chayce Leonard

3.17.25

Exercise 7.8

Repeat the computations in Example 7.1 for the case when the output transformation is the identity. You should compute $\mathbf{s}^{(l)}$, $\mathbf{x}^{(l)}$, $\boldsymbol{\delta}^{(l)}$ and $\frac{\partial e}{\partial \mathbf{W}^{(l)}}$.

Solution

Given Information

- Weight matrices: $\mathbf{W}^{(1)} = [0.1, 0.2, 0.3, 0.4]^T$, $\mathbf{W}^{(2)} = [0.2, 1, -3]^T$, $\mathbf{W}^{(3)} = [1, 2]^T$
- Data point: $x = 2$, $y = 1$
- Hidden layers use tanh activation, output layer uses identity function

Forward Propagation

Input Layer

$$\mathbf{x}^{(0)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{(with bias term)} \tag{1}$$

First Hidden Layer

$$\begin{aligned}
\mathbf{s}^{(1)} &= (\mathbf{W}^{(1)})^T \mathbf{x}^{(0)} && (2)\\
&= \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \end{bmatrix}^T \begin{bmatrix} 1 \\ 2 \end{bmatrix} && (3)\\
&= \begin{bmatrix} 0.1(1) + 0.2(2) \\ 0.3(1) + 0.4(2) \end{bmatrix} && (4)\\
&= \begin{bmatrix} 0.7 \\ 1.0 \end{bmatrix} && (5)
\end{aligned}$$

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ \tanh(0.7) \\ \tanh(1.0) \end{bmatrix} \tag{6}$$

$$= \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix} \quad \text{(with bias term)} \tag{7}$$

Second Hidden Layer

$$\mathbf{s}^{(2)} = (\mathbf{W}^{(2)})^T \mathbf{x}^{(1)} \tag{8}$$

$$= \begin{bmatrix} 0.2 & 1 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix} \tag{9}$$

$$= 0.2(1) + 1(0.60) - 3(0.76) \tag{10}$$

$$= 0.2 + 0.60 - 2.28 \tag{11}$$

$$= -1.48 \tag{12}$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ \tanh(-1.48) \end{bmatrix} \tag{13}$$

$$= \begin{bmatrix} 1 \\ -0.90 \end{bmatrix} \quad \text{(with bias term)} \tag{14}$$

Output Layer

$$\mathbf{s}^{(3)} = (\mathbf{W}^{(3)})^T \mathbf{x}^{(2)} \tag{15}$$

$$= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -0.90 \end{bmatrix} \tag{16}$$

$$= 1(1) + 2(-0.90) \tag{17}$$

$$= 1 - 1.8 \tag{18}$$

$$= -0.8 \tag{19}$$

Since the output transformation is the identity function:

$$\mathbf{x}^{(3)} = \mathbf{s}^{(3)} = -0.8 \tag{20}$$

Backpropagation

Output Layer Sensitivity

For the identity function, the derivative is 1, so:

$$\begin{aligned}
\boldsymbol{\delta}^{(3)} &= (\mathbf{x}^{(3)} - y) \cdot \frac{d}{dz} z && (21)\\
&= (-0.8 - 1) \cdot 1 && (22)\\
&= -1.8 && (23)
\end{aligned}$$

Second Hidden Layer Sensitivity

$$\begin{aligned}
\boldsymbol{\delta}^{(2)} &= (1 - \tanh^2(-1.48)) \cdot \mathbf{W}_2^{(3)} \cdot \boldsymbol{\delta}^{(3)} && (24)\\
&= (1 - (-0.90)^2) \cdot 2 \cdot (-1.8) && (25)\\
&= 0.19 \cdot 2 \cdot (-1.8) && (26)\\
&= -0.684 && (27)
\end{aligned}$$

First Hidden Layer Sensitivity

$$\begin{aligned}
\boldsymbol{\delta}^{(1)} &= \begin{bmatrix} 1 - \tanh^2(0.7) \\ 1 - \tanh^2(1) \end{bmatrix} \odot \left(\begin{bmatrix} 1 \\ -3 \end{bmatrix} \cdot (-0.684) \right) && (28)\\
&= \begin{bmatrix} 0.64 \\ 0.422 \end{bmatrix} \odot \begin{bmatrix} -0.684 \\ 2.052 \end{bmatrix} && (29)\\
&= \begin{bmatrix} -0.438 \\ 0.866 \end{bmatrix} && (30)
\end{aligned}$$

Partial Derivatives

For the First Layer

$$\begin{aligned}
\frac{\partial e}{\partial \mathbf{W}^{(1)}} &= \mathbf{x}^{(0)} (\boldsymbol{\delta}^{(1)})^T && (31)\\
&= \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} -0.438 & 0.866 \end{bmatrix} && (32)\\
&= \begin{bmatrix} -0.438 & 0.866 \\ -0.876 & 1.732 \end{bmatrix} && (33)
\end{aligned}$$

For the Second Layer

$$\frac{\partial e}{\partial \mathbf{W}^{(2)}} = \mathbf{x}^{(1)}(\boldsymbol{\delta}^{(2)})^T \tag{34}$$

$$= \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix} \cdot (-0.684) \tag{35}$$

$$= \begin{bmatrix} -0.684 \\ -0.410 \\ -0.520 \end{bmatrix} \tag{36}$$

For the Output Layer

$$\frac{\partial e}{\partial \mathbf{W}^{(3)}} = \mathbf{x}^{(2)}(\boldsymbol{\delta}^{(3)})^T \tag{37}$$

$$= \begin{bmatrix} 1 \\ -0.90 \end{bmatrix} \cdot (-1.8) \tag{38}$$

$$= \begin{bmatrix} -1.8 \\ 1.62 \end{bmatrix} \tag{39}$$

Conclusion

The key difference from Example 7.1 is that using the identity function in the output layer makes $\boldsymbol{\delta}^{(3)} = -1.8$ (versus -1.855 with tanh), which propagates through the network affecting all subsequent calculations.