Homework 2 Written Solutions (39 points)

Kathleen McKeown, Fall 2019
COMS W4705: Natural Language Processing

1 Dense Network – Forward (10 points)

Suppose we have a simple dense network. The forward pass of this network is described by:

$$Z_1 = W_1 A_{in} + b_1 \tag{1}$$
$$A_1 = f(Z_1) \tag{2}$$
$$Z_{out} = W_{out} A_1 + b_{out} \tag{3}$$
$$A_{out} = f_{out}(Z_{out}) \tag{4}$$

where 1-2 describe a feed-forward layer and 3-4 describe the output layer.

Suppose that you are given the following:

$$W_1 = \begin{bmatrix} 1 & -1 & 2 & 3 & 0 \\ 4 & 0 & -1 & 1 & 3 \\ 2 & 1 & 3 & -5 & -4 \\ 4 & -3 & 2 & 1 & -3 \end{bmatrix} \quad b_1 = \begin{bmatrix} -1 \\ 2 \\ -4 \\ 3 \end{bmatrix}$$

$$W_{out} = \begin{bmatrix} 2 & -2 & -1 & 3 \\ -2 & 1 & -5 & 4 \end{bmatrix} \quad b_{out} = \begin{bmatrix} 12 \\ 3 \end{bmatrix}$$

$$A_{in} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 3 \\ 1 & 1 \\ 4 & 2 \end{bmatrix}$$

$$f_1(x) = f_{out}(x) = relu(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

That is, for simplicity, we will assume that f_1 and f_{out} are the same function.

1. Calculate the following: Z_1 (3 points), A_1 (2 points), Z_{out} (3 points), and A_{out} (2 points). Show your work.

1.1 Calculating Z_1 (3 points)

Solution:

$$Z_1 = W_1 A_{in} + b_1 \tag{5}$$

$$= \begin{bmatrix} 1 & -1 & 2 & 3 & 0 \\ 4 & 0 & -1 & 1 & 3 \\ 2 & 1 & 3 & -5 & -4 \\ 4 & -3 & 2 & 1 & -3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 3 \\ 1 & 1 \\ 4 & 2 \end{bmatrix} + \begin{bmatrix} -1 & -1 \\ 2 & 2 \\ -4 & -4 \\ 3 & 3 \end{bmatrix} \tag{6}$$

$$= \begin{bmatrix} 12 & 6 \\ 16 & 8 \\ 1 & 2 \\ -2 & -7 \end{bmatrix} + \begin{bmatrix} -1 & -1 \\ 2 & 2 \\ -4 & -4 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} 11 & 5 \\ 18 & 10 \\ -3 & -2 \\ 1 & -4 \end{bmatrix}. \tag{7}$$

Where on line (6) we broadcast b_1 to a 4×2 matrix.

1.2 Calculating A_1 (2 points)

Solution:

$$A_1 = f_1(Z_1) = relu(Z_1) \tag{8}$$

$$= \begin{bmatrix} relu(11) & relu(5) \\ relu(18) & relu(11) \\ relu(-3) & relu(-2) \\ relu(1) & relu(-4) \end{bmatrix} = \begin{bmatrix} 11 & 5 \\ 18 & 10 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}. \tag{9}$$

1.3 Calculating Z_{out} (3 points)

Solution:

$$Z_{out} = W_{out} A_1 + b_{out} \tag{10}$$

$$= \begin{bmatrix} 2 & -2 & -1 & 3 \\ -2 & 1 & -5 & 4 \end{bmatrix} \begin{bmatrix} 11 & 5 \\ 18 & 10 \\ 0 & 0 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 12 & 12 \\ 3 & 3 \end{bmatrix} \tag{11}$$

$$= \begin{bmatrix} -11 & -10 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 12 & 12 \\ 3 & 3 \end{bmatrix} \tag{12}$$

$$= \begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix} \tag{13}$$

where we again broadcast b_{out} to a 2×2 matrix on line (11).

1.4 Calculating A_{out} (2 points)

Solution:

$$A_{out} = f_{out}(Z_{out}) = relu(Z_{out}) \tag{14}$$

$$= \begin{bmatrix} relu(1) & relu(2) \\ relu(3) & relu(3) \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix}. \tag{15}$$

2 Backpropagation (19 points)

When the network learns via stochastic gradient descent, each time we see a training example, we make our prediction, calculate the loss with respect to the gold label, and then improve the network's parameters. We calculate the proportion of the loss attributable to each parameter—the gradient of the loss with respect to that parameter—and move that parameter a tiny amount in the opposite direction.

This is described (for a generic network) by:

Parameters: W, b
Inputs: $\hat{x}$
Iterate until convergence (for given learning rate η)
$W \leftarrow W - \eta \frac{\partial Loss}{\partial W}$
$b \leftarrow b - \eta \frac{\partial Loss}{\partial b}$

Notice that stochastic gradient descent updates all of the parameters of the model.

Consider a neural network (Network N) with 4 inputs (x_1, x_2, x_3, x_4) defined below. Note: this is not a dense network but it is still a neural network.

Network N
Inputs: x_1, x_2, x_3, x_4
Hidden units: $x_5 = f_5(x_1), \quad x_6 = f_6(x_2, x_3), \quad x_7 = f_7(x_4)$
Output unit: $x_8 = f_8(x_5, x_6, x_7)$
Given by:
$f_5(x_1) = \sigma(x_1) = \frac{1}{1+\exp(-x_1)}$
$f_6(x_2, x_3) = a * x_2 + b * x_3 + c * x_2 * x_3$
$f_7(x_4) = (x_4)^2 + d$
$f_8(x_5, x_6, x_7) = \frac{\exp(x_6)}{\sum_{i=5}^{7} \exp(x_i)}$
Where:
$a, b, c, d \in \mathbb{R}$ are learned parameters.
End Network

Note: $\exp(x) = e^x$. Suppose that

$$a = 3, \quad b = 4, \quad c = 2, \quad d = 2$$

$$x_1 = 0, \quad x_2 = 2, \quad x_3 = -1, \quad x_4 = 2$$

learning rate

$$\eta = 0.1$$

and

$$\frac{\partial Loss(x_1, x_2, x_3, x_4)}{\partial x_8} = 3.$$

Answer the following:

1. For $i = 1, \ldots, 7$, write the formula to calculate

   $$\frac{\partial Loss}{\partial x_i}$$

   . Show your work. (1 point each)

2. Calculate

   $$\frac{\partial Loss}{\partial a}, \quad \frac{\partial Loss}{\partial b}, \quad \frac{\partial Loss}{\partial c}, \quad \frac{\partial Loss}{\partial d}$$

   and update the learned parameters a, b, c, d. Show your work. (2 points for each partial derivative, 1 point for each update)

2.1 Part 1: x Gradients (7 points)

Solution:

$$\frac{\partial Loss}{\partial x_7} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_7} = 3 * \Big(- \frac{e^{x_6+x_7}}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big) \tag{16}$$

$$\frac{\partial Loss}{\partial x_6} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_6} = 3 * \Big(\frac{e^{x_6}(e^{x_5}+e^{x_7})}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big) \tag{17}$$

$$\frac{\partial Loss}{\partial x_5} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_5} = 3 * \Big(- \frac{e^{x_6+x_5}}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big) \tag{18}$$

$$\frac{\partial Loss}{\partial x_4} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_7}\frac{\partial x_7}{\partial x_4} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_7} * 2x_4 = 3 * \Big(- \frac{e^{x_6+x_7}}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big) * 2x_4 \tag{19}$$

$$\frac{\partial Loss}{\partial x_3} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_6}\frac{\partial x_6}{\partial x_3} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_6} * (b + cx_2) \tag{20}$$

$$= 3 * \Big(\frac{e^{x_6}(e^{x_5}+e^{x_7})}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big) * (b + cx_2) \tag{21}$$

$$\frac{\partial Loss}{\partial x_2} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_6}\frac{\partial x_6}{\partial x_2} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_6} * (a + cx_3) \tag{22}$$

$$= 3 * \Big(\frac{e^{x_6}(e^{x_5}+e^{x_7})}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big) * (a + cx_3) \tag{23}$$

$$\frac{\partial Loss}{\partial x_1} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_5}\frac{\partial x_5}{\partial x_1} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_5} * \frac{e^{-x_5}}{(1+e^{-x_5})^2} \tag{24}$$

$$= 3 * \Big(- \frac{e^{x_6+x_5}}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big)\frac{e^{-x_1}}{(1+e^{-x_1})^2} \tag{25}$$

$$= 3 * \Big(- \frac{e^{x_6+x_5}}{(e^{x_5}+e^{x_6}+e^{x_7})^2}\Big)\sigma(x_1)(1-\sigma(x_1)) \tag{26}$$

where on line (26), $\sigma(x)$ is the sigmoid function, since $f_5(x) = \sigma(x_1)$.

2.2 Part 2: Parameter Gradients (8 points)

Solution:

$$\frac{\partial Loss}{\partial a} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_6}\frac{\partial x_6}{\partial a} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_6} * x_2 = \frac{\partial Loss}{\partial x_6} * x_2 \quad \text{(See 17)} \tag{27}$$

$$\frac{\partial Loss}{\partial b} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_6}\frac{\partial x_6}{\partial b} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_6} * x_3 = \frac{\partial Loss}{\partial x_6} * x_3 \quad \text{(See 17)} \tag{28}$$

$$\frac{\partial Loss}{\partial c} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_6}\frac{\partial x_6}{\partial c} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{x_6} * x_2 * x_3 = \frac{\partial Loss}{\partial x_6} * x_2 * x_3 \quad \text{(See 17)} \tag{29}$$

$$\frac{\partial Loss}{\partial d} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_7}\frac{\partial x_7}{\partial d} = \frac{\partial Loss}{\partial x_8}\frac{\partial x_8}{\partial x_7} * 1 = \frac{\partial Loss}{\partial x_7} * 1 \quad \text{(See 16)} \tag{30}$$

2.3 Part 3: Updates (4 points)

Solution:
Note first that $x_5 = \sigma(x_1) = 0.5$, $x_6 = 3(2)+4(-1)+2(2)(-1) = -2$ and $x_7 = (2)^2+2 = 6$, simply by plugging in the values given for x_1, x_2, x_3, x_4 and a, b, c, d.

$$a \leftarrow a - \eta \frac{\partial Loss}{\partial a} = a - \eta \left(3 * \left(\frac{e^{-2}(e^{0.5} + e^6)}{(e^{0.5} + e^{-2} + e^6)^2} \right) * 2 \right) \quad \text{(See 27)} \tag{31}$$

$$= 3 - 0.1(.0020032) \approx 2.9997997 \tag{32}$$

$$b \leftarrow b - \eta \frac{\partial Loss}{\partial b} = b - \eta \left(3 * \left(\frac{e^{-2}(e^{0.5} + e^6)}{(e^{0.5} + e^{-2} + e^6)^2} \right) * -1 \right) \quad \text{(See 28)} \tag{33}$$

$$= 4 - 0.1(-.0010016) \approx 4.00010016 \tag{34}$$

$$c \leftarrow c - \eta \frac{\partial Loss}{\partial c} = c - \eta \left(3 * \left(\frac{e^{-2}(e^{0.5} + e^6)}{(e^{0.5} + e^{-2} + e^6)^2} \right) * 2 * -1 \right) \quad \text{(See 29)} \tag{35}$$

$$= 2 - 0.1(-.0020032) \approx 2.000200324 \tag{36}$$

$$d \leftarrow d - \eta \frac{\partial Loss}{\partial d} = d - \eta \left(3 * \left(- \frac{e^{x_6+x_7}}{(e^{x_5} + e^{x_6} + e^{x_7})^2} \right) * 1 \right) \quad \text{(See 30)} \tag{37}$$

$$= 2 - 0.1(-.000997546) \approx 2.0000997546 \tag{38}$$