HW1

Yue Zhou

February 2021

1 Problem 1.2 Regression Task

1.1 1.2.a

Step 1 is to predict the outcome by feeding the input to the model. It is to feed forward to get the logits.

```
y_pred = model(X)
```

or

```
output = model(input)
```

In summary, It is the model's forward pass, which takes the input and generates the output.

Step 2 is to compute the loss. We provide prediction y_pred and label y to my criterion to get the loss.

```
loss = criterion(y_pred, y)
```

or

```
J = loss(output, target <or> label)
```

In summary, It takes the model's output and calculates the training loss with respect to the true target or label.

Step 3 is to zero out whatever it has stored before, by zero_grad().The reasoning behind is that we never compute gradients from scratch, we always accumulate the result to whatever we have had before. Since we just want to compute the gradient, we have to zero out whatever it has stored before.

```
optimizer.zero_grad()
```

or

```
model.zero_grad()
```

In summary, It cleans up the gradient calculations, so that they are not accumulated for the next pass.

Step 4 is to compute the backward. It is to compute the partial derivatives of the loss with respect to the parameters.

```
loss.backward()
```

or

```
J.backward()
```

In summary, it does back-propagation and accumulation: It computes $\nabla_x J$ for every variable x for which we have specified. These are accumulated into the gradient of each variable: x.grad ← x.grad + $\nabla_x J$.

Step 5 is stepping. This is going and jumping to the opposite direction of gradient. This is the direction of maximum increment of my function.

```
optimizer.step()
```

In summary, It takes a step in gradient descent: $\vartheta \leftarrow \vartheta - \eta \nabla_\vartheta J$.

1.2 1.2.b

Layer	Input	Output
$Linear_1$	x	$W^{(1)}x + b^{(1)} = z_1$
f	$W^{(1)}x + b^{(1)}$	$max(0, W^{(1)}x_i + b^{(1)})$ for each element x_i in vector
$Linear_2$	$max(0, W^{(1)}x + b^{(1)})$	$W^{(2)}max(0, W^{(1)}x + b^{(1)}) + b^{(2)}$
g	$W^{(2)}max(0, W^{(1)}x + b^{(1)}) + b^{(2)}$	$W^{(2)}max(0, W^{(1)}x + b^{(1)}) + b^{(2)}$
Loss	$W^{(2)}max(0, W^{(1)}x + b^{(1)}) + b^{(2)}$ and y	$\frac{1}{N} \sum_{i=1}^{N}(W^{(2)}max(0, W^{(1)}x_i + b^{(1)}) + b^{(2)} - y_i)^2$
Loss(concise)	$\hat{y}$ and y	$\frac{1}{N} \sum_{i=1}^{N}(\hat{y_i}$ - $y_i)^2$

Table 1: 1.2.b

1.3 1.2.c

Since we have

$\frac{\partial z_2}{\partial z_1} = ReLU^{'}(z_1)$

$\frac{\partial z_3}{\partial z_2} = \frac{\partial(W^{(2)}z_2 + b^{(2)})}{\partial z_2} = W^{(2)}$

$\frac{\partial z_3}{\partial b^{(2)}} = \frac{\partial(W^{(2)}z_2 + b^{(2)})}{\partial b^{(2)}} = 1$

$\frac{\partial z_1}{\partial b^{(1)}} = \frac{\partial(W^{(1)}z_1 + b^{(1)})}{\partial b^{(1)}} = 1$

$\frac{\partial z_3}{\partial W^{(2)}} = \frac{\partial(W^{(2)}z_2 + b^{(2)})}{\partial W^{(2)}} = z_2$

$\frac{\partial z_1}{\partial W^{(1)}} = \frac{\partial(W^{(1)}x + b^{(1)})}{\partial W^{(1)}} = x$

Parameter	Gradient
$W^{(1)}$	$\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial W^{(1)}} = x \ (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1})^T$
$b^{(1)}$	$\frac{\partial l}{\partial b^{(1)}} = (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b^{(1)}})^T = (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1})^T$
$W^{(2)}$	$\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial W^{(2)}} = (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3})^T z_2^T$
$b^{(2)}$	$\frac{\partial l}{\partial b^{(2)}} = (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial b^{(2)}})^T = (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3})^T$

Table 2: 1.2.c

1.4 1.2.d

$$\frac{\partial z_2}{\partial z_1} = \mathrm{ReLU}^{'}(z_1) = (\frac{\partial z_2}{\partial z_1})_{\mathrm{ij}} = \frac{\partial z_{2_i}}{\partial z_{1_j}} = \frac{\partial}{\partial z_{1j}} f(z_1) = \begin{cases} 1 & \text{if } z_{1_j} > 0 \ and \ i = j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\frac{\partial \hat{y}}{\partial z_3} = (\frac{\partial \hat{y}}{\partial z_3})_{ij} = \frac{\partial \hat{y}_i}{\partial z_{3_j}} = \begin{cases} 1 & \text{if i=j} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\frac{\partial l}{\partial \hat{y}} = ||2\hat{y} - 2y|| \tag{3}$$

2 Problem 1.3

2.1 1.3.a

Set function f,g = σ as $\sigma(z) = (1+\exp{(-z)})^{-1}$.What to change from the original network?

Layer	Input	Output
$Linear_1$	x	$W^{(1)}x + b^{(1)}$
f	$W^{(1)}x + b^{(1)}$	$\sigma(W^{(1)}x + b^{(1)})$
$Linear_2$	$\sigma(W^{(1)}x + b^{(1)})$	$W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}$
g	$W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}$	$\sigma(W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)})$
Loss	$\sigma(W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)})$ and y	$\frac{1}{N}\ \sum_{i=1}^{N}\ (\sigma(W^{(2)}\sigma(W^{(1)}x_i + b^{(1)}) + b^{(2)})$ - $y_i)^2$
Loss(concise)	$\hat{y}$ and y	$\frac{1}{N}\ \sum_{i=1}^{N}(\hat{y_i}$ - $y_i)^2$

Table 3: 1.3.a-1

$$\frac{\partial z_2}{\partial z_1} = f'(z_1) = \sigma(z_1)(1 - \sigma(z_1)) \tag{4}$$

$$\frac{\partial \hat{y}}{\partial z_3} = g'(z_3) = \sigma(z_3)(1 - \sigma(z_3)) \tag{5}$$

Parameter	Gradient
$W^{(1)}$	$\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial W^{(1)}} = x \ (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1})^T = x \ (\frac{\partial l}{\partial \hat{y}} \sigma(z_3) \ (1-\sigma(z_3)) W^{(2)} \sigma(z_1) \ (1-\sigma(z_1)))^T$
$b^{(1)}$	$\frac{\partial l}{\partial b^{(1)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b^{(1)}} = (\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1} \ 1 \ )^T = (\frac{\partial l}{\partial \hat{y}} \ \sigma(z_3) \ (1-\sigma(z_3)) W^{(2)} \sigma(z_1) \ (1-\sigma(z_1)))^T$
$W^{(2)}$	$\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial W^{(2)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} z_2 = (\frac{\partial l}{\partial \hat{y}} \ 1 \ g'(z_3))^T \ z_2^T = (\frac{\partial l}{\partial \hat{y}} \ \sigma(z_3) \ (1-\sigma(z_3)))^T \ z_2^T$
$b^{(2)}$	$\frac{\partial l}{\partial b^{(2)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial b^{(2)}} = (\frac{\partial l}{\partial \hat{y}} \ 1 \ g'(z_3) \ 1)^T = (\frac{\partial l}{\partial \hat{y}} \ \sigma'(z_3))^T = (\frac{\partial l}{\partial \hat{y}} \ \sigma(z_3) \ (1-\sigma(z_3)))^T$

Table 4: 1.3.a-2

(p.s. As z_3 is the only input of the function g, hence partial derivative of $\hat{y}$ toward z_3 is in the representation as the derivative of g.)

$$\frac{\partial l}{\partial \hat{y}} = ||2\hat{y} - 2y|| \tag{6}$$

2.2 1.3.b

$l_{\text{BCE}}(\hat{y}, y) = \frac{1}{K} \sum_{i=1}^{K} -[y_i log(\hat{y_i}) + (1-y_i) log(1-\hat{y_i})]$

Layer	Input	Output
$Linear_1$	x	$W^{(1)}x + b^{(1)}$
f	$W^{(1)}x + b^{(1)}$	$\sigma(W^{(1)}x + b^{(1)})$
$Linear_2$	$\sigma(W^{(1)}x + b^{(1)})$	$W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}$
g	$W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}$	$\sigma(W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)})$
Loss	$\sigma(W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)})$ and y	$\frac{1}{K} \sum_{i=1}^{K} -[y_i log(\sigma(W^{(2)}\sigma(W^{(1)}x_i + b^{(1)}) + b^{(2)})) + (1-y_i) log(1-\sigma(W^{(2)}\sigma(W^{(1)}x_i + b^{(1)}) + b^{(2)}))]$
Loss(concise)	$\hat{y}$ and y	$\frac{1}{K} \sum_{i=1}^{K} -[y_i log(\hat{y_i}) + (1-y_i) log(1-\hat{y_i})]$

Table 5: 1.3.b-1

$\frac{\partial l_{\text{BCE}}}{\partial \hat{y}} = -(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}) = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$
$\frac{\partial \hat{y}}{\partial z_3} = \hat{y}(1-\hat{y}) = \sigma(z_3)(1-\sigma(z_3))$
$\frac{\partial z_3}{\partial W^{(2)}} = z_2$
$\frac{\partial l_{\text{BCE}}}{\partial W^{(2)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial W^{(2)}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \ \hat{y}(1-\hat{y}) z_2 = z_2(\hat{y}-y)$
$\frac{\partial l_{\text{BCE}}}{\partial b^{(2)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial b^{(2)}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \ \hat{y}(1-\hat{y}) \ 1 = (\hat{y}-y)$
$\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial W^{(1)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1} x = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \ \hat{y}(1-\hat{y}) W^{(2)} z_2(1-z_2)x = (\hat{y}-y) W^{(2)} z_2(1-z_2)x$
$\frac{\partial l_{\text{BCE}}}{\partial b^{(1)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b^{(1)}} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1} \ 1 = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \ \hat{y}(1-\hat{y}) W^{(2)} z_2(1-z_2)x = (\hat{y}-y) W^{(2)} z_2(1-z_2)$

Parameter	Gradient
$W^{(1)}$	$\frac{\partial l_{\text{BCE}}}{\partial W^{(1)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial W^{(1)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1} x = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \sigma(z_3) \ (1-\sigma(z_3)) W^{(2)} \sigma(z_1) \ (1-\sigma(z_1))x$
$b^{(1)}$	$\frac{\partial l_{\text{BCE}}}{\partial b^{(1)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b^{(1)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1} \ 1 = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \ \sigma(z_3) \ (1-\sigma(z_3)) W^{(2)} \sigma(z_1) \ (1-\sigma(z_1))$
$W^{(2)}$	$\frac{\partial l_{\text{BCE}}}{\partial W^{(2)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial W^{(2)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} z_2 = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \ 1 \ g'(z_3) \ z_2 = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \ \sigma(z_3) \ (1-\sigma(z_3)) \ z_2$
$b^{(2)}$	$\frac{\partial l_{\text{BCE}}}{\partial b^{(2)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial b^{(2)}} = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \ 1 \ g'(z_3) \ 1 = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \ \sigma'(z_3) = \frac{\partial l_{\text{BCE}}}{\partial \hat{y}} \ \sigma(z_3) \ (1-\sigma(z_3))$

Table 6: 1.3.b-2

$$\frac{\partial z_2}{\partial z_1} = f'(z_1) = \sigma(z_1)(1 - \sigma(z_1)) = z_2(1 - z_2) \tag{7}$$

$$\frac{\partial \hat{y}}{\partial z_3} = g'(z_3) = \sigma(z_3)(1 - \sigma(z_3)) = \hat{y}(1 - \hat{y}) \tag{8}$$

$$\frac{\partial l_{\text{BCE}}}{\partial \hat{y}} = -(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}) = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \tag{9}$$

2.3 1.3.c

There are two main reasons why ReLU benefits the deep neural networks:
1.Reduce the likelihood of the gradient to vanish.
2.Generate sparsity.
Explanations:
1. As ReLu function is defined as $f(z) = max(0, z)$ where $z = Wx + b$, when the value of x increases, the output value of ReLU function increases. Hence, the gradient of sigmoid function g becomes smaller as the absolute value of x increases. The constant gradient of ReLU functions contribute on faster learning speed.
2. ReLU neutralizes the effect of sigmoid function w.r.t the sparsity. ReLU returns the value of zero when $z \leq 0$. The more zeros we get from ReLU functions the more sparsity we have. On the other hand, sigmoid functions have the tendency to generate non-zero values, which increase the density in the representation. Hence, the sparsity generated from ReLU functions neutralize the density generated from sigmoid functions.