Deep Learning:
Homework 1: Backpropagation

Due on 02/25/2021

Xinyi Zhao

Problem 1

1.2 Regression Task

(a) **Answer:**

1 Feed data forward to the model to get the logits. Namely, put input $\boldsymbol{x}$ to the model and get the corresponding output $\hat{\boldsymbol{y}}$.

2 Compute the loss using loss function and accuracy of the model. Namely, calculate $\boldsymbol{\ell}_{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y})$ in this example. And calculate the ratio of the number of accurate $\hat{\boldsymbol{y}}$ over the number of all outputs.

3 Zero the gradients before running the backward pass, namely, clean up the gradient calculations, so that they are not accumulated for the next pass.

4 Backward pass to compute the gradient of loss w.r.t our learnable parameters. Namely, do back-propagation to calculate $\frac{\partial \ell}{\partial \boldsymbol{W}}$.

5 Use SGD to update parameters. Namely, $\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta \frac{\partial \ell}{\partial \boldsymbol{W}}$.

(b) **Answer:**

Layer	Input	Output
$Linear_1$	$\boldsymbol{x}$	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$	$(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+$
$Linear_2$	$(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+$	$\boldsymbol{W}^{(2)}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+ + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{W}^{(2)}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+ + \boldsymbol{b}^{(2)}$	$\boldsymbol{W}^{(2)}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+ + \boldsymbol{b}^{(2)}$
Loss	$\boldsymbol{W}^{(2)}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+ + \boldsymbol{b}^{(2)}$	$\left\|\boldsymbol{W}^{(2)}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})^+ + \boldsymbol{b}^{(2)} - \boldsymbol{y}\right\|^2$

Problem 1 continued on next page...

(c) **Answer:**

Parameter	Gradient
$\boldsymbol{W}^{(1)}$	$\boldsymbol{x}\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W}^{(2)}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$
$\boldsymbol{b}^{(1)}$	$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W}^{(2)}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$
$\boldsymbol{W}^{(2)}$	$(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)})^{+}\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$
$\boldsymbol{b}^{(2)}$	$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$

(d) **Answer:**
Assume $\boldsymbol{z_1} \in \mathbb{R}^d$, then $\boldsymbol{z_2} \in \mathbb{R}^d, \boldsymbol{z_3} \in \mathbb{R}^k$.

(1) Since $\boldsymbol{z_2} = (\boldsymbol{z_1})^{+}$, then $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} = \begin{bmatrix} \frac{\partial z2_1}{\partial z1_1} & \frac{\partial z2_1}{\partial z1_2} & \cdots & \frac{\partial z2_1}{\partial z1_d} \\ \frac{\partial z2_2}{\partial z1_1} & \frac{\partial z2_2}{\partial z1_2} & \cdots & \frac{\partial z2_2}{\partial z1_d} \\ \cdots & & & \\ \frac{\partial z2_d}{\partial z1_1} & \frac{\partial z2_d}{\partial z1_2} & \cdots & \frac{\partial z2_d}{\partial z1_d} \end{bmatrix}$.

$$\frac{\partial z_{2i}}{\partial z_{1j}} = \begin{cases} 1, & \text{if i = j and } z_{1j} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

(2) Since $\hat{\boldsymbol{y}} = g(\boldsymbol{z_3})$, then $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial z3_1} & \frac{\partial \hat{y}_1}{\partial z3_2} & \cdots & \frac{\partial \hat{y}_1}{\partial z3_k} \\ \frac{\partial \hat{y}_2}{\partial z3_1} & \frac{\partial \hat{y}_2}{\partial z3_2} & \cdots & \frac{\partial \hat{y}_2}{\partial z3_k} \\ \cdots & & & \\ \frac{\partial \hat{y}_k}{\partial z3_1} & \frac{\partial \hat{y}_k}{\partial z3_2} & \cdots & \frac{\partial \hat{y}_k}{\partial z3_k} \end{bmatrix}$.

$$\frac{\partial \hat{y}_i}{\partial z_{3j}} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Namely, } \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} = \begin{bmatrix} 1 & 0 & ... & 0 & 0 \\ 0 & 1 & ... & 0 & 0 \\ ... & & & & \\ 0 & 0 & ... & 1 & 0 \\ 0 & 0 & ... & 0 & 1 \end{bmatrix}$$

(3) $\frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} = \begin{bmatrix} \frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_1} & \frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_2} & ... & \frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_k} \end{bmatrix}$, where $\frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_i} = 2(\hat{y}_i - y_i)$.

Problem 2

1.3 Classification Task

(a) **Answer:**
The equations of $\boldsymbol{z_2}, \boldsymbol{z_3}, \hat{\boldsymbol{y}}, \ell(\hat{\boldsymbol{y}}, \boldsymbol{y}), \frac{\partial \ell}{\partial \boldsymbol{W}^{(2)}}, \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}, \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$ need to change.

Layer	Input	Output
$Linear_1$	$\boldsymbol{x}$	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
σ	$\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$	$\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$Linear_2$	$\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$	$\boldsymbol{W}^{(2)}\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$
σ	$\boldsymbol{W}^{(2)}\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$	$\sigma(\boldsymbol{W}^{(2)}\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)})$
Loss	$\sigma(\boldsymbol{W}^{(2)}\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)})$	$\left\|\sigma(\boldsymbol{W}^{(2)}\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}) - \boldsymbol{y}\right\|^2$

$$\frac{\partial \ell}{\partial \boldsymbol{W}^{(2)}} = \sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$$

Assume $\boldsymbol{z_1} \in \mathbb{R}^d$, then $\boldsymbol{z_2} \in \mathbb{R}^d, \boldsymbol{z_3} \in \mathbb{R}^k$.

(1) Since $\boldsymbol{z_2} = \sigma(\boldsymbol{z_1})$, then $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} = \begin{bmatrix} \frac{\partial z2_1}{\partial z1_1} & \frac{\partial z2_1}{\partial z1_2} & \cdots & \frac{\partial z2_1}{\partial z1_d} \\ \frac{\partial z2_2}{\partial z1_1} & \frac{\partial z2_2}{\partial z1_2} & \cdots & \frac{\partial z2_2}{\partial z1_d} \\ \cdots & & & \\ \frac{\partial z2_d}{\partial z1_1} & \frac{\partial z2_d}{\partial z1_2} & \cdots & \frac{\partial z2_d}{\partial z1_d} \end{bmatrix}$.

$$\frac{\partial z_{2i}}{\partial z_{1j}} = \begin{cases} \frac{\exp(-z_{1i})}{(1+\exp(-z_{1i}))^2}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

(2) Since $\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z_3})$, then $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial z3_1} & \frac{\partial \hat{y}_1}{\partial z3_2} & \cdots & \frac{\partial \hat{y}_1}{\partial z3_k} \\ \frac{\partial \hat{y}_2}{\partial z3_1} & \frac{\partial \hat{y}_2}{\partial z3_2} & \cdots & \frac{\partial \hat{y}_2}{\partial z3_k} \\ \cdots & & & \\ \frac{\partial \hat{y}_k}{\partial z3_1} & \frac{\partial \hat{y}_k}{\partial z3_2} & \cdots & \frac{\partial \hat{y}_k}{\partial z3_k} \end{bmatrix}$.

$$\frac{\partial \hat{y}_i}{\partial z_{3j}} = \begin{cases} \frac{\exp(-z_{3i})}{(1+\exp(-z_{3i}))^2}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

(b) **Answer:**

The equations of $\ell(\hat{\boldsymbol{y}}, \boldsymbol{y})$, $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$ need to change.

$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} = \begin{bmatrix} \frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_1} & \frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_2} & \cdots & \frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_k} \end{bmatrix}$, where $\frac{\partial \boldsymbol{\ell}}{\partial \hat{y}_i} = -\frac{1}{k}(\frac{y_i}{\hat{y}_i} - \frac{1-y_i}{1-\hat{y}_i})$.

(c) **Answer:**

Because the use of binary activation function always leaves the network susceptible to vanishing gradients. Vanishing gradient prevents weights downstream from being modified by the neural network, which may completely stop the neural network from further training. For example, the logistic sigmoid function $\sigma(s)$ shows that when s is large, $\sigma(s)$ is 1, and when s is small, $\sigma(s)$ is 0. Because the sigmoid function is flat at $\sigma(s) = 1$ and $\sigma(s) = 0$, the gradient is 0, which results in a vanishing gradient. This tends to lose more information during training. Instead, ReLU can be used for intermediate layers to keep more useful information to improve accuracy. So in classification, we can use sigmoid function only on the output of the last layer instead of ReLU. In general, final layers of the networks usually have different kinds of activation functions to accommodate for different types of output, and ReLU is used for intermediate layers. This choice can be beneficial for training a (deeper) network.