DS-GA 1008 Homework Assignment 1

Chengze(Kol) Zuo
cz1565@nyu.edu

February 14, 2019

Part I

1.1

According to the instructions, let $\mathbf{x}$ be the column vector $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, let $\mathbf{y}$ be the column vector $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, and let the weight matrix $\mathbf{W}$ be the 2x2 matrix $\begin{bmatrix} w_{1,1} & w_{1,2} \\ x_{2,1} & w_{2,2} \end{bmatrix}$. Applying the $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$ can give us $\mathbf{y} = \begin{bmatrix} w_{1,1}x_1 + w_{1,2}x_2 + b_1 \\ w_{2,1}x_1 + w_{2,2}x_2 + b_2 \end{bmatrix}$. For the $\frac{\partial L}{\partial \mathbf{W}}$, its shape would be same as the $\mathbf{W}$, and $\frac{\partial L}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial L}{\partial w_{1,1}} & \frac{\partial L}{\partial w_{1,2}} \\ \frac{\partial L}{\partial w_{2,1}} & \frac{\partial L}{\partial w_{2,2}} \end{bmatrix}$. For now, we focus on solving $\frac{\partial L}{\partial w_{1,1}}$, and then apply the same approach to $\frac{\partial L}{\partial w_{1,2}}$, $\frac{\partial L}{\partial w_{2,1}}$, etc.According to the chain rule, we can infer that:

$$\frac{\partial L}{\partial w_{1,1}} = \frac{\partial L}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial w_{1,1}}$$

, where:

$$\frac{\partial L}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial L}{\partial y_1} \\ \frac{\partial L}{\partial y_2} \end{bmatrix} = \begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix}$$

, and $\frac{\partial \mathbf{y}}{\partial w_{1,1}}$ equals to:

$$\frac{\partial \mathbf{y}}{\partial w_{1,1}} = \begin{bmatrix} \frac{\partial y_1}{\partial w_{1,1}} \\ \frac{\partial y_2}{\partial w_{1,1}} \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$$

Therefore, after applying the dot product between two vectors, we can get:

$$\frac{\partial L}{\partial w_{1,1}} = dy_1 x_1$$

Therefore, we can infer that:

$$\frac{\partial L}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial L}{\partial w_{1,1}} & \frac{\partial L}{\partial w_{1,2}} \\ \frac{\partial L}{\partial w_{2,1}} & \frac{\partial L}{\partial w_{2,2}} \end{bmatrix} = \begin{bmatrix} dy_1 x_1 & dy_1 x_2 \\ dy_2 x_1 & dy_2 x_2 \end{bmatrix} = \begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{X}^T$$

Using the same approach, for the $\frac{\partial L}{\partial \mathbf{b}}$, we can infer that:

$$\frac{\partial L}{\partial \mathbf{b}} = \frac{\partial L}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathbf{b}}$$

where:

$$\frac{\partial \mathbf{L}}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial \mathbf{L}}{\partial b_1} \\ \frac{\partial \mathbf{L}}{\partial b_2} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial b_1} = \begin{bmatrix} \frac{\partial y_1}{\partial b_1} \\ \frac{\partial y_2}{\partial b_1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Therefore, $\frac{\partial L}{\partial b_1}$ would be:

$$\frac{\partial L}{\partial b_1} = \begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = dy_1$$

Also, $\frac{\partial L}{\partial b_2}$ would be dy_2. Thus:

$$\frac{\partial L}{\partial \mathbf{b}} = \begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix} = \frac{\partial \mathbf{L}}{\partial \mathbf{y}}$$

(i.e., $\mathbf{I}\frac{\partial \mathbf{L}}{\partial \mathbf{y}}$, the Jacobian matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{b}}$ is an Identity matrix)

1.2

The softmax expression which indicates the probability of the j-th class is as follows:

$$P(z = j|\mathbf{x}) = y_i = \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}$$

To compute the Jacobian matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ looks like:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}$$

For computing any given partial derivative, i.e., the rate of the change for j-th element y_j in the output vector w.r.t i-th element x_i in the input vector, the $\frac{\partial y_j}{\partial x_i}$ would be computed as follow:

$$\frac{\partial y_j}{\partial x_i} = \frac{\partial \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}}{\partial x_i}$$

Given the quotient rule for computing derivatives, suppose:

$$f(x) = \frac{h(x)}{g(x)}$$

we know that:

$$f'(x) = \frac{h'(x)g(x) - g'(x)h(x)}{[g(x)]^2}$$

For now, $h_i = e^{\beta x_i}$, and $g_i = \sum_i e^{\beta x_i}$. For the convenience, we simply use the summation symbol $\sum$ to represent $\sum_i e^{\beta x_i}$. Note that no matter what x_i we compute the derivative of $g(x)$, the result would always be $\beta e^{\beta x_i}$. However for computing the derivative of $e^{\beta x_j}$ w.r.t x_i, if i = j, the result would be $\beta e^{\beta x_i}$, otherwise it would be zero.

Therefore, if i = j:

$$\begin{aligned}\frac{\partial y_j}{\partial x_i} = \frac{\partial \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}}{\partial x_i} &= \frac{\beta e^{\beta x_i} \sum - e^{\beta x_j} \beta e^{\beta x_i}}{\sum^2} \\ &= \frac{\beta e^{\beta x_i}}{\sum} \frac{\sum - e^{\beta x_j}}{\sum} \\ &= \beta y_i (1 - y_j)\end{aligned}$$

or if i $\neq$ j:

$$\begin{aligned}\frac{\partial y_j}{\partial x_i} = \frac{\partial \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}}{\partial x_i} &= \frac{0 - e^{\beta x_j} \beta e^{\beta x_i}}{\sum^2} \\ &= -\frac{e^{\beta x_j}}{\sum} \frac{\beta e^{\beta x_i}}{\sum} \\ &= -y_j \beta y_i \\ &= -\beta y_j y_i\end{aligned}$$

To conclude, the expression for $\frac{\partial y_j}{\partial xi}$ would be:

$$\frac{\partial y_j}{\partial xi} = \begin{cases} \beta y_i (1 - y_j) & \text{i} = j \\ -\beta y_j y_i & \text{i} \neq j \end{cases}$$

Or we can formulate it more concisely:

$$\frac{\partial y_j}{\partial xi} = \beta y_i (1(i = j) - y_j)$$