计算两层神经网络参数迭代的 JACOBI 矩阵：以 MNIST 数据集为例

ZEYU XIE[1], ANGXIU NI[2,3]

1. 神经网络的结构

输入层 - 隐藏层 - 输出层，其中隐藏层的激活函数为 $sigmoid$，输出层的激活函数为 $softmax$

输入层的维度为 784，隐藏层的维度为 50，输出层的维度为 10

2. 变量定义

假设一个 batch 的共有 k 个样本，输入层的输入为 k 行 784 列的矩阵 $X \in \mathbb{R}^{k\times 784}$，输出层的输出为 k 行 10 列的矩阵 $Y \in \mathbb{R}^{k\times 10}$

$$
\begin{aligned}
A_1 &= XW_1 + b_1 \quad k\times 50\\
Z_1 &= sigmoid(Z_1) \quad k\times 50\\
A_2 &= A_1W_2 + b_2 \quad k\times 10\\
Y &= softmax(Z_2) \quad k\times 10
\end{aligned} \tag{1}
$$

记

$$
X = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^k \end{bmatrix} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix} \tag{2}
$$

$$
A_1 = \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix},\ Z_1 = \begin{bmatrix} z_1^1 & z_2^1 & \cdots & z_{50}^1 \\ z_1^2 & z_2^2 & \cdots & z_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ z_1^k & z_2^k & \cdots & z_{50}^k \end{bmatrix} \tag{3}
$$

[1] Department of Mathematics, Tsinghua University, Beijing, China.

[2] Department of Mathematics, University of California, Irvine, USA

[3] Yau Mathematical Sciences Center, Tsinghua University, Beijing, China.

E-mail address: niangxiu@gmail.com.

Date: 2024 年 4 月 5 日.

(4)
$$A_2 = \begin{bmatrix} {a'}_1^1 & {a'}_2^1 & \cdots & {a'}_{10}^1 \\ {a'}_1^2 & {a'}_2^2 & \cdots & {a'}_{10}^2 \\ \vdots & \vdots & \ddots & \vdots \\ {a'}_1^k & {a'}_2^k & \cdots & {a'}_{10}^k \end{bmatrix},\ Y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^k \end{bmatrix} = \begin{bmatrix} y_1^1 & y_2^1 & \cdots & y_{10}^1 \\ y_1^2 & y_2^2 & \cdots & y_{10}^2 \\ \vdots & \vdots & \ddots & \vdots \\ y_1^k & y_2^k & \cdots & y_{10}^k \end{bmatrix}$$

由 1 式可得 X, A_1, Z_1, A_2, Y 之间的关系

(5)
$$A_1 = \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix} \begin{bmatrix} w_1^1 & w_2^1 & \cdots & w_{50}^1 \\ w_1^2 & w_2^2 & \cdots & w_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ w_1^{784} & w_2^{784} & \cdots & w_{50}^{784} \end{bmatrix} + \begin{bmatrix} b_1^1 & b_2^1 & \cdots & b_{50}^1 \\ b_1^2 & b_2^2 & \cdots & b_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1^k & b_2^k & \cdots & b_{50}^k \end{bmatrix}$$

因此 A_1 对 W_1 和 b_1 的 Jacobi 矩阵为

(6)
$$\frac{\partial A_1}{\partial W_1} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix}$$
$$\frac{\partial A_1}{\partial b_1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

因为

(7)
$$Z_1 = sigmoid(A_1)$$
$$\frac{\partial Z_1}{\partial A_1} = sigmoid'(A_1)$$

所以 Z_1 对 W_1 和 b_1 的 Jacobi 矩阵为

$$
\begin{aligned}
\frac{\partial Z_1}{\partial W_1} &= \frac{\partial Z_1}{\partial A_1}\frac{\partial A_1}{\partial W_1} = sigmoid'(A_1)\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix} \\
\frac{\partial Z_1}{\partial b_1} &= \frac{\partial Z_1}{\partial A_1}\frac{\partial A_1}{\partial b_1} = sigmoid'(A_1)\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}
\end{aligned} \tag{8}
$$

其中

$$
sigmoid'(A_1) = sigmoid(A_1) \odot (1 - sigmoid(A_1)) \tag{9}
$$

同样地，A_2 对 W_2 和 b_2 的 Jacobi 矩阵为

$$
\begin{aligned}
\frac{\partial A_2}{\partial W_2} &= \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix} \\
\frac{\partial A_2}{\partial b_2} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}
\end{aligned} \tag{10}
$$

因为

$$
\begin{aligned}
Y &= softmax(A_2) \\
\frac{\partial Y}{\partial A_2} &= softmax'(A_2)
\end{aligned} \tag{11}
$$

所以 Y 对 W_2 和 b_2 的 Jacobi 矩阵为

(12)

$$\frac{\partial Y}{\partial W_2} = \frac{\partial Y}{\partial A_2}\frac{\partial A_2}{\partial W_2} = softmax'(A_2)\begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix}$$

$$\frac{\partial Y}{\partial b_2} = \frac{\partial Y}{\partial A_2}\frac{\partial A_2}{\partial b_2} = softmax'(A_2)\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

其中

(13)
$$softmax'(A_2) = softmax(A_2) \odot (1 - softmax(A_2))$$

Y 对 W_1 和 b_1 的 Jacobi 矩阵为

(14)

$$\frac{\partial Y}{\partial W_1} = \frac{\partial Y}{\partial A_2}\frac{\partial A_2}{\partial Z_1}\frac{\partial Z_1}{\partial A_1}\frac{\partial A_1}{\partial W_1} = softmax'(A_2)W_2^T sigmoid'(A_1)\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix}$$

$$\frac{\partial Y}{\partial b_1} = \frac{\partial Y}{\partial A_2}\frac{\partial A_2}{\partial Z_1}\frac{\partial Z_1}{\partial A_1}\frac{\partial A_1}{\partial b_1} = softmax'(A_2)W_2^T sigmoid'(A_1)\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

3. 总结

(15)
$$\begin{cases} \dfrac{\partial Y}{\partial W_2} = softmax'(A_2)A_1 \\ \dfrac{\partial Y}{\partial b_2} = softmax'(A_2)1 \\ \dfrac{\partial Y}{\partial W_1} = softmax'(A_2)W_2^T sigmoid'(A_1)X \\ \dfrac{\partial Y}{\partial b_1} = softmax'(A_2)W_2^T sigmoid'(A_1)1 \end{cases}$$

4. 迭代

现在考虑对于一组参数 W_1, b_1, W_2, b_2，以及一个 batch 的输入 X，如何计算 Jacobi 矩阵

也即 $Df(W_1,b_1,W_2,b_2)=\begin{bmatrix}\frac{\partial W_1'}{\partial W_1} & \frac{\partial W_1'}{\partial b_1} & \frac{\partial W_1'}{\partial W_2} & \frac{\partial W_1'}{\partial b_2}\\ \frac{\partial b_1'}{\partial W_1} & \frac{\partial b_1'}{\partial b_1} & \frac{\partial b_1'}{\partial W_2} & \frac{\partial b_1'}{\partial b_2}\\ \frac{\partial W_2'}{\partial W_1} & \frac{\partial W_2'}{\partial b_1} & \frac{\partial W_2'}{\partial W_2} & \frac{\partial W_2'}{\partial b_2}\\ \frac{\partial b_2'}{\partial W_1} & \frac{\partial b_2'}{\partial b_1} & \frac{\partial b_2'}{\partial W_2} & \frac{\partial b_2'}{\partial b_2}\end{bmatrix}$

其中，W_1', b_1', W_2', b_2' 是 W_1, b_1, W_2, b_2 在一次梯度下降后的值

取 $\alpha=0.1$ 为学习率，迭代公式为

$$
\begin{aligned}
{W_1}' &= W_1-\alpha\frac{\partial Y}{\partial W_1}\\
{b_1}' &= b_1-\alpha\frac{\partial Y}{\partial b_1}\\
{W_2}' &= W_2-\alpha\frac{\partial Y}{\partial W_2}\\
{b_2}' &= b_2-\alpha\frac{\partial Y}{\partial b_2}
\end{aligned}
\tag{16}
$$

于是

$$
\begin{aligned}
\frac{\partial {W_1}'}{\partial W_1} &= I-\alpha\frac{\partial^2 Y}{\partial W_1\partial W_1}\\
&= I-\alpha\frac{\partial}{\partial W_1}\Big(softmax'(A_2)W_2^T sigmoid'(A_1)X\Big)\\
&= I-\alpha\frac{\partial}{\partial W_1}\Big(softmax'(A_2)W_2^T sigmoid'(A_1)\Big)X\\
&= I-\alpha\left(\frac{\partial softmax'(A_2)}{\partial A_2}\frac{\partial A_2}{\partial Z_1}\frac{\partial Z_1}{\partial A_1}\frac{\partial A_1}{\partial W_1}\right)X\\
&= I-\alpha\left(softmax'(A_2)\frac{\partial A_2}{\partial Z_1}\frac{\partial Z_1}{\partial A_1}\frac{\partial A_1}{\partial W_1}\right)X\\
&= I-\alpha\left(softmax'(A_2)W_2^T sigmoid'(A_1)\frac{\partial A_1}{\partial W_1}\right)X\\
&= I-\alpha\left(softmax'(A_2)W_2^T sigmoid'(A_1)\begin{bmatrix}x_1^1 & x_2^1 & \cdots & x_{784}^1\\ x_1^2 & x_2^2 & \cdots & x_{784}^2\\ \vdots & \vdots & \ddots & \vdots\\ x_1^k & x_2^k & \cdots & x_{784}^k\end{bmatrix}\right)X
\end{aligned}
\tag{17}
$$

$$\frac{\partial {W_1}'}{\partial b_1} = -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right) X \tag{18}$$

$$\begin{aligned} \frac{\partial {W_1}'}{\partial W_2} &= -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix} \right) \\ \frac{\partial {W_1}'}{\partial b_2} &= -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right) \end{aligned} \tag{19}$$

$$\begin{aligned} \frac{\partial {b_1}'}{\partial W_1} &= -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix} \right) \\ \frac{\partial {b_1}'}{\partial b_1} &= -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right) \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial {b_1}'}{\partial W_2} &= -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{784}^1 \\ x_1^2 & x_2^2 & \cdots & x_{784}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{784}^k \end{bmatrix} \right) \\ \frac{\partial {b_1}'}{\partial b_2} &= -\alpha \left(softmax'(A_2) W_2^T sigmoid'(A_1) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right) \end{aligned} \tag{21}$$

$$\frac{\partial {W_2}'}{\partial W_1} = -\alpha \left(softmax'(A_2) \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix} \right)$$

$$\frac{\partial {W_2}'}{\partial b_1} = -\alpha \left(softmax'(A_2) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right) \tag{22}$$

$$\begin{aligned} \frac{\partial {W_2}'}{\partial W_2} &= I - \alpha \frac{\partial^2 Y}{\partial W_2 \partial W_2} \\ &= I - \alpha \frac{\partial}{\partial W_2} \left(softmax'(A_2) A_1 \right) \\ &= I - \alpha \frac{\partial}{\partial W_2} \left(softmax'(A_2) \right) A_1 \\ &= I - \alpha \left(\frac{\partial softmax'(A_2)}{\partial A_2} \frac{\partial A_2}{\partial W_2} \right) A_1 \\ &= I - \alpha \left(softmax'(A_2) \frac{\partial A_2}{\partial W_2} \right) A_1 \\ &= I - \alpha \left(softmax'(A_2) \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix} \right) A_1 \end{aligned} \tag{23}$$

$$\frac{\partial {W_2}'}{\partial b_2} = -\alpha \left(softmax'(A_2) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right) \tag{24}$$

(25)
$$\frac{\partial {b_2}'}{\partial W_1} = -\alpha \left(softmax'(A_2) \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix} \right)$$
$$\frac{\partial {b_2}'}{\partial b_1} = -\alpha \left(softmax'(A_2) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right)$$

(26)
$$\frac{\partial {b_2}'}{\partial W_2} = -\alpha \left(softmax'(A_2) \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_{50}^1 \\ a_1^2 & a_2^2 & \cdots & a_{50}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^k & a_2^k & \cdots & a_{50}^k \end{bmatrix} \right)$$
$$\frac{\partial {b_2}'}{\partial b_2} = -\alpha \left(softmax'(A_2) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \right)$$

将以上 $4 \times 4 = 16$ 个 Jacobi 矩阵计算公式代入 $Df(W_1, b_1, W_2, b_2)$ 的定义，即得到 $Df(W_1, b_1, W_2, b_2)$ 的表达式