

Softmax Regression Math

Ngày 2 tháng 1 năm 2021

Gọi k là số loại label, d là số chiều của dữ liệu, m là số mẫu dữ liệu trong 1 mini-batch ta quy ước Θ , X và Y (dạng one-hot) như sau:

$$\Theta = \begin{bmatrix} w_{01} & w_{02} & \dots & w_{1k} \\ w_{11} & w_{12} & \dots & w_{2k} \\ \dots & \dots & \dots & \dots \\ w_{d1} & w_{d2} & \dots & w_{dk} \end{bmatrix} = [\theta_1 \quad \theta_2 \quad \dots \quad \theta_k] \in \mathbb{R}^{(d+1) \times k}$$

$$X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ \dots & \dots & \dots & \dots \\ x_d^{(1)} & x_d^{(2)} & \dots & x_d^{(m)} \end{bmatrix} = [\mathbf{x}^{(1)} \quad \mathbf{x}^{(2)} \quad \dots \quad \mathbf{x}^{(m)}] \in \mathbb{R}^{(d+1) \times m}$$

$$Y = \begin{bmatrix} y_1^{(1)} & y_1^{(2)} & \dots & y_1^{(m)} \\ y_2^{(1)} & y_2^{(2)} & \dots & y_2^{(m)} \\ \dots & \dots & \dots & \dots \\ y_k^{(1)} & y_k^{(2)} & \dots & y_k^{(m)} \end{bmatrix} = [\mathbf{y}^{(1)} \quad \mathbf{y}^{(2)} \quad \dots \quad \mathbf{y}^{(m)}] \in \mathbb{R}^{k \times m}$$

Từ Θ và X ta tính Z :

$$Z = \Theta^T X = \begin{bmatrix} z_1^{(1)} & z_1^{(2)} & \dots & z_1^{(m)} \\ z_2^{(1)} & z_2^{(2)} & \dots & z_2^{(m)} \\ \dots & \dots & \dots & \dots \\ z_k^{(1)} & z_k^{(2)} & \dots & z_k^{(m)} \end{bmatrix} \in \mathbb{R}^{k \times m}$$

Gọi \mathbf{s} là vector dòng chứa các phần tử là nghịch đảo của tổng các phần tử theo từng cột của Z , ta có:

$$\mathbf{s} = \mathbf{1} \oslash \left(\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} e^Z \right) = \left[\frac{1}{\sum_{j=1}^k e^{z_j^{(1)}}} \quad \frac{1}{\sum_{j=1}^k e^{z_j^{(2)}}} \quad \dots \quad \frac{1}{\sum_{j=1}^k e^{z_j^{(m)}}} \right] \in \mathbb{R}^{1 \times m}$$

\oslash là ký hiệu của Hadamard Division.

Từ đó ta tính được \hat{Y} :

$$\hat{Y} = \mathbf{s} \circ \mathbf{Z} = \text{softmax}(\mathbf{Z}) = \begin{bmatrix} \frac{e^{z_1^{(1)}}}{\sum_{j=1}^k e^{z_j^{(1)}}} & \frac{e^{z_1^{(2)}}}{\sum_{j=1}^k e^{z_j^{(2)}}} & \cdots & \frac{e^{z_1^{(m)}}}{\sum_{j=1}^k e^{z_j^{(m)}}} \\ \frac{e^{z_2^{(1)}}}{\sum_{j=1}^k e^{z_j^{(1)}}} & \frac{e^{z_2^{(2)}}}{\sum_{j=1}^k e^{z_j^{(2)}}} & \cdots & \frac{e^{z_2^{(m)}}}{\sum_{j=1}^k e^{z_j^{(m)}}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{e^{z_k^{(1)}}}{\sum_{j=1}^k e^{z_j^{(1)}}} & \frac{e^{z_k^{(2)}}}{\sum_{j=1}^k e^{z_j^{(2)}}} & \cdots & \frac{e^{z_k^{(m)}}}{\sum_{j=1}^k e^{z_j^{(m)}}} \end{bmatrix} = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \cdots & \hat{y}_1^{(m)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \cdots & \hat{y}_2^{(m)} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_k^{(1)} & \hat{y}_k^{(2)} & \cdots & \hat{y}_k^{(m)} \end{bmatrix} \in \mathbb{R}^{k \times m}$$

Hàm loss $L(\Theta) = -\frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \delta(i, c^{(n)}) \log \hat{y}_i^{(n)}$

Ta quy ước 2 vector \mathbf{k} và \mathbf{c} như sau:

$$\mathbf{k} = \begin{bmatrix} 1 & 2 & \dots & k \end{bmatrix} \in \mathbb{R}^{1 \times k}$$

$$\mathbf{c} = \mathbf{kY} = \begin{bmatrix} c^{(1)} & c^{(2)} & \dots & c^{(m)} \end{bmatrix} \in \mathbb{R}^{1 \times m}$$

Mỗi phần tử trong vector \mathbf{c} biểu thị index của label.

Đạo hàm $L(\Theta)$ theo w_{qj} :

$$\frac{\partial L(\Theta)}{\partial w_{qj}} = \frac{\partial L}{\partial \hat{y}_i^{(n)}} \frac{\partial \hat{y}_i^{(n)}}{\partial z_j^{(n)}} \frac{\partial z_j^{(n)}}{\partial w_{qj}}$$

- Tính $\frac{\partial L}{\partial \hat{y}_i^{(n)}}$:

$$\frac{\partial L}{\partial \hat{y}_i^{(n)}} = -\frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \frac{\delta(i, c^{(n)})}{\hat{y}_i^{(n)}}$$

- Tính $\frac{\partial \hat{y}_i^{(n)}}{\partial z_j^{(n)}}$:

$$\hat{y}_i^{(n)} = \frac{e^{z_i^{(n)}}}{\sum_{j=1}^k e^{z_j^{(n)}}}$$

$$\frac{\partial \hat{y}_i^{(n)}}{\partial z_j^{(n)}} = \frac{\delta(i, j) e^{z_i^{(n)}} \sum_{j=1}^k e^{z_j^{(n)}} - e^{z_i^{(n)}} e^{z_j^{(n)}}}{\left(\sum_{j=1}^k e^{z_j^{(n)}} \right)^2}$$

$$= \delta(i, j) \hat{y}_i^{(n)} - \hat{y}_i^{(n)} \hat{y}_j^{(n)}$$

$$= \hat{y}_i^{(n)} (\delta(i, j) - \hat{y}_j^{(n)})$$

- Tính $\frac{\partial z_j^{(n)}}{\partial w_{qj}}$:

$$\frac{\partial z_j^{(n)}}{\partial w_{qj}} = x_q^{(n)}$$

- Tính $\frac{\partial L(\Theta)}{\partial w_{qj}}$:

$$\begin{aligned}
\frac{\partial L(\Theta)}{\partial w_{qj}} &= -\frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \frac{\delta(i, c^{(n)})}{\hat{y}_i^{(n)}} \hat{y}_i^{(n)} (\delta(i, j) - \hat{y}_j^{(n)}) x_q^{(n)} \\
&= \frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \delta(i, c^{(n)}) (\hat{y}_j^{(n)} - \delta(i, j)) x_q^{(n)} \\
&= \frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \delta(i, c^{(n)}) \hat{y}_j^{(n)} x_q^{(n)} - \frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \delta(i, c^{(n)}) \delta(i, j) x_q^{(n)} \\
&= \frac{1}{m} \sum_{n=1}^m \hat{y}_j^{(n)} x_q^{(n)} - \frac{1}{m} \sum_{n=1}^m \delta(j, c^{(n)}) x_q^{(n)} \\
&= \frac{1}{m} \sum_{n=1}^m (\hat{y}_j^{(n)} - \delta(j, c^{(n)})) x_q^{(n)}
\end{aligned}$$

Gradient của $L(\Theta)$ theo Θ :

$$\begin{aligned}
\nabla_{\Theta} L(\Theta) &= \frac{\partial L(\Theta)}{\partial \Theta} \\
&= \frac{1}{m} \begin{bmatrix} \frac{\partial L}{\partial w_{01}} & \frac{\partial L}{\partial w_{02}} & \dots & \frac{\partial L}{\partial w_{0k}} \\ \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} & \dots & \frac{\partial L}{\partial w_{1k}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial L}{\partial w_{d1}} & \frac{\partial L}{\partial w_{d2}} & \dots & \frac{\partial L}{\partial w_{dk}} \end{bmatrix} \in \mathbb{R}^{(d+1) \times k} \\
&= \frac{1}{m} \begin{bmatrix} \sum_{n=1}^m (\hat{y}_1^{(n)} - \delta(1, c^{(n)})) x_0^{(n)} & \sum_{n=1}^m (\hat{y}_2^{(n)} - \delta(2, c^{(n)})) x_0^{(n)} & \dots & \sum_{n=1}^m (\hat{y}_k^{(n)} - \delta(k, c^{(k)})) x_0^{(n)} \\ \sum_{n=1}^m (\hat{y}_1^{(n)} - \delta(1, c^{(n)})) x_1^{(n)} & \sum_{n=1}^m (\hat{y}_2^{(n)} - \delta(2, c^{(n)})) x_1^{(n)} & \dots & \sum_{n=1}^m (\hat{y}_k^{(n)} - \delta(k, c^{(k)})) x_1^{(n)} \\ \dots & \dots & \dots & \dots \\ \sum_{n=1}^m (\hat{y}_1^{(n)} - \delta(1, c^{(n)})) x_d^{(n)} & \sum_{n=1}^m (\hat{y}_2^{(n)} - \delta(2, c^{(n)})) x_d^{(n)} & \dots & \sum_{n=1}^m (\hat{y}_k^{(n)} - \delta(k, c^{(k)})) x_d^{(n)} \end{bmatrix} \\
&= \frac{1}{m} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ \dots & \dots & \dots & \dots \\ x_d^{(1)} & x_d^{(2)} & \dots & x_d^{(m)} \end{bmatrix} \begin{bmatrix} \hat{y}_1^{(1)} - \delta(1, c^{(1)}) & \hat{y}_1^{(2)} - \delta(1, c^{(2)}) & \dots & \hat{y}_1^{(m)} - \delta(1, c^{(m)}) \\ \hat{y}_2^{(1)} - \delta(2, c^{(1)}) & \hat{y}_2^{(2)} - \delta(2, c^{(2)}) & \dots & \hat{y}_2^{(m)} - \delta(2, c^{(m)}) \\ \dots & \dots & \dots & \dots \\ \hat{y}_k^{(1)} - \delta(k, c^{(1)}) & \hat{y}_k^{(2)} - \delta(k, c^{(2)}) & \dots & \hat{y}_k^{(m)} - \delta(k, c^{(m)}) \end{bmatrix}^T \\
&= \frac{1}{m} \mathbf{X} \mathbf{E}^T
\end{aligned}$$

Với \mathbf{E} là ma trận như sau:

$$\mathbf{E} = \begin{bmatrix} \hat{y}_1^{(1)} - \delta(1, c^{(1)}) & \hat{y}_1^{(2)} - \delta(1, c^{(2)}) & \dots & \hat{y}_1^{(m)} - \delta(1, c^{(m)}) \\ \hat{y}_2^{(1)} - \delta(2, c^{(1)}) & \hat{y}_2^{(2)} - \delta(2, c^{(2)}) & \dots & \hat{y}_2^{(m)} - \delta(2, c^{(m)}) \\ \dots & \dots & \dots & \dots \\ \hat{y}_k^{(1)} - \delta(k, c^{(1)}) & \hat{y}_k^{(2)} - \delta(k, c^{(2)}) & \dots & \hat{y}_k^{(m)} - \delta(k, c^{(m)}) \end{bmatrix} \in \mathbb{R}^{k \times m}$$

Ở bước cuối cùng ta chỉ cần cập nhật Θ với tốc độ học η :

$$\begin{aligned}
\Theta &= \Theta - \eta \nabla_{\Theta} L(\Theta) \\
&= \Theta - \frac{\eta}{m} \mathbf{X} \mathbf{E}^T
\end{aligned}$$

Hàm loss $L(\Theta) = -\frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k y_i^{(n)} \log \hat{y}_i^{(n)}$

Đạo hàm $L(\Theta)$ theo w_{qj} :

$$\frac{\partial L(\Theta)}{\partial w_{qj}} = \frac{\partial L}{\partial \hat{y}_i^{(n)}} \frac{\partial \hat{y}_i^{(n)}}{\partial z_j^{(n)}} \frac{\partial z_j^{(n)}}{\partial w_{qj}}$$

- Tính $\frac{\partial L}{\partial \hat{y}_i^{(n)}}$:

$$\frac{\partial L}{\partial \hat{y}_i^{(n)}} = -\frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \frac{y_i^{(n)}}{\hat{y}_i^{(n)}}$$

- Tính $\frac{\partial \hat{y}_i^{(n)}}{\partial z_j^{(n)}}$:

$$\frac{\partial \hat{y}_i^{(n)}}{\partial z_j^{(n)}} = \hat{y}_i^{(n)} (\delta(i, j) - \hat{y}_j^{(n)})$$

- Tính $\frac{\partial z_j^{(n)}}{\partial w_{qj}}$:

$$\frac{\partial z_j^{(n)}}{\partial w_{qj}} = x_q^{(n)}$$

- Tính $\frac{\partial L(\Theta)}{\partial w_{qj}}$:

$$\begin{aligned} \frac{\partial L(\Theta)}{\partial w_{qj}} &= -\frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k \frac{y_i^{(n)}}{\hat{y}_i^{(n)}} \hat{y}_i^{(n)} (\delta(i, j) - \hat{y}_j^{(n)}) x_q^{(n)} \\ &= \frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k y_i^{(n)} (\hat{y}_j^{(n)} - \delta(i, j)) x_q^{(n)} \\ &= \frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k y_i^{(n)} \hat{y}_j^{(n)} x_q^{(n)} - \frac{1}{m} \sum_{n=1}^m \sum_{i=1}^k y_i^{(n)} \delta(i, j) x_q^{(n)} \\ &= \frac{1}{m} \sum_{n=1}^m \hat{y}_j^{(n)} x_q^{(n)} - \frac{1}{m} \sum_{n=1}^m y_j^{(n)} x_q^{(n)} \\ &= \frac{1}{m} \sum_{n=1}^m (\hat{y}_j^{(n)} - y_j^{(n)}) x_q^{(n)} \end{aligned}$$

Gradient của $L(\Theta)$ theo Θ :

$$\begin{aligned}
\nabla_{\Theta} L(\Theta) &= \frac{\partial L(\Theta)}{\partial \Theta} \\
&= \frac{1}{m} \begin{bmatrix} \frac{\partial L}{\partial w_{01}} & \frac{\partial L}{\partial w_{02}} & \cdots & \frac{\partial L}{\partial w_{0k}} \\ \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} & \cdots & \frac{\partial L}{\partial w_{1k}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial L}{\partial w_{d1}} & \frac{\partial L}{\partial w_{d2}} & \cdots & \frac{\partial L}{\partial w_{dk}} \end{bmatrix} \in \mathbb{R}^{(d+1) \times k} \\
&= \frac{1}{m} \begin{bmatrix} \sum_{n=1}^m (\hat{y}_1^{(n)} - y_1^{(n)}) x_0^{(n)} & \sum_{n=1}^m (\hat{y}_2^{(n)} - y_2^{(n)}) x_0^{(n)} & \cdots & \sum_{n=1}^m (\hat{y}_k^{(n)} - y_k^{(n)}) x_0^{(n)} \\ \sum_{n=1}^m (\hat{y}_1^{(n)} - y_1^{(n)}) x_1^{(n)} & \sum_{n=1}^m (\hat{y}_2^{(n)} - y_2^{(n)}) x_1^{(n)} & \cdots & \sum_{n=1}^m (\hat{y}_k^{(n)} - y_k^{(n)}) x_1^{(n)} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{n=1}^m (\hat{y}_1^{(n)} - y_1^{(n)}) x_d^{(n)} & \sum_{n=1}^m (\hat{y}_2^{(n)} - y_2^{(n)}) x_d^{(n)} & \cdots & \sum_{n=1}^m (\hat{y}_k^{(n)} - y_k^{(n)}) x_d^{(n)} \end{bmatrix} \\
&= \frac{1}{m} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ \cdots & \cdots & \cdots & \cdots \\ x_d^{(1)} & x_d^{(2)} & \cdots & x_d^{(m)} \end{bmatrix} \begin{bmatrix} \hat{y}_1^{(1)} - y_1^{(1)} & \hat{y}_1^{(2)} - y_1^{(2)} & \cdots & \hat{y}_1^{(m)} - y_1^{(m)} \\ \hat{y}_2^{(1)} - y_2^{(1)} & \hat{y}_2^{(2)} - y_2^{(2)} & \cdots & \hat{y}_2^{(m)} - y_2^{(m)} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_k^{(1)} - y_k^{(1)} & \hat{y}_k^{(2)} - y_k^{(2)} & \cdots & \hat{y}_k^{(m)} - y_k^{(m)} \end{bmatrix}^T \\
&= \frac{1}{m} \mathbf{X} \mathbf{E}^T
\end{aligned}$$

Với \mathbf{E} là ma trận như sau:

$$\mathbf{E} = \begin{bmatrix} \hat{y}_1^{(1)} - y_1^{(1)} & \hat{y}_1^{(2)} - y_1^{(2)} & \cdots & \hat{y}_1^{(m)} - y_1^{(m)} \\ \hat{y}_2^{(1)} - y_2^{(1)} & \hat{y}_2^{(2)} - y_2^{(2)} & \cdots & \hat{y}_2^{(m)} - y_2^{(m)} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_k^{(1)} - y_k^{(1)} & \hat{y}_k^{(2)} - y_k^{(2)} & \cdots & \hat{y}_k^{(m)} - y_k^{(m)} \end{bmatrix} \in \mathbb{R}^{k \times m}$$

Ở bước cuối cùng ta chỉ cần cập nhật Θ với tốc độ học η :

$$\begin{aligned}
\Theta &= \Theta - \eta \nabla_{\Theta} L(\Theta) \\
&= \Theta - \frac{\eta}{m} \mathbf{X} \mathbf{E}^T
\end{aligned}$$

Kết luận:

Ma trận \mathbf{E} ở cả hai cách trên thực chất là giống nhau chỉ khác nhau về cách kí hiệu. Thông qua hai cách kí hiệu trên ta dễ dàng thấy kí hiệu ở cách thứ hai rất phù hợp với đầu ra là một one-hot vector. Còn kí hiệu ở cách thứ nhất phù hợp cho việc đầu ra là một số nguyên (phần tử của vector \mathbf{c}). Dù là kí hiệu như thế nào thì kết quả cập nhật Θ cuối cùng đều giống nhau.