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 $\mathbf{\Theta}$, $\boldsymbol{X}$ và $\boldsymbol{Y}$ (dạng one-hot) như sau:

$$\mathbf{\Theta} = \begin{bmatrix} w_{01} & w_{02} & ... & w_{0k} \\ w_{11} & w_{12} & ... & w_{1k} \\ ... & ... & ... & ... \\ w_{d1} & w_{d2} & ... & w_{dk} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & ... & \boldsymbol{\theta}_k \end{bmatrix} \in \mathbb{R}^{(d+1)\times k}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & ... & 1 \\ x_1^{(1)} & x_1^{(2)} & ... & x_1^{(m)} \\ ... & ... & ... & ... \\ x_d^{(1)} & x_d^{(2)} & ... & x_d^{(m)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & ... & \boldsymbol{x}^{(m)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times m}$$

$$\boldsymbol{Y} = \begin{bmatrix} y_1^{(1)} & y_1^{(2)} & ... & y_1^{(m)} \\ y_2^{(1)} & y_2^{(2)} & ... & y_2^{(m)} \\ ... & ... & ... & ... \\ y_k^{(1)} & y_k^{(2)} & ... & y_k^{(m)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{(1)} & \boldsymbol{y}^{(2)} & ... & \boldsymbol{y}^{(m)} \end{bmatrix} \in \mathbb{R}^{k\times m}$$

Từ $\mathbf{\Theta}$ và $\boldsymbol{X}$ ta tính $\boldsymbol{Z}$:

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

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

$$\boldsymbol{s} = 1 \oslash \left(\begin{bmatrix} 1 & 1 & ... & 1 \end{bmatrix} e^{\boldsymbol{Z}}\right) = \begin{bmatrix} \frac{1}{\sum_{j=1}^k e^{z_j^{(1)}}} & \frac{1}{\sum_{j=1}^k e^{z_j^{(2)}}} & \cdots & \frac{1}{\sum_{j=1}^k e^{z_j^{(m)}}} \end{bmatrix} \in \mathbb{R}^{1\times m}$$

$\oslash$ là ký hiệu của Hadamard Division.

Từ đó ta tính được $\hat{Y}$:

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

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

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

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

Mỗi phần tử trong vector $\boldsymbol{c}$ biểu thị index của label.
Đạo hàm $L(\boldsymbol{\Theta})$ theo w_{qj}:

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

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

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

Gradient của $L(\mathbf{\Theta})$ theo $\mathbf{\Theta}$:

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

Với $\boldsymbol{E}$ là ma trận như sau:

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

Ở bước cuối cùng ta chỉ cần cập nhật $\mathbf{\Theta}$ với tốc độ học η:

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

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

Đạo hàm $L(\mathbf{\Theta})$ theo w_{qj}:

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

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

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

Gradient của $L(\mathbf{\Theta})$ theo $\mathbf{\Theta}$:

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

Với $\boldsymbol{E}$ là ma trận như sau:

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

Ở bước cuối cùng ta chỉ cần cập nhật $\mathbf{\Theta}$ với tốc độ học η:

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

Kết luận:

Ma trận $\boldsymbol{E}$ 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 $\boldsymbol{c}$). Dù là kí hiệu như thế nào thì kết quả cập nhật $\mathbf{\Theta}$ cuối cùng đều giống nhau.