# Exercise 5 - Softmax regression

Đặng Linh Anh

January 2021

**1.** Xây dựng công thức tính forward và backward (tính đạo hàm cho từng tham số) cho bài toán softmax regression dùng phương pháp dựa trên hàm delta.

(a) Stochastic gradient descent

(b) Batch gradient descent

**Lời giải**

Cho dataset có: $n$ features, $N$ samples, $k$ outputs

$$\mathbf{\Theta} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{k1} \\ w_{12} & w_{22} & \cdots & w_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{kn} \\ b_1 & b_2 & \cdots & b_k \end{bmatrix} \tag{1}$$

(a) Stochastic gradient descent:
Cài đặt:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{bmatrix} \tag{2}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_k \end{bmatrix} \tag{3}$$

$$\Delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_k \end{bmatrix} \tag{4}$$

$$\delta_i = \begin{cases} 1, & \text{nếu } y = i \\ 0, & \text{khác} \end{cases} \tag{5}$$

i. Tính toán forward

$$\mathbf{z} = \mathbf{\Theta}^T \mathbf{x} \tag{6}$$

$$\hat{\mathbf{y}} = \frac{e^{\mathbf{z}}}{\sum e^{\mathbf{z}}} \tag{7}$$

Loss function:

$$
\begin{aligned}
L &= \sum_{i=1}^{k} \delta_i \log \hat{y} && (8)\\
\Rightarrow L &= \sum (\Delta \odot \log \hat{\mathbf{y}}) && (9)
\end{aligned}
$$

ii. Tính toán backward

- Gradient cho từng tham số:

$$
\frac{\partial L}{\partial \Theta_{ij}} = \sum_{u=1}^{k} \frac{\partial L}{\partial \hat{y}_u} \times \sum_{v=1}^{k} \frac{\partial \hat{y}_v}{\partial z_i} \times \frac{\partial z_i}{\partial \Theta_{ij}} \tag{10}
$$

Trong đó:

$$
\begin{aligned}
\frac{\partial z_i}{\partial \Theta_{ij}} &= x_j, && (11)\\
\sum_{v=1}^{k} \frac{\partial \hat{y}_v}{\partial z_i} &= \sum_{v=1}^{k} \frac{\partial}{\partial z_i} \left( \frac{e^{z_v}}{\sum_{j=1}^{k} e^{z_j}} \right) && (12)\\
\Leftrightarrow \sum_{v=1}^{k} \frac{\partial \hat{y}_v}{\partial z_i} &= \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} - \sum_{v=1}^{k} \frac{e^{z_i} e^{z_v}}{\left(\sum_{j=1}^{k} e^{z_j}\right)^2} && (13)\\
\Leftrightarrow \sum_{v=1}^{k} \frac{\partial \hat{y}_v}{\partial z_i} &= \sum_{v=1}^{k} \hat{y}_v (\delta_v - \hat{y}_i), && (14)\\
\sum_{u=1}^{k} \frac{\partial L}{\partial \hat{y}_u} &= -\sum_{u=1}^{k} \frac{\delta_u}{\hat{y}_u} && (15)
\end{aligned}
$$

Từ đó, ta tìm được gradient:

$$
\begin{aligned}
\frac{\partial L}{\partial \Theta_{ij}} &= -\sum_{u=1}^{k} \frac{\delta_u}{\hat{y}_u} \times \sum_{v=1}^{k} \hat{y}_v (\delta_i - \hat{y}_i) \times x_j && (16)\\
\Leftrightarrow \frac{\partial L}{\partial \Theta_{ij}} &= x_j (\hat{y}_i - \delta_i) && (17)
\end{aligned}
$$

Vectorization:

$$
\frac{\partial L}{\partial \Theta} = \mathbf{x}^T (\hat{\mathbf{y}} - \Delta) \tag{18}
$$

- Cập nhật tham số:

$$
\boldsymbol{\Theta} := \boldsymbol{\Theta} - \eta \times \mathbf{x}^T (\hat{\mathbf{y}} - \Delta) \tag{19}
$$

(b) Batch gradient descent

Cài đặt:

$$
\begin{aligned}
\mathbf{X} &= \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \\ 1 & 1 & \cdots & 1 \end{bmatrix} && (20) \\
\mathbf{Y} &= \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(m)} \end{bmatrix} && (21) \\
\hat{\mathbf{Y}} &= \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \cdots & \hat{y}_1^{(m)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \cdots & \hat{y}_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_k^{(1)} & \hat{y}_k^{(2)} & \cdots & \hat{y}_k^{(m)} \end{bmatrix} && (22) \\
\Delta &= \begin{bmatrix} \delta_1^{(1)} & \delta_1^{(2)} & \cdots & \delta_1^{(m)} \\ \delta_2^{(1)} & \delta_2^{(2)} & \cdots & \delta_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_k^{(1)} & \delta_k^{(2)} & \cdots & \delta_k^{(m)} \end{bmatrix} && (23) \\
\delta_i^{(u)} &= \begin{cases} 1, & \text{nếu } y^{(u)} = i \\ 0, & \text{khác} \end{cases} && (24)
\end{aligned}
$$

i. Tính toán forward

$$
\begin{aligned}
\mathbf{Z} &= \mathbf{\Theta}^T \mathbf{x} && (25) \\
\hat{y}^{(u)} &= \frac{e^{\mathbf{z}^{(u)}}}{\sum e^{\mathbf{z}^{(u)}}} && (26) \\
\Rightarrow \hat{\mathbf{Y}} &= \begin{bmatrix} \hat{\mathbf{y}}^{(1)} & \hat{\mathbf{y}}^{(2)} & \cdots & \hat{\mathbf{y}}^{(u)} \cdots \hat{\mathbf{y}}^{(m)} \end{bmatrix} && (27)
\end{aligned}
$$

Loss function:

$$
\begin{aligned}
L &= \frac{1}{m} \sum_{u=1}^{m} \sum_{i=1}^{k} \delta_i^{(u)} \log \hat{y}_i^{(u)} && (28) \\
\Rightarrow L &= \frac{1}{m} \underbrace{\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}}_{k} (\Delta \odot \log \hat{\mathbf{y}}) \underbrace{\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T}_{m} && (29)
\end{aligned}
$$

ii. Tính toán backward

- Gradient cho từng tham số tính trên sample thứ $w$:

$$
\frac{\partial L^{(w)}}{\partial \Theta_{ij}} = \sum_{u=1}^{k} \frac{\partial L^{(w)}}{\partial \hat{y}_u^{(w)}} \times \sum_{v=1}^{k} \frac{\partial \hat{y}_v^{(w)}}{\partial z_i^{(w)}} \times \frac{\partial z_i^{(w)}}{d\Theta_{ij}} \quad (30)
$$

Trong đó:

$$
\begin{aligned}
\frac{\partial z_i^{(w)}}{\partial \Theta_{ij}} &= x_j^{(w)}, && (31)\\
\sum_{v=1}^{k} \frac{\partial \hat{y}_v^{(w)}}{\partial z_i^{(w)}} &= \sum_{v=1}^{k} \frac{\partial}{\partial z_i^{(w)}} \left( \frac{e^{z_v^{(w)}}}{\sum_{j=1}^{k} e^{z_j^{(w)}}} \right) && (32)\\
\Leftrightarrow \sum_{v=1}^{k} \frac{\partial \hat{y}_v^{(w)}}{\partial z_i^{(w)}} &= \frac{e^{z_i^{(w)}}}{\sum_{j=1}^{k} e^{z_j^{(w)}}} - \sum_{v=1}^{k} \frac{e^{z_i^{(w)}} e^{z_v^{(w)}}}{\left(\sum_{j=1}^{k} e^{z_j^{(w)}}\right)^2} && (33)\\
\Leftrightarrow \sum_{v=1}^{k} \frac{\partial \hat{y}_u^{(w)}}{\partial z_i^{(w)}} &= \sum_{v=1}^{k} \hat{y}_v^{(w)} (\delta_v^{(w)} - \hat{y}_i^{(w)}), && (34)\\
\sum_{u=1}^{k} \frac{\partial L^{(w)}}{\partial \hat{y}_u^{(w)}} &= -\sum_{u=1}^{k} \frac{\delta_u^{(w)}}{\hat{y}_u^{(w)}} && (35)
\end{aligned}
$$

Do đó:

$$
\begin{aligned}
\frac{\partial L^{(w)}}{\partial \Theta_{ij}} &= -\sum_{u=1}^{k} \frac{\delta_u^{(w)}}{\hat{y}_u^{(w)}} \times \sum_{v=1}^{k} \hat{y}_v^{(w)} (\delta_v^{(w)} - \hat{y}_i^{(w)}) \times x_j^{(w)} && (36)\\
\Leftrightarrow \frac{\partial L^{(w)}}{\partial \Theta_{ij}} &= x_j^{(w)} (\hat{y}_i^{(w)} - \delta_i^{(w)}), && (37)\\
\frac{\partial L}{\partial \Theta_{ij}} &= \sum_{w=1}^{m} \frac{\partial L^{(w)}}{\partial \Theta_{ij}} && (38)\\
\Leftrightarrow \frac{\partial L}{\partial \Theta_{ij}} &= \frac{1}{m} \sum_{w=1}^{m} x_j^{(w)} (\hat{y}_i^{(w)} - \delta_i^{(w)}) && (39)
\end{aligned}
$$

Vectorization:

$$\frac{\partial L}{\partial \Theta} = \frac{1}{m} \mathbf{X} (\hat{\mathbf{Y}} - \Delta)^T \tag{40}$$

- Cập nhật tham số:

$$\boldsymbol{\Theta} := \boldsymbol{\Theta} - \eta \times \frac{1}{m} \mathbf{X} (\hat{\mathbf{Y}} - \Delta)^T \tag{41}$$

**2.** Xây dựng công thức tính forward và backward (tính đạo hàm cho từng tham số) cho bài toán softmax regression dùng phương pháp dựa trên one-hot encoding.

(a) Stochastic gradient descent

(b) Batch gradient descent

**Lời giải**

(a) Stochastic gradient descent

Cài đặt:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{bmatrix} \tag{42}$$

$$\mathbf{y} = \begin{cases} y^{(1)}, \\ y^{(2)}, \\ \cdots, \\ y^{(k)} \end{cases} \tag{43}$$

Với:

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(k)} \end{bmatrix} = \mathbf{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_k \end{bmatrix} \tag{44}$$

i. Tính toán forward

$$\mathbf{z} = \mathbf{\Theta}^T \mathbf{x} \tag{45}$$

$$\hat{\mathbf{y}} = \frac{e^{\mathbf{z}}}{\sum e^{\mathbf{z}}} \tag{46}$$

Loss function:

$$L = -\mathbf{y} \log \hat{\mathbf{y}} \tag{47}$$

ii. Tính toán backward

- Gradient cho từng tham số:

$$\frac{\partial L}{\partial \Theta_{ij}} = \sum_{u=1}^{k} \frac{\partial L}{\partial \hat{y}_u} \times \sum_{v=1}^{k} \frac{\partial \hat{y}_v}{\partial z_i} \times \frac{\partial z_i}{\partial \Theta_{ij}} \tag{48}$$

Trong đó:

$$\frac{\partial z_i}{\partial \Theta_{ij}} = x_j, \tag{49}$$

$$\sum_{v=1}^{k} \frac{\partial \hat{y}_v}{\partial z_i} = \sum_{v=1}^{k} \hat{y}_v (y_v - \hat{y}_i), \tag{50}$$

$$\sum_{u=1}^{k} \frac{\partial L}{\partial \hat{y}_u} = -\sum (\mathbf{y}^T \oslash \hat{\mathbf{y}}), \quad \forall \hat{y} \in \hat{\mathbf{y}} \neq 0 \tag{51}$$

Từ đó, ta tìm được gradient:

$$\begin{aligned}\frac{\partial L}{\partial \Theta_{ij}} &= -\sum(\mathbf{y}^T \oslash \hat{\mathbf{y}}) \times \sum_{v=1}^{k} \hat{y}_v(y_v - \hat{y}_i) \times x_j & (52)\\ \Leftrightarrow \frac{\partial L}{\partial \Theta_{ij}} &= x_j(\hat{y}_i - y_i) & (53)\end{aligned}$$

Vectorization:

$$\frac{\partial L}{\partial \Theta} = \mathbf{x}(\hat{\mathbf{y}} - \mathbf{y}^T) \quad (54)$$

- Cập nhật tham số:

$$\boldsymbol{\Theta} := \boldsymbol{\Theta} - \eta \times \mathbf{x}(\hat{\mathbf{y}} - \mathbf{y}^T) \quad (55)$$

(b) Batch gradient descent
Cài đặt:

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad (56)$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \cdots \\ \mathbf{y}^{(m)} \end{bmatrix} \quad (57)$$

$$(58)$$

Với $u \in [1, m]$:

$$\mathbf{y}^{(u)} = \begin{cases} \mathbf{y}^{[1]}, \\ \mathbf{y}^{[2]}, \\ \vdots \\ \mathbf{y}^{[k]} \end{cases}$$

$$\begin{bmatrix} \mathbf{y}^{[1]} \\ \mathbf{y}^{[2]} \\ \vdots \\ \mathbf{y}^{[k]} \end{bmatrix} = \mathbf{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \cdots & \hat{y}_1^{(m)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \cdots & \hat{y}_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_k^{(1)} & \hat{y}_k^{(2)} & \cdots & \hat{y}_k^{(m)} \end{bmatrix} \quad (59)$$

i. Tính toán forward

$$\begin{aligned}\mathbf{Z} &= \boldsymbol{\Theta}^T \mathbf{x} & (60)\\ \hat{\mathbf{y}}^{(u)} &= \frac{e^{\mathbf{z}^{(u)}}}{\sum e^{\mathbf{z}^{(u)}}} & (61)\\ \Rightarrow \hat{\mathbf{Y}} &= \begin{bmatrix} \hat{\mathbf{y}}^{(1)} & \hat{\mathbf{y}}^{(2)} & \cdots & \hat{\mathbf{y}}^{(u)} \cdots \hat{\mathbf{y}}^{(m)} \end{bmatrix} & (62)\end{aligned}$$

Loss function:

$$
\begin{aligned}
L &= -\frac{1}{m}\sum_{u=1}^{m}\sum_{i=1}^{k}\delta_i^{(u)}\log\hat{y}_i^{(u)} && (63)\\
\Rightarrow L &= -\frac{1}{m}\sum_{u=1}^{m}\sum \mathbf{y}^{[u]}\log\hat{\mathbf{y}}^{(u)} && (64)\\
\Rightarrow L &= -\frac{1}{m}\sum\left(\mathbf{y}^T\odot\log\hat{\mathbf{y}}\right) && (65)
\end{aligned}
$$

ii. Tính toán backward

- Gradient cho từng tham số tính trên sample thứ $w$:

$$
\frac{\partial L^{(w)}}{\partial\Theta_{ij}} = \sum_{u=1}^{k}\frac{\partial L^{(w)}}{\partial\hat{y}_u^{(w)}}\times\sum_{v=1}^{k}\frac{\partial\hat{y}_v^{(w)}}{\partial z_i^{(w)}}\times\frac{\partial z_i^{(w)}}{d\Theta_{ij}} \quad (66)
$$

Trong đó:

$$
\begin{aligned}
\frac{\partial z_i^{(w)}}{\partial\Theta_{ij}} &= x_j^{(w)}, && (67)\\
\sum_{v=1}^{k}\frac{\partial\hat{y}_v^{(w)}}{\partial z_i^{(w)}} &= \sum_{v=1}^{k}\hat{y}_v^{(w)}(y_v^{(w)}-\hat{y}_i^{(w)}), && (68)\\
\sum_{u=1}^{k}\frac{\partial L^{(w)}}{\partial\hat{y}_u^{(w)}} &= -\sum_{u=1}^{k}\frac{y_u^{(w)}}{\hat{y}_u^{(w)}} && (69)\\
\Rightarrow\sum_{u=1}^{k}\frac{\partial L^{(w)}}{\partial\hat{y}_u^{(w)}} &= \sum\mathbf{y}^{(w)}\oslash(\hat{\mathbf{y}}^{(w)})^T && (70)
\end{aligned}
$$

Do đó:

$$
\begin{aligned}
\frac{\partial L^{(w)}}{\partial\Theta_{ij}} &= -\sum\mathbf{y}^{(w)}\oslash(\hat{\mathbf{y}}^{(w)})^T\times\sum_{u=1}^{k}\frac{y_u^{(w)}}{\hat{y}_u^{(w)}}\times x_j^{(w)} && (71)\\
\Leftrightarrow\frac{\partial L^{(w)}}{\partial\Theta_{ij}} &= x_j^{(w)}(\hat{y}_i^{(w)}-y_i^{(w)}), && (72)\\
\frac{\partial L}{\partial\Theta_{ij}} &= \sum_{w=1}^{m}\frac{\partial L^{(w)}}{\partial\Theta_{ij}} && (73)\\
\Leftrightarrow\frac{\partial L}{\partial\Theta_{ij}} &= \frac{1}{m}\sum_{w=1}^{m}x_j^{(w)}(\hat{y}_i^{(w)}-y_i^{(w)}) && (74)
\end{aligned}
$$

Vectorization:

$$
\frac{\partial L}{\partial\boldsymbol{\Theta}} = \frac{1}{m}\mathbf{X}(\hat{\mathbf{Y}}^T-\mathbf{Y}) \quad (75)
$$

- Cập nhật tham số:

$$
\boldsymbol{\Theta} := \boldsymbol{\Theta}-\eta\times\frac{1}{m}\mathbf{X}(\hat{\mathbf{Y}}^T-\mathbf{Y}) \quad (76)
$$

**3.** Cài đặt bài toán softmax regression cho data `iris_1D_2c.csv` bằng phương pháp dựa vào one-hot encoding.

1. Stochastic gradient descent

2. Batch gradient descent

**4.** Cài đặt bài toán softmax regression cho data `iris_full.csv` bằng phương pháp dựa vào one-hot encoding.

1. Stochastic gradient descent

2. Batch gradient descent

**Lời giải cho bài 3 & 4**

LINK NOTEBOOK