Assignment 04

SoftMax Regression

Vũ Lê Thế Anh (20C13002)

Exercise 1 *Prove that softmax function maps a vector to a probability distribution.*

The softmax function is a function $\sigma(.)$ mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ defined by

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, i = 1, 2, \ldots, n \tag{1}$$

This mean that it takes in a vector $\mathbf{z} = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^n$ and outputs a new vector $\sigma(\mathbf{z})$ with the i-th element computed using (1).

We have that

$$\sum_{j=1}^{n} e^{z_j} = e^{z_i} + \sum_{1 \leq j \leq n, j \neq i} e^{z_j} \geq e^{z_i}$$

The inequality uses the fact that $e^x \geq 0, \forall x \in \mathbb{R}$. Using this fact, it is also obvious that $\sigma(\mathbf{z})_i \geq 0, \forall \mathbf{z} \in \mathbb{R}^n$ since it is the sum and division of non-negative numbers.

We can conclude that

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \leq 1, \forall \mathbf{z} \in \mathbb{R}^n.$$

In other words, the softmax function maps a vector in $\mathbb{R}^n$ to a vector in $[0, 1]^n$ (a probability vector). To be a distribution, we show also that the sum of its element is 1. It is indeed, since

$$\sum_{i=1}^{n} \sigma(\mathbf{z})_i = \sum_{i=1}^{n} \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} = \frac{1}{\sum_{j=1}^{n} e^{z_j}} \sum_{i=1}^{n} e^{z_i} = 1, \forall \mathbf{z} \in \mathbb{R}^n.$$

We conclude that the softmax function maps a vector to a probability distribution.

Exercise 2 *Find the gradient vector of the loss function in SoftMax Regression model.*

The SoftMax Regression model is that, for a given input $\mathbf{x}^{(i)} \in \mathbb{R}^M$, we can compute the output as follows:

1. Compute $z_c^{(i)} = \mathbf{w}^{(c)} \cdot \mathbf{x}^{(i)}$ where $\mathbf{w}^{(c)} \in \mathbb{R}^M$ is the c-th row of our weight matrix. Here, $\mathbf{z}^{(i)} = (z_1^{(i)}, z_2^{(i)}, \ldots, z_C^{(i)})$ is called the **logits** of the i-th sample.
2. Compute $y_c^{(i)} = \sigma(\mathbf{z}^{(i)})_c$ as our predicted probability for the i-th training data to belong to the c-th class

Given a set training data $\mathcal{D}$, the loss function in SoftMax Regression model is

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \left(- \sum_{c=1}^{C} \hat{y}_c^{(i)} \log y_c^{(i)} \right)$$

where $\hat{y}_c^{(i)}$ is our ground truth label for the i-th training data (1 if the input belongs to the c-th class, 0 otherwise).

To find the gradients, we must first find the gradient of the softmax function.

$$\frac{\partial}{\partial z_j} \sigma(\mathbf{z})_i = \frac{\partial}{\partial z_j} \frac{e^{z_i}}{\sum_{k=1}^{n} e^{z_k}} = \frac{1}{\left(\sum_{k=1}^{n} e^{z_k}\right)^2} \left[\left(\frac{\partial}{\partial z_j} e^{z_i} \right) \left(\sum_{k=1}^{n} e^{z_k} \right) - e^{z_i} \frac{\partial}{\partial z_j} \left(\sum_{k=1}^{n} e^{z_k} \right) \right]$$

Using the delta function $\delta_{i,j}$ which equals 1 when $i = j$ and 0 otherwise, we have

$$\frac{\partial}{\partial z_j} \sigma(\mathbf{z})_i = \frac{1}{\left(\sum_{k=1}^{n} e^{z_k}\right)^2} \left[\delta_{i,j} e^{z_i} \left(\sum_{k=1}^{n} e^{z_k} \right) - e^{z_i} e^{z_j} \right] = \frac{e^{z_i}}{\sum_{k=1}^{n} e^{z_k}} \left[\delta_{i,j} - \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}} \right]$$

We conclude the derivatives of the softmax function can be computed as

$$\frac{\partial}{\partial z_j} \sigma(\mathbf{z})_i = \sigma(\mathbf{z})_i [\delta_{i,j} - \sigma(\mathbf{z})_j]$$

Each term in the overall loss function is as follows

$$l(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{c=1}^{C} \hat{y}_c \log y_c = - \sum_{c=1}^{C} \hat{y}_c \log \sigma(\mathbf{z})_c$$

The derivative with respect to the prediction of the k-th element of the c-th weight vector

$$\frac{\partial}{\partial w_j^{(i)}} l(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{c=1}^{C} \frac{\partial}{\partial w_j^{(i)}} \hat{y}_c \log \sigma(\mathbf{z})_c = - \sum_{c=1}^{C} \frac{1}{\sigma(\mathbf{z})_c} \hat{y}_c \frac{\partial \sigma(\mathbf{z})_c}{\partial w_j^{(i)}}$$

Applying the chain rule and the above

$$\frac{\partial}{\partial w_j^{(i)}} l(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{c=1}^{C} \frac{1}{\sigma(\mathbf{z})_c} \hat{y}_c \sum_{k=1}^{C} \frac{\partial \sigma(\mathbf{z})_c}{\partial z_k} \frac{\partial z_k}{\partial w_j^{(i)}} = - \sum_{c=1}^{C} \hat{y}_c \sum_{k=1}^{C} [\delta_{c,k} - \sigma(\mathbf{z})_k] \frac{\partial z_k}{\partial w_j^{(i)}}$$

Since $z_k = \sum_{t=1}^{M} w_t^{(k)} x_t$, we have

$$\frac{\partial}{\partial w_j^{(i)}} l(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} \hat{y}_c \sum_{k=1}^{C} [\delta_{c,k} - \sigma(\mathbf{z})_k] \delta_{i,k} x_j = -\sum_{c=1}^{C} \hat{y}_c [\delta_{c,i} - \sigma(\mathbf{z})_i] x_j$$

In conclusion, the derivative with respect to $w_j^{(i)}$ (or $W_{i,j}$ of the weight matrix) of the softmax loss is

$$\frac{\partial}{\partial W_{i,j}} l(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} \hat{y}_c [\delta_{c,i} - \sigma(\mathbf{z})_i] x_j$$

In the case of single-label classification, only one component of $\hat{\mathbf{y}}$ is equal to 1: the component corresponds to the ground truth class.

References

[1] "Softmax function." [Online]. Available: https://en.wikipedia.org/wiki/Softmax_function

[2] E. Bendersky, "The softmax function and its derivative." [Online]. Available: https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/