Gradient Computation on Multi-class classifier

Zhihan Hu

February 2025

1 Introduction

In this document, we will talk about a more complex case where there are more than 2 labels. In this case, the predicted output layer will be calculated using softmax.

2 Math setup

Suppose we have a total of l labels.

- $X \in R^{d \times n}$: input training data set, where each data point has d features, and each col in X represents one data point.
- $\theta \in R^{L \times d}$: the learnable parameters that projects each data point in X to an l-dimensional vector
- $Z = \theta^T X + b \in R^{L \times n}$: the output value of applying θ to X
- $Y \in R^{L \times n}$: Each column is a one hot distribution where there is a 1 on the corresponding label of that data point.
- $b \in R^{L \times 1}$: the bias

Propagate:

$$A_{li} = \frac{e^{Z_{li}}}{\sum_{l'=1}^{L} e^{Z_{l'i}}} \in R^{L \times n} \tag{1}$$

Loss:

The loss function is cross-entropy loss, which shows as follows:

$$L = -\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} Y_{li} \log A_{li} \tag{2}$$

3 Gradient Computation

In this model, we primarily need to compute two gradients:

- $\frac{\delta L}{\delta \theta}$: The gradient of loss w.r.t the parameters
- $\frac{\delta L}{\delta b}$: The gradient of loss w.r.t the bias

By noticing that L depends on A, A depends on Z and Z depends on θ, we have the following chain rule:

$$\frac{\delta L}{\delta \theta} = \frac{\delta L}{\delta A} \frac{\delta A}{\delta Z} \frac{\delta Z}{\delta \theta} \tag{3}$$

The first term is rather simple:

$$\frac{\delta L}{\delta A_{li}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} \frac{Y_{li}}{A_{li}} \tag{4}$$

For the second term, let's assume the corresponding label for i^{th} data point is l. Noticing that only A_{li} will contribute to the Loss, so we only need to figure out **How Z contribute to** A_{li} for each data point. Note that, only the i^{th} column of Z contributes to A_{li}, as stated by equation 1. Then what we need to compute is

- $\frac{\delta A_{li}}{\delta Z_{li}}$, the element that corresponding to the true label
- $\frac{\delta A_{li}}{\delta Z_{l'i}}$, where l' $\neq$ l

The detailed derivation of the gradient of softmax is shown in appendix A and I will show here the result:

$$\frac{\delta A_{li}}{\delta Z_{li}} = A_{li}(1 - A_{li}) \tag{5}$$

$$\frac{\delta A_{li}}{\delta Z_{l'i}} = -A_{li} A_{l'i} \tag{6}$$

By having these equation, we can know that a change in Z_{li} will lead to a change of value $A_{li}(1 - A_{li})\delta Z_{li}$ in A_{li}, and thus a value change of

$$\frac{Y_{li}}{A_{li}}(A_{li}(1 - A_{li})) = Y_{li}(1 - A_{li}) = 1 - A_{li} \tag{7}$$

in the loss. Here I omit the sum because I assume that l is the corresponding label for data point i. and Y_{li} is thus 1.

Similarly, we can have a value change of L due to the value change of $A_{l'i}$ as

$$\frac{Y_{li}}{A_{li}}(-A_{li} A_{l'i}) = -Y_{li} A_{l'i} = -A_{l'i} \tag{8}$$

Thus, the gradient of the value on i^{th} column of Z is

$$\frac{\delta L}{\delta Z_i} = \frac{1}{n}[A_{1i}, A_{2i}, ..., A_{li} - 1, ..., A_{Li}]^T \tag{9}$$

To write it in matrix form, the seemingly complicated computation gracefully becomes a really simple form:

$$\frac{\delta L}{\delta Z} = \frac{1}{n}(A - Y) \tag{10}$$

Notice that, although this model is much more complex than logistic, the gradient w.r.t Z is the same.

For the final term, how θ change Z, the thought flow is similar to the logistic model, but a little bit more complicated. Let's start from a parameter θ_{lj}.

- It will affect the l^{th} row of Z, i.e, $Z_{l1},\ Z_{l2},\ ...,\ Z_{ln}$
- A change of $\delta\theta_{lj}$ will change the l^{th} row of Z by $X_{j1}\delta\theta_{li}, ..., X_{jn}\delta\theta_{li}$
- The overall impact on L will be

$$\sum_{i=1}^{n}(\frac{\delta L}{\delta Z})_{li}X_{ji}\delta\theta_{li} \tag{11}$$

To rewrite the above equation, we have

$$\frac{\delta L}{\delta \theta_{lj}} = (\frac{\delta L}{\delta Z})_l(X_j)^T \tag{12}$$

Meaning that the impact of θ_{lj} is the inner product of l^{th} row of $\frac{\delta L}{\delta Z}$ and j^{th} column of X^T.

Then, we can have

$$\frac{\delta L}{\delta \theta} = \frac{1}{n}(A - Y)X^T \tag{13}$$

For the gradient w.r.t biases. A change of the value in b_i will lead to change of value for the i^{th} row of Z. So the total change will be

$$\sum_{j=1}^{n}\frac{\delta L}{\delta Z}_{ij}\delta b \tag{14}$$

And thus we can get the gradient as

$$\frac{\delta L}{\delta b_i} = \mathbf{sum}(\frac{\delta L}{\delta Z})_i \tag{15}$$

Which is the sum of i^{th} row in $\frac{\delta L}{\delta Z}$

And thus we have

$$\frac{\delta L}{\delta b} = \mathbf{sum}(\frac{1}{n}(A - Y), axis = 1) \tag{16}$$

4 Key takeaway

3 equations are worth to be remembered and understood

$$\frac{\delta L}{\delta \theta} = \frac{1}{n}(A - Y)X^T \tag{17}$$

$$\frac{\delta L}{\delta b} = \textbf{sum}(\frac{1}{n}(A - Y), axis = 1) \tag{18}$$

$$\frac{\delta L}{\delta Z} = \frac{1}{n}(A - Y) \tag{19}$$

One other important thing to notice is that if

$$Z = \theta X + b \textbf{ or } Z = \theta^T X + b \tag{20}$$

Then

$$\frac{\delta Z}{\delta \theta} = X^T \tag{21}$$

5 Appendix A

Suppose

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{22}$$

Then, by using quotient rule, we have

$\frac{\delta y_i}{\delta x_i} = \frac{(\sum_{j=1}^{n} e^{x_j})e^{x_i} - e^{x_i}e^{x_i}}{(\sum_{j=1}^{n} e^{x_j})^2} = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} - (\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}})^2 = y_i - y_i^2 = y_i(1 - y_i)$

$\frac{\delta y_i}{\delta x_j} = \frac{-e^{x_i}e^{x_j}}{(\sum_{j=1}^{n} e^{x_j})^2} = -\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \frac{e^{x_j}}{\sum_{j=1}^{n} e^{x_j}} = -y_i y_j$