Gradient Computation on logistic model

Zhihan Hu

February 2025

1 Introduction

In this document, I will present a detailed derivation of gradients of the logistic model.

2 Math setup

- $X \in R^{d \times n}$: input training data set, where each data point has d features, and each col in X represents one data point.
- $\theta \in R^{d \times 1}$: the learnable parameters
- $Z = -\theta^T X + b \in R^{1 \times n}$: the output value of applying θ to X
- $Y \in R^{1 \times n}$: the corresponding label of each training data point.
- $b \in R$: the bias

Propagate:

$$A = \frac{1}{1 + e^Z} \in R^{1 \times n} \tag{1}$$

The predicted probability of label 0 of each training data point. The reason I use A here is that we can consider A as the activated output from a neural network.

Loss:

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

$$L = -\frac{1}{n} \sum_{i=1}^{n} Y_i \log A_i + (1 - Y_i) \log (1 - A_i) \tag{2}$$

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

Let's begin with the gradient of loss w.r.t the parameters. Notice that, L depends on A, A depends on Z, and Z depends on θ, and thus we can have the 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 partial derivative of each A_i is

$$\frac{\delta L}{\delta A_i} = -\frac{1}{n}(\frac{Y_i}{A_i} - \frac{1-Y_i}{1-A_i}) \tag{4}$$

For the second term, we're interested in how change of Z will affect the value of A. Note that, since each Z_i only affects the corresponding A_i ($A_i = \frac{1}{1+e^{Z_i}}$), we can have, by simple derivative, that

$$\frac{\delta A_i}{\delta Z_i} = A_i(1 - A_i) \tag{5}$$

From equation 4 and 5, what information needs to be extracted is that:

- Given a value change δZ_i, A_i will change by $A_i(1 - A_i)\delta Z_i$
- Given a value change $A_i(1 - A_i)\delta Z_i$ in A_i, the value change in L will be $-\frac{1}{n}(\frac{Y_i}{A_i} - \frac{1-Y_i}{1-A_i})A_i(1 - A_i)\delta Z_i$

And thus, we have the partial gradient

$$\frac{\delta L}{\delta Z_i} = -\frac{1}{n}(\frac{Y_i}{A_i} - \frac{1-Y_i}{1-A_i})A_i(1 - A_i) = \frac{1}{n}(A_i - Y_i) \tag{6}$$

To write in matrix form, we have

$$\frac{\delta L}{\delta Z} = \frac{1}{n}(A - Y) \tag{7}$$

For the third term, we are going to figure out the effect of a change in θ_i to the value change in Z. Since $Z = \theta^T X + b$, meaning that θ is the coefficient for i^{th} feature of each data point. i.e, **A change of $\delta\theta_i$ will cause a change of $\delta\theta_i X_{ij}$ of** Z_j. Then we can think of it as

- A change of $\delta\theta_i X_{ij}$ for Z_j cause a change of $\frac{1}{n}(A_j - Y_j)\delta\theta_i X_{ij}$ in L
- All Z_j value will be changed due to change of θ_i, so the overall effect will be $\sum_{j=1}^{n} \frac{1}{n}(A_j - Y_j)\delta\theta_i X_{ij}$

- The above equation can be written as $\frac{\delta L}{\delta \theta_i} = \frac{1}{n}(A - Y)X_i^T$, where X_i^T is the transpose of i^{th} row of X

Based on the above derivation, we can write out the matrix form as

$$\frac{\delta L}{\delta \theta} = \frac{1}{n}(A - Y)X^T \tag{8}$$

However, in order to keep the dimensions, we use the transpose as the gradient:

$$\frac{\delta L}{\delta \theta} = \frac{1}{n}X(A - Y)^T \tag{9}$$

We can do the same analysis on the gradient of L w.r.t the bias. The chain rule is

$$\frac{\delta L}{\delta \theta} = \frac{\delta L}{\delta A}\frac{\delta A}{\delta Z}\frac{\delta Z}{\delta b} \tag{10}$$

The first two terms are the same as them in the gradient w.r.t theta, and thus is equal to $\frac{1}{n}(A - Y)$. When there is a value change of δb, it will

- Change each Z_i by δb
- Change of δb in Z_i cause a change of $\frac{1}{n}(A_i - Y_i)\delta b$ in L
- The overall change will be $\frac{1}{n}\sum_{i=1}^{n}(A_i - Y_i)\delta b$

The above equation can be written as

$$\frac{\delta L}{\delta b} = \frac{1}{n}\mathbf{sum}(A - Y) \tag{11}$$

4 Key takeaways

Three equations are worth to remember:

$$\frac{\delta L}{\delta \theta} = \frac{1}{n}X(A - Y)^T \tag{12}$$

$$\frac{\delta L}{\delta b} = \frac{1}{n}\mathbf{sum}(A - Y) \tag{13}$$

$$\frac{\delta L}{\delta Z} = \frac{1}{n}(A - Y) \tag{14}$$

Later when we talk about gradient computation of deep learning models, you will get an intuition on the values of remembering and understanding these 3 equation.