G Gianluca Turcatel

Dec 23, 2021 1 min read

Derivation of the Binary Cross Entropy Loss Gradient

The binary cross entropy loss function is the preferred loss function in binary classification tasks, and is utilized to estimate the value of the model's parameters through gradient descent. In order to apply gradient descent we must calculate the derivative (gradient) of the loss function w.r.t. the model's parameters. Deriving the gradient is usually the most tedious part of training a machine learning model.
In this article we will derive the derivative of the binary cross entropy loss function w.r.t. W, step by step.

The binary cross entropy loss is given by

$$L(\boldsymbol{W}) = -[\boldsymbol{y} * \log(\hat{\boldsymbol{y}}) + (1 - \mathbf{y}) * \log(1 - \hat{\boldsymbol{y}})] \quad (1)$$

y is the observed class, y_hat the prediction, W the model's parameters. Predictions are given by:

$$\hat{\boldsymbol{y}} = \sigma(z) = \frac{1}{1 + e^{-\mathbf{z}}} \quad (2)$$

z is equal to:

$$\boldsymbol{z} = \boldsymbol{w_0} + \boldsymbol{x}_1 * \boldsymbol{w_1} + \boldsymbol{x_2} * \boldsymbol{w_2} + \cdots + \boldsymbol{x_k} * \boldsymbol{w_k} = \boldsymbol{W}^T \boldsymbol{X} \quad (3)$$

To calculate the gradient of L(W) w.r.t. W we will use the chain rule:

$$\frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}} = \frac{\partial L(\boldsymbol{W})}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} \quad (4)$$

Let's derive the first term:

$$\frac{\partial L(\boldsymbol{W})}{\partial \hat{\boldsymbol{y}}} = -\left(\frac{\boldsymbol{y}}{\hat{\boldsymbol{y}}} - \frac{1-\boldsymbol{y}}{1-\hat{\boldsymbol{y}}}\right) = \frac{\hat{\boldsymbol{y}} - \boldsymbol{y}}{\hat{\boldsymbol{y}}(1-\hat{\boldsymbol{y}})} \qquad (5)$$

The second term is a little more complicated:

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}} = \frac{\partial}{\partial \boldsymbol{z}}\left[\frac{1}{1+e^{-\boldsymbol{z}}}\right] = \frac{\partial}{\partial \boldsymbol{z}}(1+e^{-\boldsymbol{z}})^{-1} = \qquad (6)$$

$$= -(1+e^{-\boldsymbol{z}})^{-2}(-e^{-\boldsymbol{z}}) = \frac{e^{-\boldsymbol{z}}}{(1+e^{-\boldsymbol{z}})^{2}} = \qquad (7)$$

$$= \frac{1}{(1+e^{-\boldsymbol{z}})}\,\frac{e^{-\boldsymbol{z}}}{(1+e^{-\boldsymbol{z}})} = \frac{1}{1+e^{-\boldsymbol{z}}}\,\frac{(1+e^{-\boldsymbol{z}})-1}{(1+e^{-\boldsymbol{z}})} = (8)$$

$$= \frac{1}{1+e^{-\boldsymbol{z}}}\left(\frac{1+e^{-\boldsymbol{z}}}{1+e^{-\boldsymbol{z}}} - \frac{1}{1+e^{-\boldsymbol{z}}}\right) = \tag{9}$$

$$= \frac{1}{1+e^{-\boldsymbol{z}}}\left(1 - \frac{1}{1+e^{-\boldsymbol{z}}}\right) = \hat{\boldsymbol{y}}(1-\hat{\boldsymbol{y}}) \tag{10}$$

Done with the second term. The derivative of the third term is straight forward:

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} = \boldsymbol{X} \tag{11}$$

Now let's put everything together:

$$\frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}} = \frac{\hat{\boldsymbol{y}} - \boldsymbol{y}}{\hat{\boldsymbol{y}}(1-\hat{\boldsymbol{y}})}\hat{\boldsymbol{y}}(1-\hat{\boldsymbol{y}})\boldsymbol{X} = (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{X} \tag{12}$$

And there you have it: the derivative of the binary cross entropy loss function w.r.t. the model's parameters.

Follow me on Twitter and Facebook to stay updated.

#Python #MachineLearning #CrossEntropy #BinaryClassification