

Medium Test 3

Avinash Barnwal

March 2019

Problem Derive the formula for first- and second-order partial derivatives of the loss function for binary classification. The probability for obtaining the i -th label (y_i) given the i -th training data point (x_i) is as follows:

First Expression

$$P(y_i|x_i) = \begin{cases} \sigma(\hat{y}_i) & \text{if } y_i = 1 \\ 1 - \sigma(\hat{y}_i) & \text{else} \end{cases} \quad (1)$$

Second Expression

$$P(y_i|x_i) = \sigma(\hat{y}_i)^{(y_i)} * (1 - \sigma(\hat{y}_i))^{(1-y_i)} \quad (2)$$

where \hat{y}_i is a prediction score (range between -inf to inf) for x_i produced by our model the label y_i is either 0 or 1 $\sigma(*)$ is the sigmoid function. Note that the sigmoid function converts any real number into a probability value between 0 and 1.

Q1. Explain why the first expression is equivalent to the second expression?

Ans:- Considering $\sigma(\hat{y}_i)$ being the probability operator. First expression is Bernoulli distribution with $P(y_i = 1|x_i) = \sigma(\hat{y}_i)$. As we know that probability mass function of Bernoulli distribution is as follows

$$f(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases} \quad (3)$$

It can also be written as This can also be expressed as

$$f(k;p) = p^k * (1 - p)^{1-k} \quad k \in \{0, 1\} \quad (4)$$

We can also see that if we put $k = 1$, we get the first part of First expression and $k = 0$, we get the second part of Second expression.

Using the principle of Maximum Likelihood Estimation, we will choose the best \hat{y}_i so as to maximize the value of $P(y_i|x_i)$, i.e. choose \hat{y}_i to make the training data most probable. The "distance" between the prediction \hat{y}_i and the true label y_i , is given as the negative logarithm of $P(y_i|x_i)$:

$$loss(y_i, \hat{y}_i) = -\log(P(y_i|x_i)) \quad (5)$$

$$loss(y_i, \hat{y}_i) = -\log(\sigma(\hat{y}_i)^{(y_i)} * (1 - \sigma(\hat{y}_i))^{(1-y_i)}) \quad (6)$$

Q2. Explain how minimizing the loss function $loss(y_i, \hat{y}_i)$ is equivalent to maximizing the probability $P(y_i|x_i)$?

Ans:- $loss(y_i, \hat{y}_i)$ is $-\log(P(y_i|x_i))$ which is monotonically decreasing mapping of $P(y_i|x_i)$ with negative sign. As log function is monotonically increasing function and we have multiplied it -1 which made it monotonically decreasing function. Therefore, maximizing $P(y_i|x_i)$ is equivalent to minimizing $loss(y_i, \hat{y}_i)$.

Q3. Simplify the expression for $loss(y_i, \hat{y}_i)$. Show your steps (i.e. don't just write the answer, show how you got it) ?

Ans:-

$$\begin{aligned} loss(y_i, \hat{y}_i) &= -\log(\sigma(\hat{y}_i)^{(y_i)} * (1 - \sigma(\hat{y}_i))^{(1-y_i)}) \\ &= -(y_i * \log(\sigma(\hat{y}_i)) - (1 - y_i) * \log(1 - \sigma(\hat{y}_i))) \\ &= -\log(1 - \sigma(\hat{y}_i)) - y_i \log\left(\frac{\sigma(\hat{y}_i)}{1 - \sigma(\hat{y}_i)}\right) \end{aligned} \quad (7)$$

As

$$\begin{aligned} \log\left(\frac{\sigma(\hat{y}_i)}{1 - \sigma(\hat{y}_i)}\right) &= \log\left(\frac{e^{\hat{y}_i}}{1 + e^{\hat{y}_i}}\right) \\ &= \log(e^{\hat{y}_i}) \\ &= \hat{y}_i \end{aligned} \quad (8)$$

Also

$$\log(1 - \sigma(\hat{y}_i)) = \log\left(1 - \frac{e^{\hat{y}_i}}{1 + e^{\hat{y}_i}}\right) = -\log(1 + e^{\hat{y}_i}) \quad (9)$$

Using Equation 8 and 9,

$$loss(y_i, \hat{y}_i) = \log(1 + e^{\hat{y}_i}) - y_i * \hat{y}_i \quad (10)$$

Q4. Now compute the first and second partial derivatives of $loss(y_i, \hat{y}_i)$ with respect to the second variable \hat{y}_i . Then express the two derivatives in terms of $\sigma(\hat{y}_i)$. Notice how simple the expressions become. Again, show your steps (i.e. don't just write the answer, show how you got it).

Ans:- Gradient or First Order partial derivatives of $loss(y_i, \hat{y}_i)$

$$\begin{aligned} \frac{\partial loss(y_i, \hat{y}_i)}{\partial \hat{y}_i} &= \frac{\partial(\log(1 + e^{\hat{y}_i}) - y_i * \hat{y}_i)}{\partial \hat{y}_i} \\ &= \frac{e^{\hat{y}_i}}{1 + e^{\hat{y}_i}} - y_i \\ &= \sigma(\hat{y}_i) - y_i \end{aligned} \quad (11)$$

Hessian or Second Order partial derivatives of $loss(y_i, \hat{y}_i)$

$$\begin{aligned}
\frac{\partial^2 \text{loss}(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} &= \frac{\partial}{\partial \hat{y}_i} \frac{\partial \text{loss}(y_i, \hat{y}_i)}{\partial \hat{y}_i} \\
&= \frac{\partial \frac{e^{\hat{y}_i}}{1+e^{\hat{y}_i}}}{\partial \hat{y}_i} \\
&= \frac{(1+e^{\hat{y}_i}) * e^{\hat{y}_i} - e^{\hat{y}_i} * e^{\hat{y}_i}}{(1+e^{\hat{y}_i})^2} \\
&= \frac{e^{\hat{y}_i}}{(1+e^{\hat{y}_i})^2} \\
&= \frac{e^{\hat{y}_i}}{1+e^{\hat{y}_i}} \frac{1}{1+e^{\hat{y}_i}} \\
&= \sigma(\hat{y}_i) * (1 - \sigma(\hat{y}_i))
\end{aligned} \tag{12}$$

Q5. In the source code `src/objective/regression_loss.h`, locate the structure that implements this loss function.

Ans:- Parent Structure is `LogisticRegression` starts at Line no - 40 and ends at 70 with "rmse" as default error but to have Logistic Classification , children structure is created `LogisticClassification` starts at Line no 73 and ends at 75 with log-loss as evaluation metric.