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

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

Second Expression

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

where $yhat_i$ 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 $P(y_i = 1|x_i) = \sigma(\hat{y})$. As we know that probability mass function of Bernoulli distribution is as follows

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

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

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

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 $P(y_i|x_i)$:

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

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

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:-

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

As

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

Also

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

Using Equation 8 and 9,

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

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 $loss(y_i, \hat{y}_i)$

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

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

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