Machine Learning: Logistic Regression

Tran Kha Uyen

Ngày 26 tháng 10 năm 2022

1 Problem 1:

Question: Show that $\frac{\partial L}{\partial w} = X^T(\hat{y} - y)$

Answer: Assume:

$$\hat{y_i} = \sigma(w_0 + w_1 * x_1^{(i)}) \tag{1}$$

$$\sigma_i = \frac{1}{e^{-(w_0 + w_1 * x_1^{(i)})}} \tag{2}$$

In matric form:

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ . & . \\ 1 & x_1^{(n)} \end{bmatrix} \tag{3}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \tag{4}$$

$$\Rightarrow Xw = \begin{bmatrix} w_0 + w_1 * x_1^{(1)} \\ ... \\ w_0 + w_1 * x_1^{(n)} \end{bmatrix} \tag{5}$$

$$\Rightarrow \hat{y} = \sigma(Xw) \tag{6}$$

Binary cross-entropy loss:

$$L = -(y_i * log(\hat{y_i}) - (1 - y) * log(1 - \hat{y_i}) \tag{7}$$

Take derivative of loss respect to w:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial w} \tag{8}$$

Firstly, calculate $\frac{\partial L}{\partial \hat{y}}$

$$
\begin{aligned}
\frac{\partial L}{\partial \hat{y}} &= \frac{\partial - (y_i * log(\hat{y}_i) - (1 - y_i) * log(1 - \hat{y}_i))}{\partial \hat{y}_i} && (9) \\
&= -(\frac{1}{\hat{y}_i} * y_i - \frac{-1}{1 - \hat{y}_i}(1 - y_i)) && (10) \\
&= -(\frac{y_i}{\hat{y}_i} + \frac{(1 - y_i)}{1 - \hat{y}_i}) && (11) \\
&= \frac{1 - y_i}{1 - \hat{y}_i} - \frac{y_i}{\hat{y}_i} && (12) \\
& && (13)
\end{aligned}
$$

Next, calculate the other part $\frac{\partial \hat{y}_i}{\partial w_k}$

$$
\begin{aligned}
\frac{\partial \hat{y}_i}{\partial w_k} &= \frac{\sigma(w_0 + w_1 * x_1^{(i)})}{\partial w_k} && (14) \\
&= \frac{\partial \frac{1}{1 + e^{-(w_0 + w_1 * x_1^{(i)})}}}{\partial w_k} && (15) \\
&= \frac{-1}{1 + e^{-(w_0 + w_1 * x_1^{(i)})}} - e^{-(w_0 + w_1 * x_1^{(i)})} \frac{\partial w_0 + w_1 * x_1^{(i)}}{\partial w_k} && (16) \\
&= \frac{e^{-(w_0 + w_1 * x_1^{(i)})}}{(1 + e^{-(w_0 + w_1 * x_1^{(i)})})^2} \frac{\partial w_0 + w_1 * x_1^{(i)}}{\partial w_k} && (17) \\
&= \frac{1}{1 + 1 + e^{-(w_0 + w_1 * x_1^{(i)})}} \frac{e^{-(w_0 + w_1 * x_1^{(i)})}}{1 + 1 + e^{-(w_0 + w_1 * x_1^{(i)})}} \frac{\partial w_0 + w_1 * x_1^{(i)}}{\partial w_k} && (18) \\
&= \hat{y}_i \frac{1 + e^{-(w_0 + w_1 * x_1^{(i)})} - 1}{1 + e^{-(w_0 + w_1 * x_1^{(i)})}} \frac{\partial w_0 + w_1 * x_1^{(i)}}{\partial w_k} && (19) \\
&= \hat{y}_i (1 - \hat{y}_i) x_k && (20)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{\partial L}{\partial w} &= (\frac{1 - y_i}{1 - \hat{y}_i} - \frac{y_i}{\hat{y}_i})(\hat{y}_i (1 - \hat{y}_i) x_k) && (21) \\
&= \frac{(1 - y_i)\hat{y}_i - y_i(1 - \hat{y}_i)}{(1 - \hat{y}_i)\hat{y}_i} \hat{y}_i (1 - \hat{y}_i) x_k && (22) \\
&= (\hat{y}_i - y_i) x_k && (23)
\end{aligned}
$$

So in matrix form, we can write as:

$$
\frac{\partial L}{\partial w} = X^T (\hat{y} - y) \qquad (24)
$$

2 Problem 2:

Question: Show that in logistic model, loss binary crossentropy is convex function with W while loss mean square error is not convex function with W **Answer** * loss binary crossentropy is convex function with W because:

$$\begin{aligned}\frac{\partial^2 L}{\partial w^2} &= \frac{\partial((\hat{y}_i - y_i)x_k)}{\partial w} &(25)\\ &= (\frac{\partial \hat{y}}{\partial w} - \frac{\partial y}{\partial w})x_k &(26)\\ &= (\hat{y}(1-\hat{y})x_k - 0)x_k &(27)\\ &= \hat{y}(1-\hat{y})x_k^2 \geq 0 \forall \hat{y} \in [0,1] &(28)\end{aligned}$$

So, loss binary crossentropy is convex function

* loss mean square error is not convex function with W:

$$MSE = \frac{1}{N}\sum_{i}^{N}(\hat{y} - y)^2 \tag{29}$$

Take the first order derivative:

$$\begin{aligned}\frac{\partial L}{\partial w} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial w} &(30)\\ &= 2(\hat{y} - y)\frac{\partial \hat{y}}{\partial w} &(31)\\ &= 2(\hat{y} - y)\hat{y}(1-\hat{y})x_k &(32)\\ &= 2(\hat{y} - y)(\hat{y}^2 - y\hat{y})(1-\hat{y})x_k &(33)\\ &= 2(\hat{y}^2 - \hat{y}^3 - y\hat{y} + y\hat{y}^2)x_k &(34)\end{aligned}$$

Next, calculate the 2nd order derivative:

$$\begin{aligned}\frac{\partial^2 L}{\partial w^2} &= \frac{\partial(2(\hat{y}^2 - \hat{y}^3 - y\hat{y} + y\hat{y}^2)x_k)}{\partial w} &(35)\\ &= 2(\frac{\hat{y}^2}{\partial w} - \frac{\hat{y}^3}{\partial w} - y\frac{\hat{y}}{\partial w} + y\frac{\hat{y}^2}{\partial w})x_k &(36)\\ &= 2(2\hat{y}\frac{\hat{y}}{\partial w} - 3\hat{y}\frac{\hat{y}}{\partial w} - y\frac{\hat{y}}{\partial w} + y2\hat{y}\frac{\hat{y}}{\partial w})x_k &(37)\\ &= 2\frac{\hat{y}}{\partial w}(2\hat{y} - 3\hat{y} - y + 2y\hat{y})x_k &(38)\\ &= 2(2\hat{y} - 3\hat{y})\hat{y}(1-\hat{y})x_k^2 &(39)\end{aligned}$$

$(2\hat{y} - 3\hat{y})$ is not always larger than 0 when $\forall \hat{y} \in [0,1]$. So loss mean square error is not convex function with W