Open in app

Follow

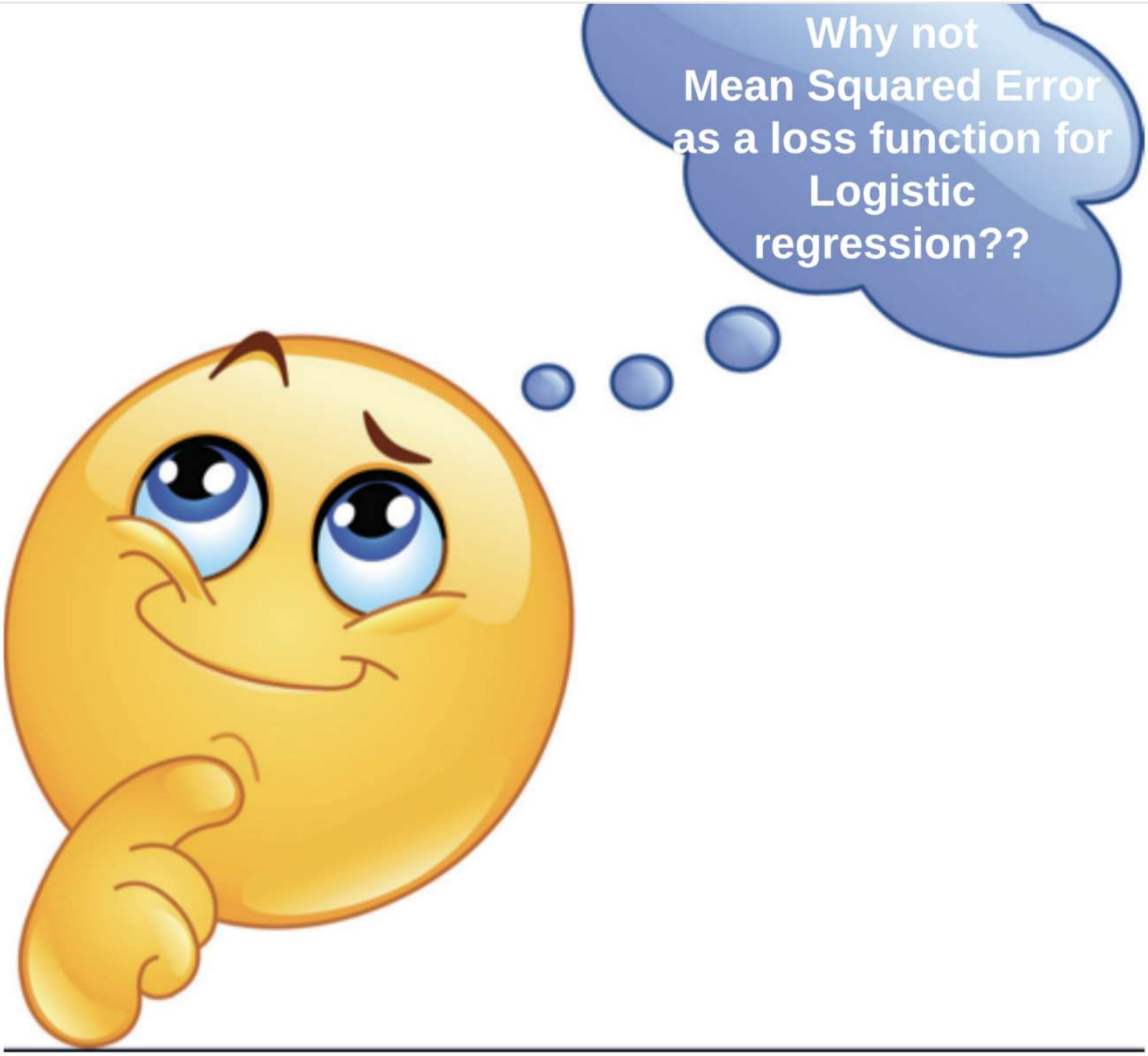
571K Followers

This is your **last** free member-only story this month. Upgrade for unlimited access.

Why not Mean Squared Error(MSE) as a loss function for Logistic Regression? 🤔

Rajesh Shreedhar Bhat Sep 16, 2019 · 5 min read ★

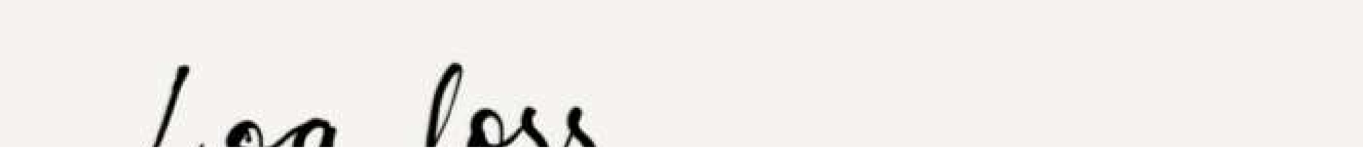
Authors: Rajesh Shreedhar Bhat*, Souradip Chakraborty* (* denotes equal contribution).



> *In this blog post, we mainly compare "**log loss**" vs **"mean squared error"** for logistic regression and show that why **log loss** is recommended for the same based on empirical and mathematical analysis.*

Equations for both the loss functions are as follows:

Log loss:



$$\mathcal{L}(y-\hat{y}) = -\sum_{i=1} y_i \log(\hat{y}_i)$$

Figure 1: Log Loss

Mean Squared Loss:

Mean Squared Loss

$$\mathcal{L}(y-\hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Figure 2: Mean Squared Error

In the above two equations

y: actual label

ŷ: predicted value

n: number of classes

Let's say we have a dataset with 2 classes(n = 2) and the labels are represented as **“0”** and **“1”.**

For example:

Let’s say

- Actual label for a given sample in a dataset is “1”
- Prediction from the model after applying sigmoid function = 0

Loss value when using MSE:

$(1- 0)^2 = 1$

Loss value when using log loss:

Before plugging in the values for loss equation, we can have a look at how the graph of ***log(x)*** looks like.

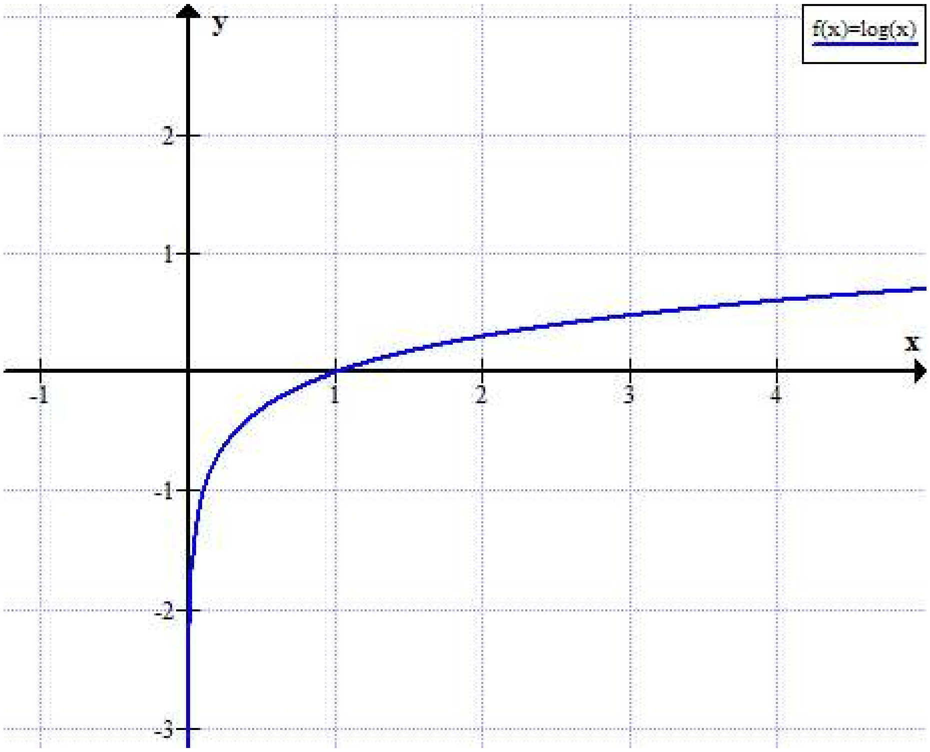


Figure 3: f(x) = log(x)

As seen from the above graph as **x tends to 0, *log(x) tends to -infinity.***

-(1 * log(0) + 0 * log(1)) = tends to infinity !!

As seen above, loss value using MSE was much much less compared to the loss value computed using the log loss function. Hence it is very clear to us that MSE doesn't strongly penalize misclassifications even for the perfect mismatch!

However, if there is a perfect match between predicted values and actual labels both the loss values would be "0" as shown below.

Actual label: **"1"**

Predicted: **"1"**

MSE: (1 - 1)² = 0

Log loss: -(1 * log(1) + 0 * log(0)) = 0

Here we have shown that MSE is not a good choice for binary classification problems. But the same can be extended for multi-class classification problems given that target values are one-hot encoded.

MSE and problem of Non-Convexity in Logistic Regression.

In classification scenarios, we often use gradient-based techniques(Newton Raphson, gradient descent, etc ..) to find the optimal values for coefficients by minimizing the loss function. Hence if the loss function is not convex, it is not guaranteed that we will always reach the global minima, rather we might get stuck at local minima.

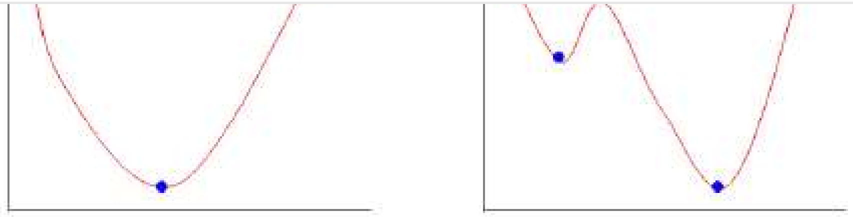


Figure 4: Convex and non-Convex functions

Before diving deep into why MSE is not a convex function when used in logistic regression, first, we will see what are the conditions for a function to be convex.

A real-valued function defined on an *n*-dimensional interval is called **convex** if the line segment between any two points on the graph of the function lies above or on the graph.

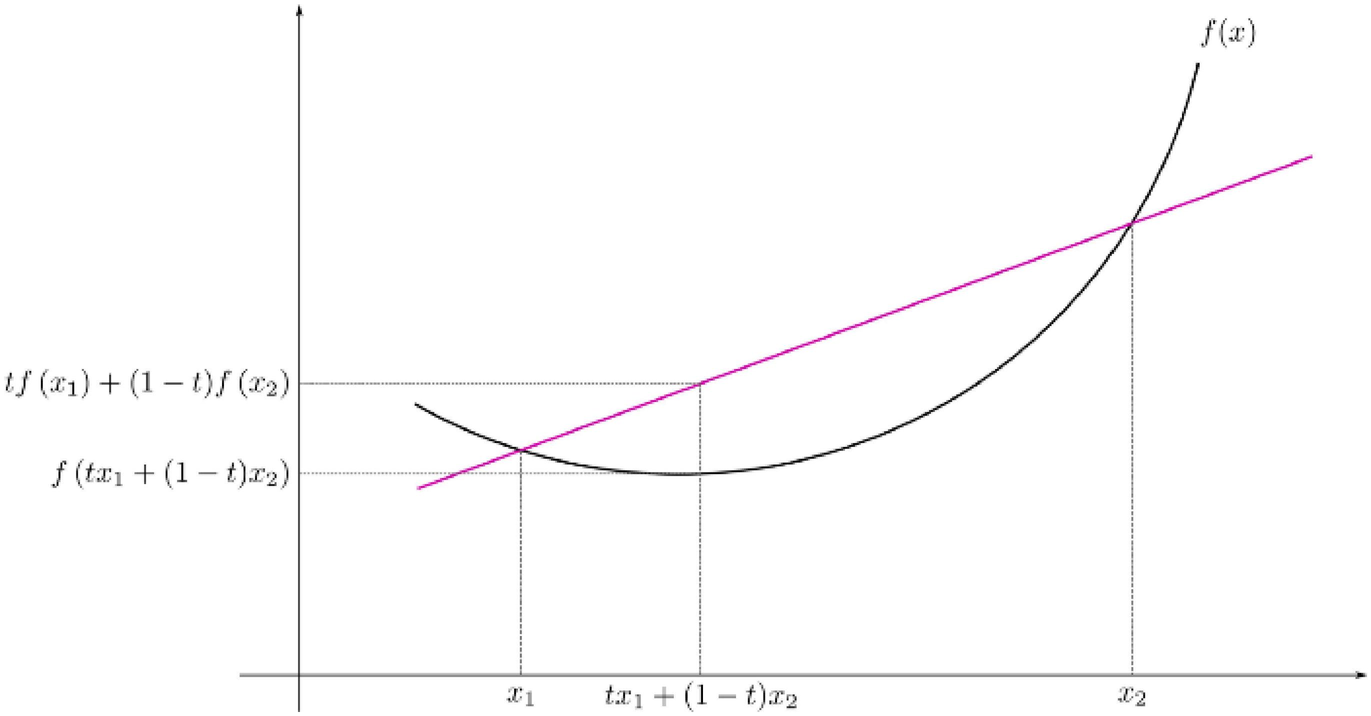
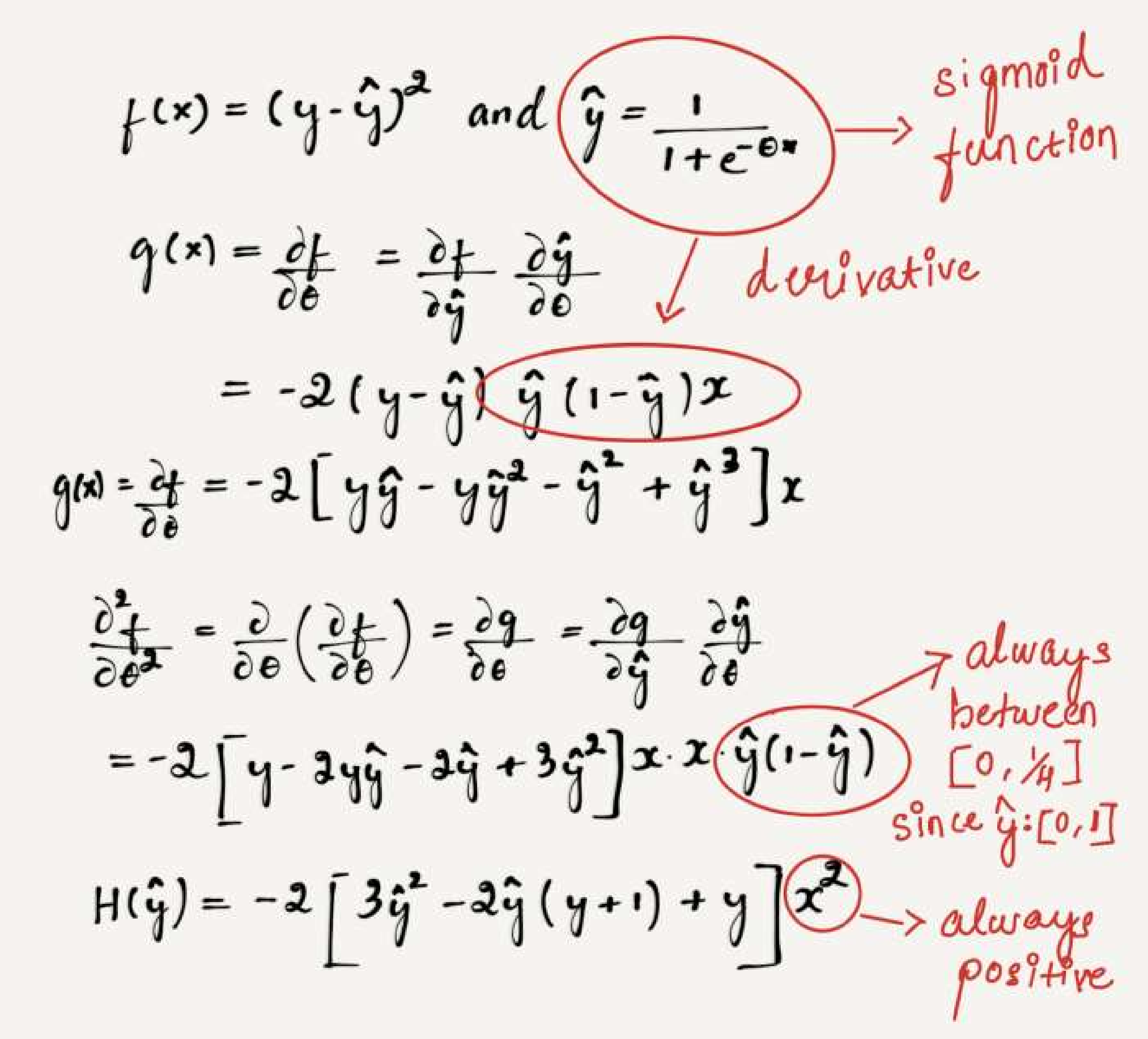


Figure 5: Convexity of a function

If **f** is twice differentiable and the domain is the real line, then we can characterize it as follows:

f is convex if and only if **f "(x) ≥ 0** for all ***x***. ***Hence if we can show that the double derivative of our loss function is ≥ 0 then we can claim it to be convex.*** *For more details, you can refer to this video.*

For simplicity, let's assume we have one feature **“x”** and **“binary labels”** for a given dataset. In the below image **f(x) = MSE** and ŷ is the predicted value obtained after applying sigmoid function.



$$f(x) = (y-\hat{y})^2 \quad \text{and} \quad \hat{y} = \frac{1}{1+e^{-\theta x}} \longrightarrow$$

sigmoid function

$$g(x) = \frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta}$$

derivative

$$= -2(y-\hat{y})\,\hat{y}(1-\hat{y})x$$

$$g(x) = \frac{\partial f}{\partial \theta} = -2\left[y\hat{y} - y\hat{y}^2 - \hat{y}^2 + \hat{y}^3\right]x$$

$$\frac{\partial^2 f}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left(\frac{\partial f}{\partial \theta}\right) = \frac{\partial g}{\partial \theta} = \frac{\partial g}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta}$$

$$= -2\left[y - 2y\hat{y} - 2\hat{y} + 3\hat{y}^2\right]x \cdot x \cdot \hat{y}(1-\hat{y})$$

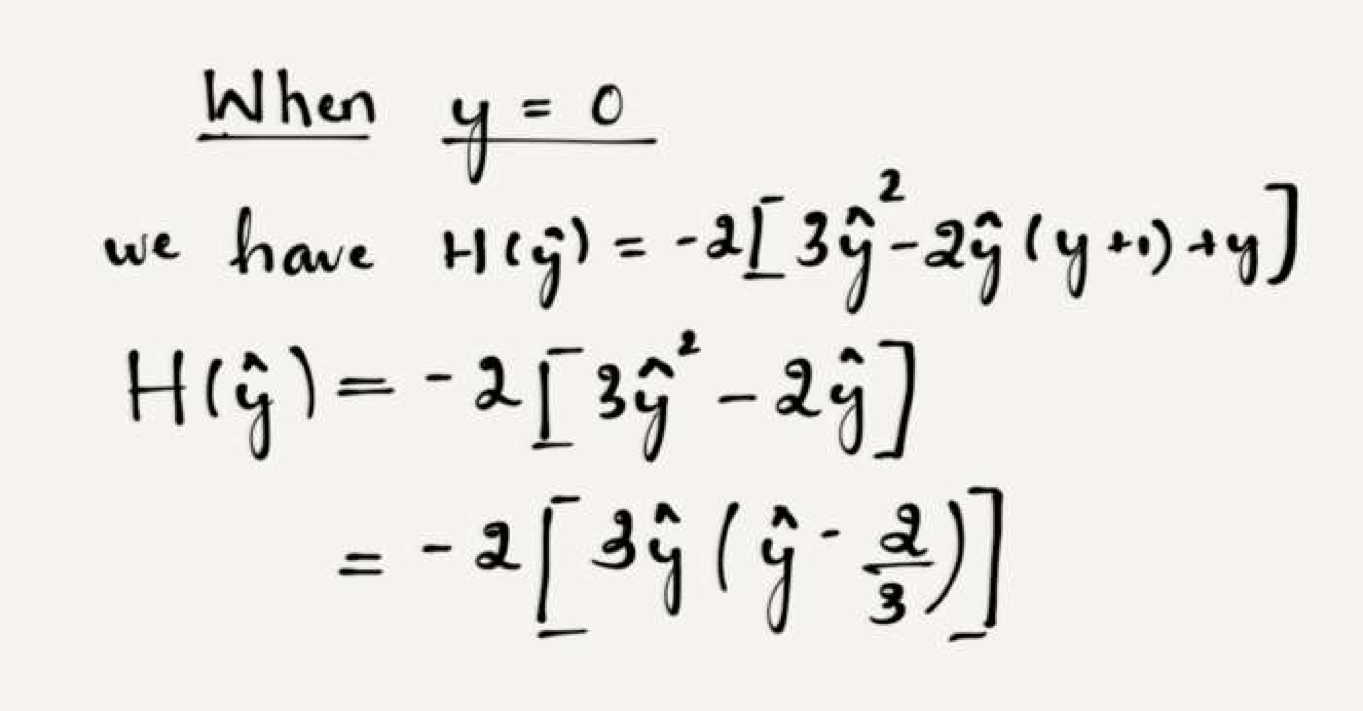
$\rightarrow$ always between $[0, \frac{1}{4}]$ since $\hat{y}:[0,1]$

$$H(\hat{y}) = -2\left[3\hat{y}^2 - 2\hat{y}(y+1) + y\right]x^2 \longrightarrow$$

always positive

Figure 6: MSE double derivative

From the above equation, **ŷ * (1 - ŷ)** lies between [0, 1]. Hence we have to check that if **H(ŷ)** is positive for all values of **“x”** or not, to be a convex function.



When $\underline{y = 0}$

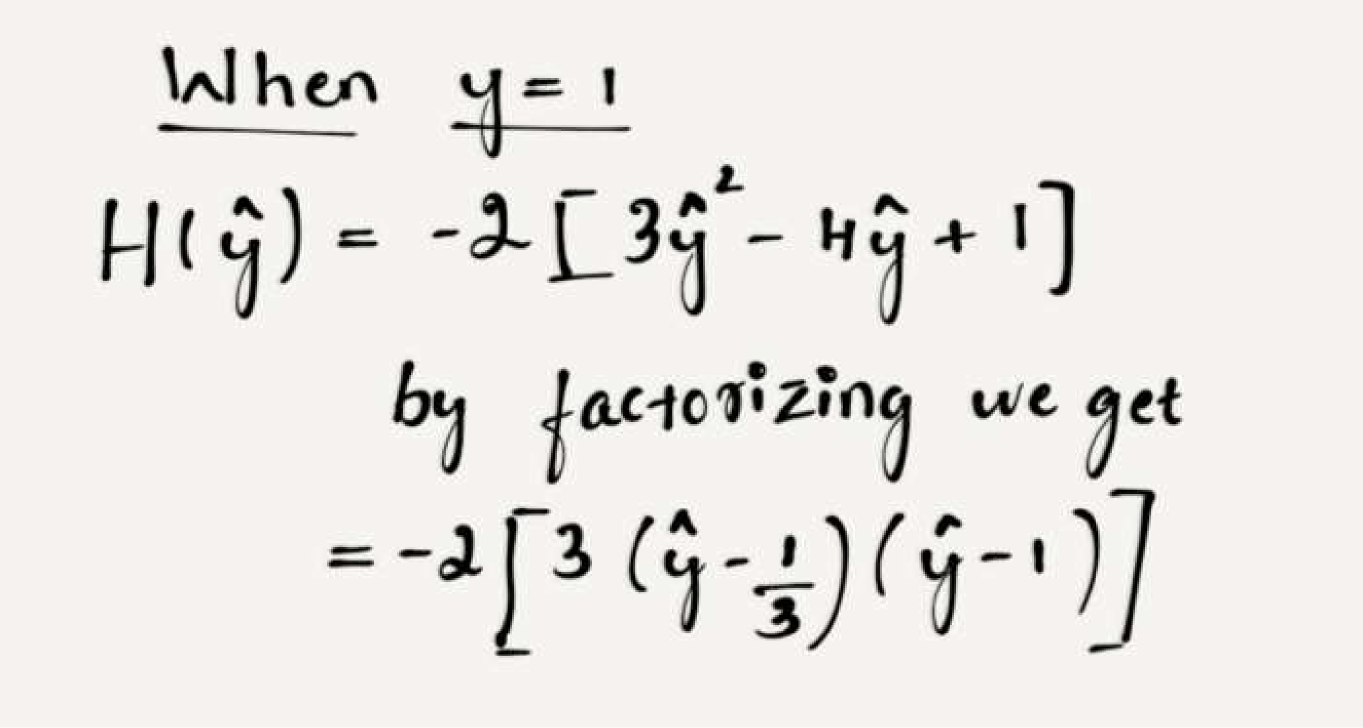
we have $H(\hat{y}) = -2[3\hat{y}^2 - 2\hat{y}(y+1) + y]$

$$H(\hat{y}) = -2[3\hat{y}^2 - 2\hat{y}]$$

$$= -2\left[3\hat{y}\left(\hat{y} - \frac{2}{3}\right)\right]$$

Figure 7: Double derivate of MSE when y=0

So in the above case when y = 0, it is clear from the equation that when ŷ lies in the range **[0, 2/3]** the function **H(ŷ) ≥ 0 and** when ŷ lies between **[2/3, 1]** the function **H(ŷ) ≤ 0.** This shows the function is not convex.



When $\underline{y = 1}$

$$H(\hat{y}) = -2[3\hat{y}^2 - 4\hat{y} + 1]$$

by factorizing we get

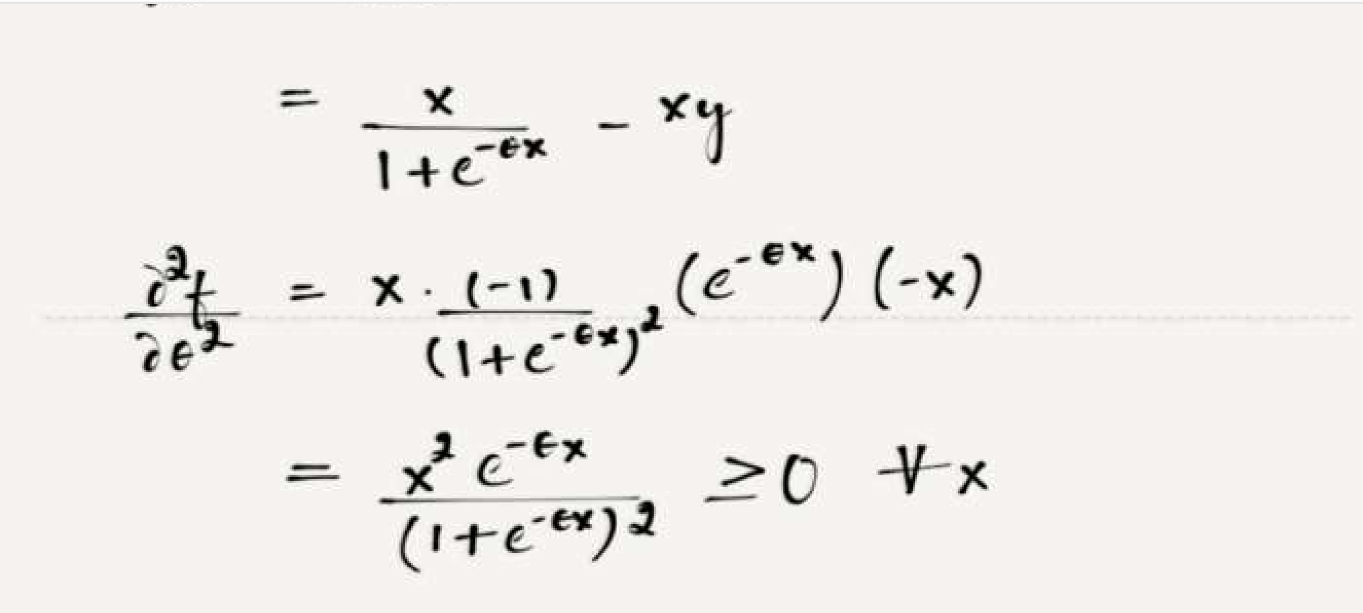
$$= -2\left[3\left(\hat{y} - \frac{1}{3}\right)(\hat{y} - 1)\right]$$

Now, when **y = 1**, it is clear from the equation that when ŷ lies in the range **[0, 1/3]** the function **H(ŷ) ≤ 0 and** when ŷ lies between **[1/3, 1]** the function **H(ŷ) ≥ 0.** This also shows the function is not convex.

Hence, based on the convexity definition we have mathematically shown the MSE loss function for logistic regression is non-convex and not recommended.

Now comes the question of **convexity of the "log-loss" function!!** We will mathematically show that log loss function is convex for logistic regression.

$$
\begin{aligned}
-f(x) &= y\log(\hat{y}) + (1-y)\log(1-\hat{y}) \\
&= y\log\left(\frac{1}{1+e^{-\theta x}}\right) + (1-y)\log\left(1-\frac{1}{1+e^{-\theta x}}\right) \\
&= y\log\left(\frac{e^{\theta x}}{1+e^{\theta x}}\right) + (1-y)\log\left(\frac{1}{1+e^{\theta x}}\right) \\
&= y\log(e^{\theta x}) - y\log(1+e^{\theta x}) \\
&\quad + \cancelto{0}{(1-y)}\log(1) - (1-y)\log(1+e^{\theta x}) \\
&= y(\theta x) - \cancel{y\log(1+e^{\theta x})} \\
&\quad - \log(1+e^{\theta x}) + \cancel{y\log(1+e^{\theta x})} \\
-f(x) &= xy\theta - \log(1+e^{\theta x}) \\
f(x) &= \log(1+e^{\theta x}) - xy\theta
\end{aligned}
$$



$$= \frac{x}{1+e^{-\theta x}} - xy$$

$$\frac{\partial^2 f}{\partial \theta^2} = x \cdot \frac{(-1)}{(1+e^{-\theta x})^2} (e^{-\theta x}) (-x)$$

$$= \frac{x^2 e^{-\theta x}}{(1+e^{-\theta x})^2} \geq 0 \quad \forall x$$

Figure 9: Double derivative of log loss

Theta: co-efficient of independent variable “x”.

As seen in the final expression(double derivative of log loss function) the squared terms are always ≥0 and also, in general, we know the range of **e^x** is **(0, infinity).** ***Hence the final term is always ≥0 implying that the log loss function is convex in such scenarios !!***

Final thoughts:

We hope this post was able to make you understand the cons of using MSE as a loss function in logistic regression. If you have any thoughts, comments or questions, please leave a comment below or contact us on LinkedIn and don’t forget to click on 👏 if you like the post.

Rajesh Shreedhar Bhat - Data Scientist - WalmartLabs India | LinkedIn

View Rajesh Shreedhar Bhat's profile on LinkedIn, the world's largest professional community.

Souradip Chakraborty - Statistical Analyst - Walmart Labs India | LinkedIn

View Souradip Chakraborty's profile on LinkedIn, the world's largest professional community.

www.linkedin.com

References:

Convex function

In mathematics, a real-valued function defined on an n-dimensional interval is called convex (or convex downward or…

en.wikipedia.org

Logistic Regression Machine Learning Data Science Convex F Towards Data Science