Assignment 2

September 14, 2023

Tingyu Wu
who is DYING

0.1 Problem 3

Why don't use MSE as loss function for classification?

0.2 Solution 3

AS for MSE we have:

$$\mathcal{L} = \frac{1}{m}(y - \hat{y})^2$$

In the problem of classification, we normally use Logistic Regression or Softmax Regression. Take Logistic Regression as an example, our fitting function will be:

$$f(z) = \frac{1}{1 + e^{-z}}$$

If we are going to use MSE as the loss function, it will be:

$$\mathcal{L}(w) = \left(y - \frac{1}{1 + e^{-(wx)}}\right)^2$$

In order to minimize this loss function above, normally we will use Gradient Descent. Let's now take derivative:

$$\mathcal{L}'(w) = (f(z) - y)f'(z)x$$

From the above Eq we could realize that **the loss function using MSE is not convex** since the function $f(z) = \frac{1}{1+e^{-z}}$ is not convex, which will cause the problem of only finding local minimum rather than absolute minimum.

On the other hand, if we use MLE as the loss function:

$$J(w) = -[y \ln(f(z)) + (1 - y)\ln(1 - f(z))]$$

Take the derivative:

$$J'(w) = \frac{f(z) - y}{f(z)(1 - f(z))} f'(z)x$$

For the function $f(z)$,

$$f'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} = f(z)(1 - f(z))$$

Thus,

$$J'(w) = (f(z) - y)x$$

Comparing $\mathcal{L}'(w)$ to $J'(w)$, we find out that $\mathcal{L}'(w)$ have one more term, $f'(z)$, which reaches maximum $\frac{1}{4}$ when $z = 0$. Therefore, **the velocity of convergence would be larger with MLE**.

0.3 Problem 4

What's the relationship between log-odds and logistics, what's the relationship between log-odds and self-information? Interpret the result you get. (l0g-odds is $\log\left(\frac{p}{1-p}\right)$)

0.4 Solution 4

- $\log\left(\frac{p}{1-p}\right)$

Let's directly prove that sigmoid(log-odds)=p.

$$\exp\left\{-\log\left(\frac{p}{1-p}\right)\right\} = \frac{1-p}{p}$$

Hence,

$$\frac{1}{1+\exp\{-z\}} = \frac{1}{1+\frac{1-p}{p}} = p$$

Therefore, we say that log-odds is the inverse function of sigmoid.

- $I = \log\frac{1}{p(x)}$

$$\begin{aligned} \text{log-odds}(x) &= \log\left(\frac{p}{1-p}\right) \\ &= \log(p) - \log(1-p) \\ &= I(1-p) - I(p) \end{aligned}$$

In this situation, log-odds can be used to show the difference between the self-information of whether an event is happened, or not.

0.5 Problem 6

Prove KL Divergence is non-negative.

0.6 Solution 6

According to KL Divergence:

$$\begin{aligned} D_P(Q) = D_{KL}(Q\|P) &= H_P(Q) - H_Q(P) \\ &= \sum_x Q(x)\log\frac{1}{P(x)} - \sum_x Q(x)\log\frac{1}{Q(x)} \\ &= \sum_x Q(x)\log\frac{Q(x)}{P(x)} \end{aligned}$$

Jensen's Inequality tells us: for any real function $f(x)$ which is convect on the interval I, the below inequality is satisfied:

$$f\left(\sum_{i=1}^{N} p_i x_i\right) \leq \sum_{i=1}^{N} p_i f(x_i)$$

while $p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$. Also, $\log(x)$ is a convex function.

Thus,

$$
\begin{aligned}
D_P(Q) &= -\sum_x Q(x) \log \frac{P(x)}{Q(x)} \\
&\leq -\log\left(\sum_x Q(x) \frac{P(x)}{Q(x)}\right) \\
&= -\log\left(\sum_x P(x)\right) \\
&= -\log(1) \\
&= 0
\end{aligned}
$$