

为什么Binary Classification 不用MSE ?

当我们优化一个binary classification 的时候，我们的输出实际为 y_i 的概率，当 $y_i = 1$ 的时候， $output$ 越大越好。反之，当 $y_i = 0$ 的时候， $output$ 越小越好。所以理论上来说，用MSE:

$$MSE = y_i(1 - p)^2 + (1 - y_i)(0 - p)^2$$

也行的通。但是为什么呢？

我搜索了一下并翻了很多答案，我觉得比较正确和靠谱的解释是，回归模型和分类模型的假设不一样。

回归模型

在回归模型里面，我们假设 $P(y_i|x_i) \sim N(h(x), \sigma)$ ，因此，回归模型的log maximum likelihood 会是：

$$\log L = \sum_{i=1}^n -\frac{(y_i - h(x_i))^2}{2\sigma^2} - n\log(\sigma) - \frac{n}{2}\log 2\pi$$

其余的 $2\sigma^2, n\log\sigma, n/2 \times \log 2\pi$ 都是常数，而我们模型唯一能优化的就是 $-(y_i - h(x))^2$ ，正好等于负的MSE，也就是说，当MSE越小，模型的MLE就能越大。

分类模型

而在分类模型里面，我们假设 $P(y_i|x_i)$ 服从伯努利分布，而非正态分布，即 $P(y_i|x_i) = h(x)^{y_i}(1 - h(x))^{1-y_i}$ ，而分布模型的log MLE会是：

$$\log L = \sum_{i=1}^n y_i \log(h(x_i)) + (1 - y_i) \log(1 - h(x_i))$$

而公式中间部分正好的负的BCE loss

$$BCE_{loss} = -y_i \log(h(x_i)) + (1 - y_i) \log(1 - h(x_i))$$

也就是说，当BCE越小，MLE会越大。

如果用MSE来评估分类问题的loss

$$MSE = y_i(1 - h(x_i))^2 + (1 - y_i)h(x_i)^2$$

$$MSE' = \begin{cases} 2h(x_i) - 2, y_i = 1 \\ 2h(x_i), y_i = 0 \end{cases}$$

而实际的MLE是：

$$MLE' = \begin{cases} \frac{1}{h(x_i)}, y_i = 1 \\ \frac{1}{h(x_i)-1}, y_i = 0 \end{cases}$$

一个是线性函数一个是反比例函数，换句话说，当 $h(x_i)$ 增加的时候，MSE无论何时都会同比例的减少，但是对于MLE, 当 $h(x_i)$ 比较大的时候，再增加对模型的改进是有限的。

在具体一点来说，假如目前目标时1，模型的输出经过一轮后由0.5->0.6, 另一个改进是从0.8->0.9, 对于regression 问题，这两个改进的幅度是一样的，模型都进步了0.1。但是对于classification 的问题，0.5->0.6是很大的提升，也就是将一个不太确定的目标改为一个有点确定的目标，而0.8->0.9这个改进时不大的，只是将一个已经比较确定的目标该问更确定了而已，这个改进是不如0.5->0.6的。

如果模型采用MSE作为评估，那么在训练的时候，0.9 的导数会比0.5的大($2h(x_i)$)，也就是模型没有把注意力放在最需要改进的地方。而如果采用BCE, 0.9 的导数比0.5($1/h(x)$)的要小，模型会将注意力放在难样本上面，符合我们的期待。