问题

线性回归损失函数为什么要用平方形式？

问题背景

这是在阿里一面中遇到的问题，当时我的回答是损失函数是是模型预测值与真实值之间的一种距离度量，我们可以计算出每个样本的预测值与真实值之间的距离，全部加起来就得到了所谓的损失函数。而距离的度量可以采用预测值与真实值之间差的绝对值，或者两者之差的平方，当然更高次的也行，只要你喜欢。正如问题所述，为什么我们一般使用的是两者之差的平方而不是两者只差的绝对值呢？其实这与模型的求解相关，举最简单的线性回归为例，如果采用的距离是两者之差的绝对值，那么求解的目标函数如下：

$$(\omega^*, b) = argmin_{(\omega,b)} \sum_{i=1}^{m} |f(x_i) - y_i| \tag{1}$$

如果采用的距离是两者之差的平方，那么求解的目标函数如下：

$$(\omega^*, b) = argmin_{(\omega,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \tag{2}$$

其中：$f(x_i) = \omega x_i + b$ 即预测值，y_i 为真实值，m 为样本总数，ω 和 b 为要求解的参数

要求得使以上损失函数最小化对应的那个 ω 和 b ，可将损失函数对 ω 和 b 求导，并令导数为0。但是当采取的距离是两者之差的绝对值时，函数在0处不可导，且还增加的一个工作量是需要判断 $f(x_i) - y_i$ 正负号。而采用的距离是两者之差的平方时就没有这方面的问题，所以解决回归问题的时候一般使用平方损失。但理论上两者都可以使用，只是如果用两者之差的绝对值的话，那么需要判断和处理的东西多点，例如人为设定0处的导数为0等等。

但其实这样的回答是不对的，下面给出一个大佬的解答：

问题解答

其实是**因为最小化平方误差本质上等同于在误差服从高斯分布的假设下的最大似然估计。**（这句话确实不好理解，我也理解不了，就先这么记着吧，说不定啥时候就理解了呢.....）

线性回归，简单点来说就是对于训练数据样本 (x_i, y_i) ，预测值 $\hat{y_i} = \theta_0 + \theta_1 * x_i$ ，并构建损失函数：

$$J = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{3}$$

损失函数表示的是每个训练数据点 (x_i, y_i) 到拟合直线 $\hat{y_i} = \theta_0 + \theta_1 * x_i$ 的竖直距离的平方和，通过最小化这个损失函数来求得拟合直线的最佳参数 θ 。那么损失函数为什么要用均方误差的形式，而不是绝对值形式，一次方，三次方，或者四次方形式呢？

简单来说，是因为使用平方形式的时候，使用的是“最小二乘法”的思想，这里的二乘指的是平方的形式来度量预测值与真实值之间的距离，“最小”指的是求得的最佳参数 θ 要保证预测值与真实值之间距离的平方和最小。

最小二乘法以预测值与真实值距离的平方和作为损失函数，在误差服从正态分布的前提下，与极大似然估计的思想在本质上是相同的（哈哈哈不太理解.....）。

我们设真实值与预测值之间的误差为：

$$\varepsilon_i = y_i - \hat{y_i} \tag{4}$$

我们通常认为误差 ε 服从标准正态分布$(\mu = 0, \sigma^2 = 1)$，即给定一个 x_i，模型输出真实值为 y_i 的概率为：

$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi}} * exp(-\frac{\varepsilon_i^2}{2}) \tag{5}$$

进一步我们假设数据集中N个样本点之间相互独立，则给定所有 x 输出所有真实值 y 的概率即似然Likeihood，为所有 $p(y_i|x_i)$ 的累乘：

$$L(x, y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} * exp[-\frac{\varepsilon_i^2}{2}] \tag{6}$$

取对数似然函数得：

$$log[L(x, y)] = -\frac{n}{2} log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \varepsilon_i^2 \tag{7}$$

去掉与 $\hat{y_i}$ 无关的第一项，然后转化为最小化负对数似然：

$$neg\_log[L(x, y)] = \frac{1}{2} \sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{8}$$

可以看到这个实际上就是均方差损失的形式。也就是说**在模型输出与真实值的误差服从高斯分布的假设下，最小化均方差损失函数与极大似然估计本质上是一致的**，因此在这个假设能被满足的场景中（比如回归），均方差损失是一个很好的损失函数选择；当这个假设没能被满足的场景中（比如分类），均方差损失不是一个好的选择，这也解释了为什么在分类问题中不使用均方误差作为损失函数而是使用交叉熵的问题。

当然回答上面下划线部分的问题还可以举的一个例子是：假设有一个样本 $[x_1]$ 的真实标签为 $[0, 0, 1]$，那么预测得到概率中第三类的概率最大即说明分类正确，例如为 $[0.2, 0.2, 0.6]$，但是平方误差却过于严格，比如当预测结果是 $[0, 0.4, 0.6]$，虽然两者在交叉熵上的结果是一样的，但是平方误差中却差别很大。

参考资料

1、线性回归损失函数为什么要用平方形式
2、机器学习常用损失函数小结

By Yee

2020.05.12