机器学习常见知识汇总之优化算法

田晓彬
xiaobin9652@163.com

损失函数

在有监督学习中，损失函数刻画了模型和训练样本的匹配程度。

令训练样本为(x_i, y_i)，其中$x_i \in \boldsymbol{X}$，为第i个训练样本；y_i为第i个训练样本的真实标签。模型可以表示为$f(\theta)$,其中模型关于第i个样本的输出记为$f(x_i|\theta)$。由此我们可以定义损失函数为$L(f(x_i|\theta), y_i)$，损失函数越小，表明模型在该样本点匹配得越好。

1. 0-1损失函数

对于二分类问题，y_i取值1或者-1，我们总是希望$sign(f(x_i|\theta)) = y_i$。0-1损失可以表示为

$$L_{0-1}(f, y) = 1_{fy \leq 0}$$

其中$1_{fy \leq 0}$是指示函数，当且仅当$fy \leq 0$为真时取值为1，否则取值为0。该函数可以直观地刻画分类的错误率，但是由于其非凸、非光滑的特点，使得算法很难直接对该函数进行优化。

2. Hinge损失函数

Hinge损失函数定义如下：

$$L_{hinge}(f, y) = \max\{0, 1 - fy\}$$

Hinge损失函数是0-1损失函数相对紧的凸上界，且当$fy \geq 1$时，该函数不对其做任何惩罚。因为Hinge函数在$fy = 1$处不可导，因此不能使用梯度下降法进行优化，而是使用次梯度下降法。

3. Logistic损失函数

Logistic损失函数定义如下：

$$L_{logistic}(f, y) = \log_2(1 + \exp(-fy))$$

Logistic函数也是0-1函数的凸上界，且该函数处处光滑，因此可以使用梯度下降算法进行优化。但是该损失函数对所有的样本点都有惩罚，因此对异常值相对来说更敏感。

4. 交叉熵损失函数

交叉熵损失函数定义如下：

$$L_{cross_entropy}(f,y)=-\log_2\left(\frac{1+fy}{2}\right)$$

交叉熵损失函数也是0-1损失函数的光滑凸上界。

5. 平方损失函数

对于回归问题，我们总是希望$f(x_i,\theta)\approx y_i$，最常用的损失函数是平方损失函数

$$L_{square}(f,y)=(f-y)^2$$

平方损失函数是光滑函数，能够用梯度下降算法进行优化。然而，当预测值距离真实值越远时，平方损失函数的惩罚力度越大，因此平方损失函数对异常点较为敏感。

5. 绝对损失函数

为了解决平方损失函数对异常点敏感的问题，引入绝对损失函数：

$$L_{absolute}(f,y)=|f-y|$$

绝对损失函数相当于是在做中值回归，相比做均值回归的平方损失函数，绝对损失函数对异常点更鲁棒一些。但是绝对损失函数在 $f=y$ 处无法求导数。

6. Huber损失函数

Huber损失函数定义如下：

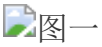
$$L_{Huber}(f,y)=\begin{cases}(f-y)^2, & |f-y|\le\delta\\ 2\delta|f-y|-\delta^2, & |f-y|\ge\delta\end{cases}$$

Huber损失函数综合考虑了可导性和对异常点的鲁棒性，在$|f-y|$较小时为平方损失，在$|f-y|$较大时为线性损失，并且处处可导，对异常点鲁棒性较好。

L1正则化与稀疏性

L1正则化可以给解带来稀疏性，为什么需要有稀疏性的解呢？模型解的稀疏性越大，说明模型参数中0的个数越多。这相当于对模型进行了一次特征选择，只留下比较重要的特征，从而提高模型的泛化能力，降低过拟合的风险。

在机器学习中，有正则化项和约束条件两个概念。实际上，正则化项和约束条件是等价的，L1正则化可以转换成约束条件$||w||_1 \leq \mu$，其中μ是一个小值。由此可以看出，在二维的情况下，L1正则化实际上为参数定义了一个菱形的解空间。


图一

如上图所示，多边形为L1正则化的解空间，等高线为凸优化问题中目标函数的等高线。由图可以看出，L1正则化的多边形解空间，更容易在尖角处与等高线相交，从而得到稀疏解。

L2正则化

同L1正则化我们可以得出L2正则化转换成的约束条件为$||w||_2^2 \leq \mu$，其中μ是一个小值。由此可以看出，在二维的情况下，L2正则化实际上为参数定义了一个圆形的解空间。


图二

如上图所示，圆形为L2正则化的解空间，等高线为凸优化问题中目标函数的等高线。由图可以看出，正则项系数越大，解空间的半径越小，模型参数的值也会越小。而在模型的学习中，通常都倾向于让权值尽可能小，最后构造一个所有参数都比较小的模型。因为一般认为参数值小的模型比较简单，在一定程度上避免过拟合现象。