带答案面经分享-L1正则&L2正则

2019-07-04 阅读 1.5K

作者：石晓文
来源：小小挖掘机

正则化也是校招中常考的题目之一，在去年的校招中，被问到了多次：

1、过拟合的解决方式有哪些，l1和l2正则化都有哪些不同，各自有什么优缺点(爱奇艺)

2、L1和L2正则化来避免过拟合是大家都知道的事情，而且我们都知道L1正则化可以得到稀疏解，L2正则化可以得到平滑解，这是为什么呢?

3、L1和L2有什么区别，从数学角度解释L2为什么能提升模型的泛化能力。（美团）

4、L1和L2的区别，以及各自的使用场景（头条）

接下来，咱们就针对上面的几个问题，进行针对性回答！

1、什么是L1正则&L2正则?

L1正则即将参数的绝对值之和加入到损失函数中，以二元线性回归为例，损失函数变为：

$$min\frac{1}{2m}\sum_{i=1}^{n}(h_w(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{2}|w_j|$$

L2正则即将参数的平方之和加入到损失函数中，以二元线性回归为例，损失函数变为：

$$min\frac{1}{2m}\sum_{i=1}^{n}(h_w(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{2}w_j^2$$

2、L1正则&L2正则的区别是什么?

二者的区别的话，咱们总结主要有以下两点，最主要的还是第二点：

1、L1正则化是指在损失函数中加入权值向量w的绝对值之和，即各个元素的绝对值之和，L2正则化指在损失函数中加入权值向量w的平方和。

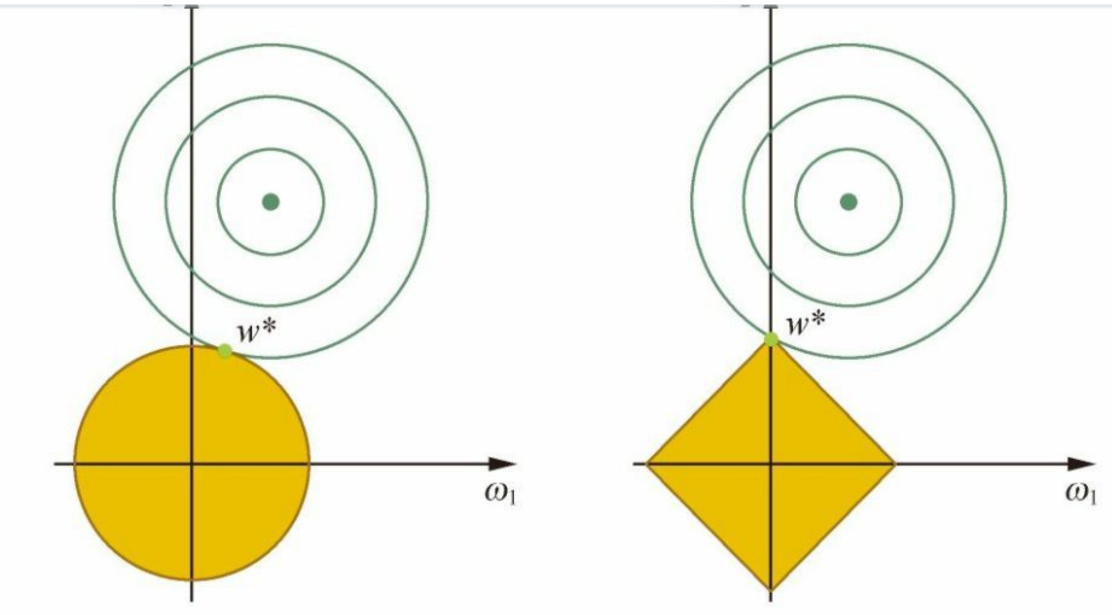
2、L1的功能是使权重稀疏，而L2的功能是使权重平滑。

3、L1正则为什么可以得到稀疏解?

这一道题是面试中最容易考到的，大家一定要理解掌握！这一部分的回答，在《百面机器学习》中给出了三种答案：

3.1 解空间形状

这是我们最常使用的一种答案，就是给面试官画如下的图：



（a）L2 正则化对应的解空间　　（b）L1 正则化对应的解空间

L2正则化相当于为参数定义了一个圆形的解空间，而L1正则化相当于为参数定义了一个菱形的解空间。L1“棱角分明”的解空间显然更容易与目标函数等高线在脚点碰撞。从而产生稀疏解。

3.2 函数叠加

我们考虑一维的情况，横轴是参数的值，纵轴是损失函数，加入正则项之后，损失函数曲线图变化如下：

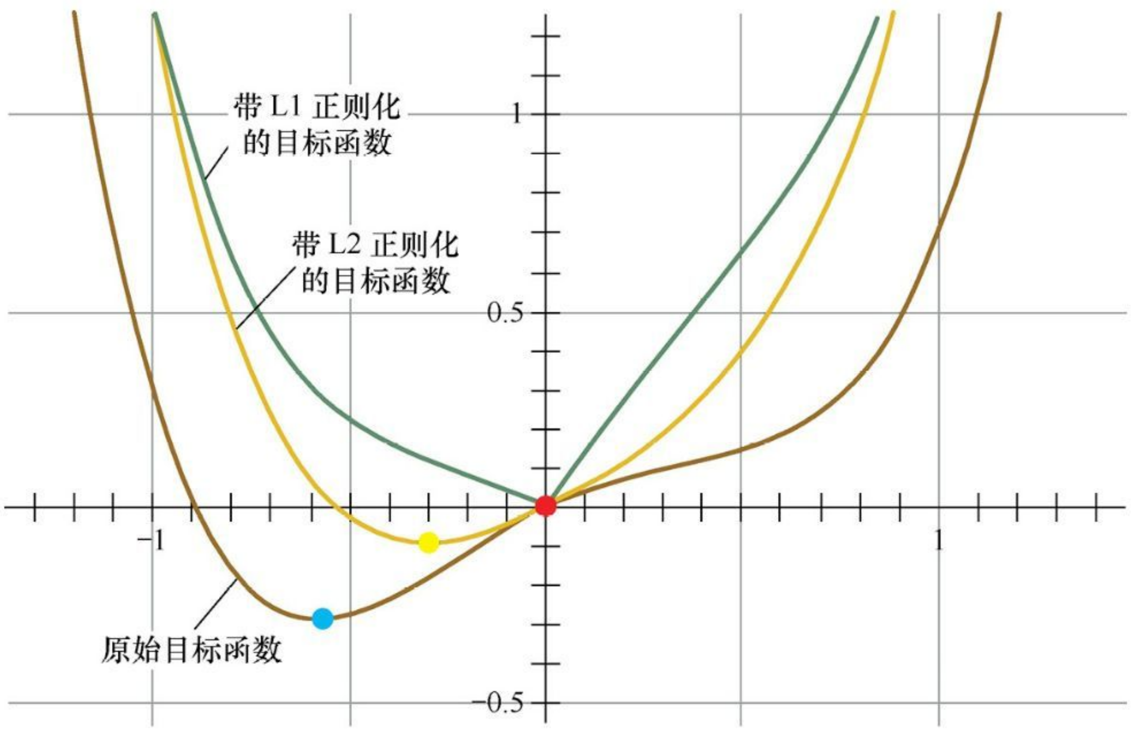


图7.7　函数曲线图

可以看到，在加入L1正则项后，最小值在红点处，对应的w是0。而加入L2正则项后，最小值在黄点处，对应的w并不为0。

为什么呢？加入L1正则项后，目标函数变为L(w)+C|w|，单就正则项部分求导，原点左边的值为-C，原点右边的值为C，因此，只要原目标函数的导数绝对值|L'(w)|<C,那么带L1正则项的目标函数在原点左边部分始终递减，在原点右边部分始终递增，最小值点自然会出现在原点处。

3.3 贝叶斯先验

从贝叶斯角度来看，L1正则化相当于对模型参数w引入了拉普拉斯先验，L2正则化相当于引入了高斯先验(为什么我们在后面详细解释)。我们来看一下高斯分布和拉普拉斯分布的形状：

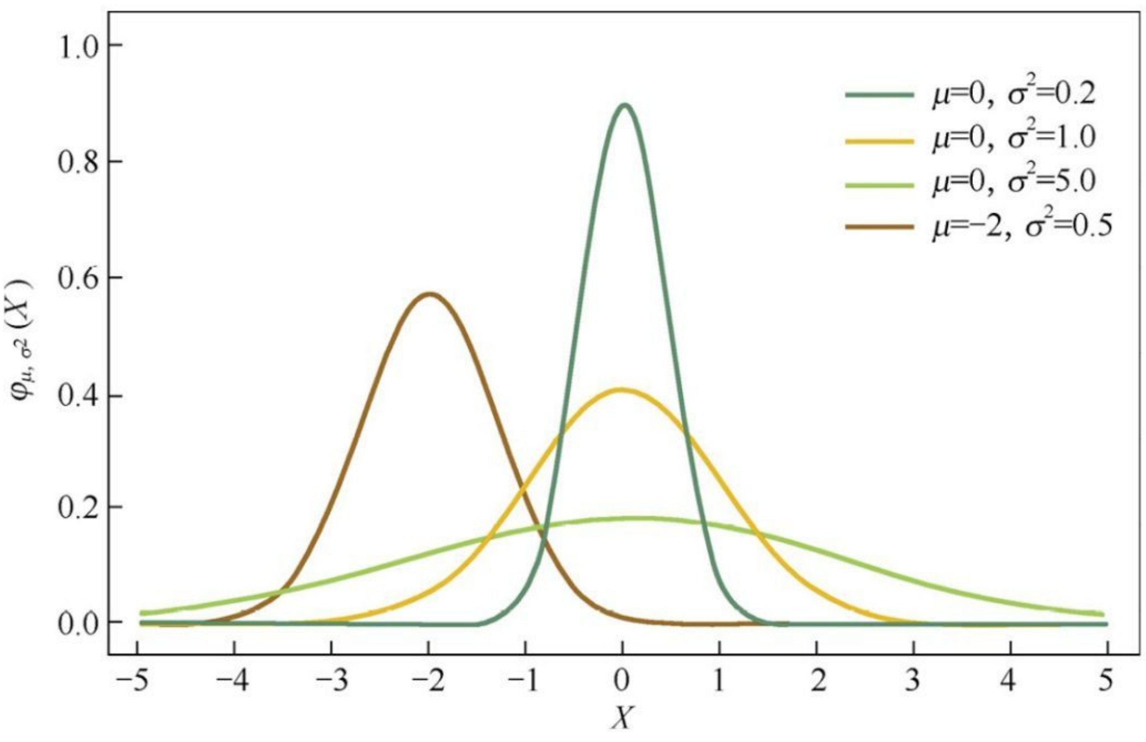


图7.8　高斯分布曲线图

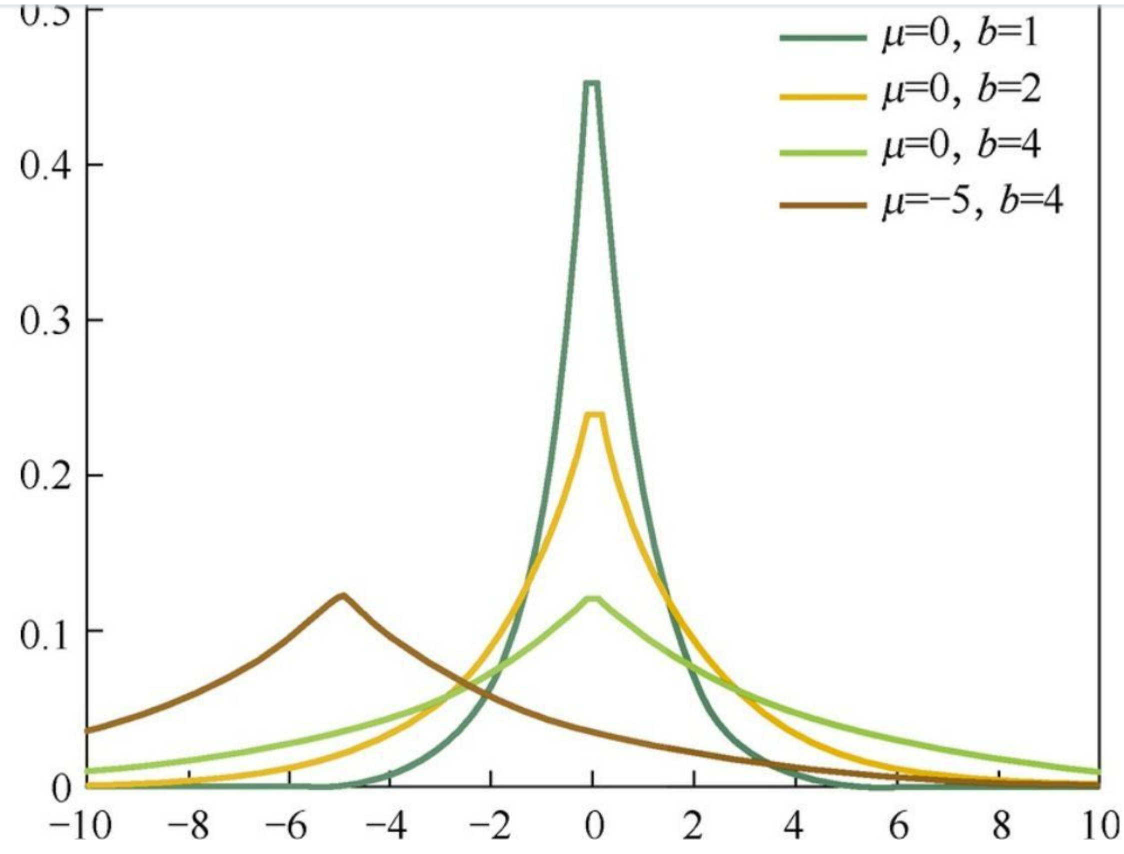


图7.9 拉普拉斯分布曲线图

可以看到，当均值为0时，高斯分布在极值点处是平滑的，也就是高斯先验分布认为w在极值点附近取不同值的可能性是接近的。但对拉普拉斯分布来说，其极值点处是一个尖峰，所以拉普拉斯先验分布中参数w取值为0的可能性要更高。

4、从数学角度解释L2为什么能提升模型的泛化能力

这里主要给出两篇博客作为参考：

https://www.zhihu.com/question/35508851

https://blog.csdn.net/zouxy09/article/details/24971995

5、为什么说“L1正则化相当于对模型参数w引入了拉普拉斯先验，L2正则化相当于引入了高斯先验”？

这一部分咱们小小推导一下，嘻嘻，如果一看数学就头大的同学，可以跳过此处。

在贝叶斯估计中，我们要求解的是参数θ的后验概率最大化：

$$P(\theta|X,Y)=\frac{P(Y,X,\theta)}{P(X,Y)}=\frac{P(Y|X,\theta)P(X,\theta)}{P(X,Y)}$$

$$=\frac{P(Y|X,\theta)P(X|\theta)P(\theta)}{P(X,Y)}=\prod_{i=1}^{m}\frac{P(Y_i|X_i,\theta)P(X_i|\theta)P(\theta)}{P(Y_i,X_i)}$$

在最后一项的分子中P(Xi|θ)和分母都是一个常数，因此，上式可以继续化简：

$$P(\theta|X,Y)=\prod_{i=1}P(Y_i|X_i,\theta)P(\theta)=\sum_{i=1}lnP(Y_i|X_i,\theta)+mlnP(\theta)$$

所以贝叶斯学派估计是使下面的式子最小化：

$$argmin_{\theta}-[\sum_{i=1}^{m}ln\,P(Y_i|X_i,\theta)+m\;lnP(\theta)]$$

关于第一项，假设我们做的是一元线性回归，那么求解过程如下：

极大似然法

真实值=估计值+误差

那么我们现在就从极大似然估计的角度来看一下线性回归的本质。现在我们假设 $Y_i=X_i^T\Theta+\epsilon_i=\hat{Y}_i+\epsilon_i$。这个式子中 ϵ 代表着误差。且 $\epsilon_i\sim\mathcal{N}(0,\sigma^2)$。这个条件也就解释了为什么线性回归是 **高斯模型** 的。

现在我们来看一下我们要求的 $P(Y_i|X_i,\Theta)$，这个先验概率表达的是什么呢？就是给定了一组样本 X_i，然后我们采用参数集 Θ 进行加权估计最终得到正确答案 Y_i 的概率。那么这个时候的误差是什么呢？给定了 X_i 和 Θ，那么也就说明误差 $\epsilon_i=Y_i-\hat{Y}_i$。

所以 $P(Y_i|X_i,\Theta)=P(\epsilon_i=Y_i-X_i^T\Theta)$。根据高斯分布的公式，我们可以得到一下结论：

这里相等是因为 x_i 和 Θ 都可以认为是定值，此时误差也就确定了

$$P(Y_i|X_i,\Theta)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(Y_i-X_i^T\Theta)^2}{2\sigma^2}}$$

因为 X_i 是相互独立的，所以：

$$P(Y|X,\Theta)=\prod_{i=1}^{m}P(Y_i|X_i,\Theta)=\prod_{i=1}^{m}P(\epsilon i=Y_i-X_i^T\Theta)$$

同时取对数后，再根据对数公式进行化简得到：

$$logP(Y|X,\Theta)=m*log(\frac{1}{\sqrt{2\pi\sigma}})-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(Y_i-X_i^T\Theta)^2$$

第二项，咱们就得分类讨论了，如果θ服从的是0均值的高斯分布，为了和上面的方差所区分，这里咱们用alpha来表示，那么有：

$$m\;lnP(\theta)=mln\frac{1}{\alpha\sqrt{2\pi}}e^{-\frac{\theta^2}{2\alpha^2}}=mln\frac{1}{\alpha\sqrt{2\pi}}-m\frac{\theta^2}{2\alpha^2}$$

所以，最终可以得到：

$$argmin_{\theta}-\left[\sum_{i=1}^{m}ln\,P(Y_i|X_i,\theta)+m\;lnP(\theta)\right]$$

$$=argmin_{\theta}-[\,mln\frac{1}{\sigma\sqrt{2\pi}}+mln\frac{1}{\alpha\sqrt{2\pi}}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(Y_i-X_i\theta)^2-m\frac{\theta^2}{2\alpha^2}]$$

我们把与θ无关的情况去掉，便得到：

$$argmin_\theta[\frac{1}{2\sigma^2}\sum_{i=1}(Y_i - X_i\theta)^2 + \frac{m}{2\alpha^2}\theta^2]$$

你可能觉得，alpha不是θ的方差么，请注意，这里是先验分布，我们可以任意指定alpha的值，所以去掉也是可以的。

同理，我们可以得到当先验是拉普拉斯分布时的情况。

$$m\ lnP(\theta) = mln\frac{1}{2b}e^{-\frac{|\theta|}{2b}} = mln\frac{1}{2b} - m\frac{|\theta|}{b}$$

$$argmin_\theta - \left[\sum_{i=1}^{m} ln\,P(Y_i|X_i,\theta) + m\ lnP(\theta)\right]$$

$$= argmin_\theta - [\ mln\frac{1}{\sigma\sqrt{2\pi}} + mln\frac{1}{2b} - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(Y_i - X_i\theta)^2 - m\frac{|\theta|}{b}]$$

$$= argmin_\theta[\frac{1}{2\sigma^2}\sum_{i=1}^{m}(Y_i - X_i\theta)^2 + \frac{m}{b}|\theta|]$$

END

举报