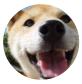
L1正则化与L2正则化



bingo酱

I am a fighter

在腾讯互娱的两面，和百度的一面中，都问到了这个**问题：**

讲讲正则化为什么能降低过拟合程度，并且说明下下L1正则化和L2正则化。

（要想看答案请直接看文章结尾）

L1和L2正则化：

我们所说的正则化，就是在原来的loss function的基础上，加上了一些正则化项或者称为惩罚项。现在我们还是以最熟悉的线性回归为例子。

优化目标：

$$\min \quad 1/N * \sum_{i=1}^{N} (y_i - \omega^T x_i)^2 \qquad 式子（1）$$

加上L1正则项（lasso回归）：

$$\min \quad 1/N * \sum_{i=1}^{N} (y_i - \omega^T x_i)^2 + C||\omega||_1 \qquad 式子（2）$$

加上L2正则项（岭回归）：

$$\min \quad 1/N * \sum_{i=1}^{N} (y_i - \omega^T x_i)^2 + C||\omega||_2 \qquad 式子（3）$$

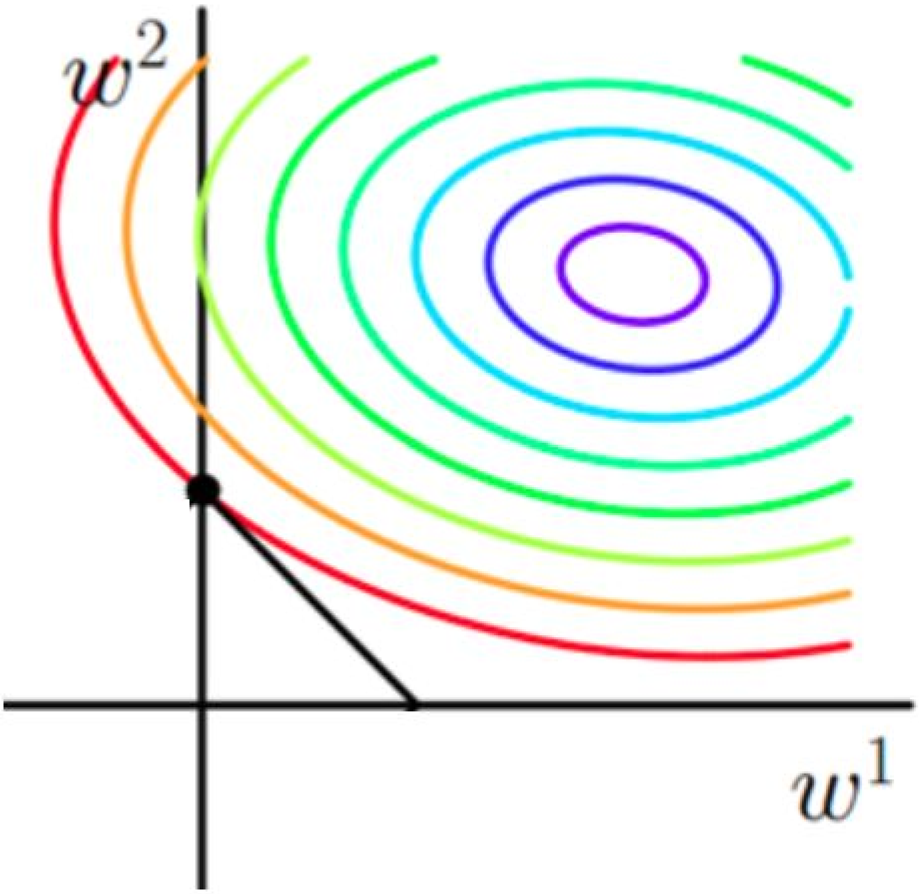
结构风险最小化角度：

结构风险最小化： 在经验风险最小化的基础上（也就是训练误差最小化），尽可能采用简单的模型，以此提高泛化预测精度。

那现在我们就看看加了L1正则化和L2正则化之后，目标函数求解的时候，最终解有什么变化。

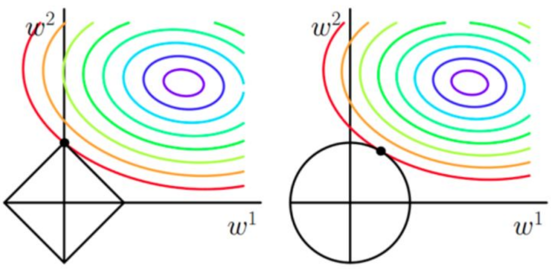
图像解释（假设X为一个二维样本，那么要求解参数 ω 也是二维）：

- 原函数曲线等高线(同颜色曲线上，每一组 ω_1 ， ω_2 带入值都相同)



目标函数等高线

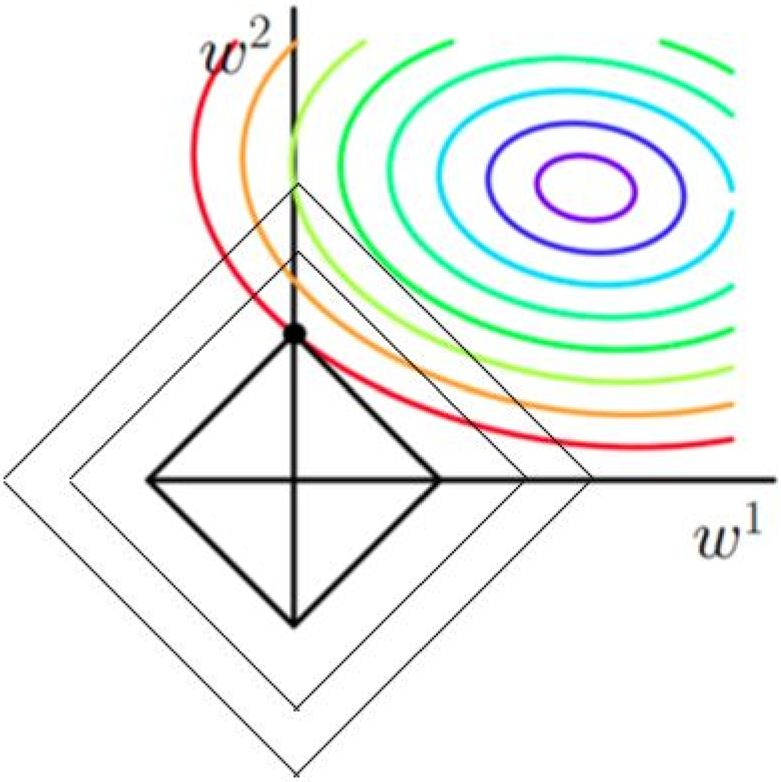
- L1和L2加入后的函数图像：



加入L1和L2正则的等高线

从上边两幅图中我们可以看出：

- 如果不加L1和L2正则化的时候，对于线性回归这种目标函数凸函数的话，我们最终的结果就是最里边的紫色的小圈圈等高线上的点。
- 当加入L1正则化的时候，我们先画出 $|\omega_1| + |\omega_2| = F$ 的图像，也就是一个菱形，代表这些曲线上的点算出来的 1范数 $|\omega_1| + |\omega_2|$ 都为F。那我们现在的目标是不仅是原曲线算得值要小（越来越接近中心的紫色圈圈），还要使得这个菱形越小越好（F越小越好）。那么还和原来一样的话，过中心紫色圈圈的那个菱形明显很大，因此我们要取到一个恰好的值。那么如何求值呢？



带L1正则化的目标函数求解

1. 以同一条原曲线目标等高线来说，现在以最外圈的红色等高线为例，我们看到，对于红色曲线上的每个点都可以做一个菱形，根据上图可知，当这个菱形与某条等高线相切（仅有一个交点）的时候，这个菱形最小，上图相割对比较大的两个菱形对应的1范数更大。

   用公式说这个时候能使得在相同的 $1/N * \sum_{i=1}^{N}(y_i - \omega^T x_i)^2$ 下，

   由于相切的时候的 $C||\omega||_1$ 小，即 $|\omega_1| + |\omega_2|$ 小，所以：

   能够使得 $1/N * \sum_{i=1}^{N}(y_i - \omega^T x_i)^2 + C||\omega||_1$ 更小。

2. 有了1.的说明，我们可以看出，最终加入L1范数得到的解，一定是某个菱形和某条原函数等高线的切点。现在有个比较重要的结论来了，**我们经过观察可以看到，几乎对于很多原函数等高曲线，和某个菱形相交的时候及其容易相交在坐标轴（比如上图），也就是说最终的结果，解的某些维度及其容易是0，比如上图最终解是 $\omega =$ （0，x） ，这也就是我们所说的L1更容易得到稀疏解（解向量中0比较多）的原因。**

3. 当然了，光看着图说，L1的菱形更容易和等高线相交在坐标轴，一点都没

说服力，只是个感性的认识，不过不要紧，其实是很严谨的，我们直接用求导来证明，具体的证明这里有一个很好的答案了，简而言之就是假设现在

我们是一维的情况下 $h(\omega)=f(\omega)+C|\omega|$ ，其中 $h(\omega)$ 是目标函数， $f(\omega)$

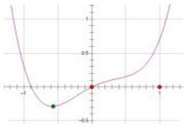
是没加L1正则化项前的目标函数， $C|\omega|$ 是L1正则项，那么要使得0点

成为最值可能的点，虽然在0点不可导，但是我们只需要让0点左右的导数

异号，即 $h'_{左}(0)*h'_{右}(0)=$ （$f'(0)+C$） $(f'(0)-C)<0$ 即可

也就是 $C>|f'(0)|$ 的情况下，0点都是可能的最值点。

l1 相比于 l2 为什么容易获得稀疏解?
www.zhihu.com



- 当加入L2正则化的时候，分析和L1正则化是类似的，也就是说我们仅仅是从菱形变成了圆形而已，同样还是求原曲线和圆形的切点作为最终解。当然与L1范数比，我们这样求的L2范数的**从图上来看，不容易交在坐标轴上，但是仍然比较靠近坐标轴。因此这也就是我们老说的，L2范数能让解比较小（靠近0），但是比较平滑（不等于0）。**

综上所述，我们可以看见，加入正则化项，在最小化经验误差的情况下，可以让我们选择解更简单（趋向于0）的解。

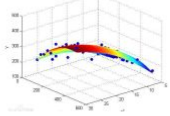
结构风险最小化： 在经验风险最小化的基础上（也就是训练误差最小化），尽可能采用简单的模型，以此提高泛化预测精度。

因此，加正则化项就是结构风险最小化的一种实现。

贝叶斯先验概率的角度：

现在再从贝叶斯学派的观点来看看正则化，即是我们先假设要求的参数服从某种先验分布，以线性回归为例子，我们之前讲过，用高斯分布的极大似然估计求线性回归。

bingo酱：线性回归求解的两种表示（最小化均方误差和基于高斯分布...
zhuanlan.zhihu.com



1. 在我们求解的时候，我们假设Y|X; ω 服从 $N(\omega^T X,\sigma)$ 的正太分布，即概率密度函数 $p(Y|X;\omega)=N(\omega^T X,\sigma)$ ，然后利用极大似然估计求解参数 ω ：

$$\max \quad log\prod_{i=1}^{m}p(y_i|x_i;\omega) \quad \text{式子（4）}$$

或者表示成常用的求极小值：

$$\min \quad -log\prod_{i=1}^{m}p(y_i|x_i;\omega) \quad \text{式子（5）}$$

2. 在贝叶斯学派的观点看来，如果我们先假设参数 ω 服从一种先验分布 $P(\omega)$ ，那么根据贝叶斯公式 $P(\omega|(X,Y)) \sim P(Y|X;\omega) * P(\omega)$ ，那我们利用极大似然估计求参数 ω 的时候，现在我们的极大似然函数就变成了：

$$\max \quad log\prod_{i=1}^{m}p(y_i|x_i;\omega) * p(\omega) = log\prod_{i=1}^{m}p(y_i|x_i;\omega) + log\prod_{i=1}^{m}p(\omega) \quad \text{式子（6）}$$

表示成求极小的情况就是：

$$\min \quad log\prod_{i=1}^{m}p(y_i|x_i;\omega) * p(\omega) = -log\prod_{i=1}^{m}p(y_i|x_i;\omega) - log\prod_{i=1}^{m}p(\omega) \quad \text{式（7）}$$

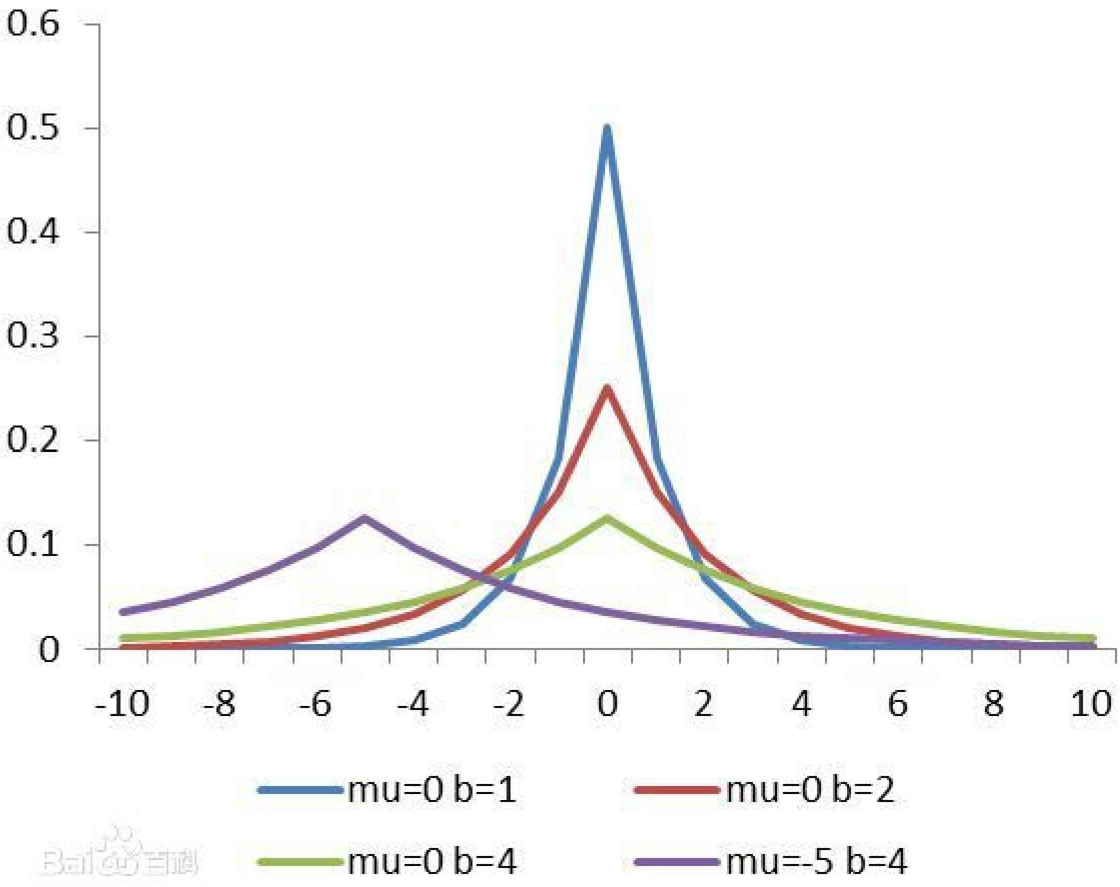
对比式子（5）和式子（7），我们看到，式子（7）比式子（5）多了最后的一个求和项。

L1范数：

假设我们让 ω 服从的分布为标准拉普拉斯分布，即概率密度函数为 $1/2 * exp(-|x|)$ ，那么式子（7）多出的项就变成了 $C||\omega||_1$,其中C为常数了，重写式子（7）：

$$\min \quad log\prod_{i=1}^{m}p(y_i|x_i;\omega) * p(\omega) = -log\prod_{i=1}^{m}p(y_i|x_i;\omega) + C||\omega||_1 \quad \text{式子（8）}$$

熟悉吧，这不就是加了L1范数的优化目标函数么。假设 ω 服从拉普拉斯分布的话，从下图可以看出 ω 的值取到0的概率特别大。也就是说我们提前先假设了 ω 的解更容易取到0。



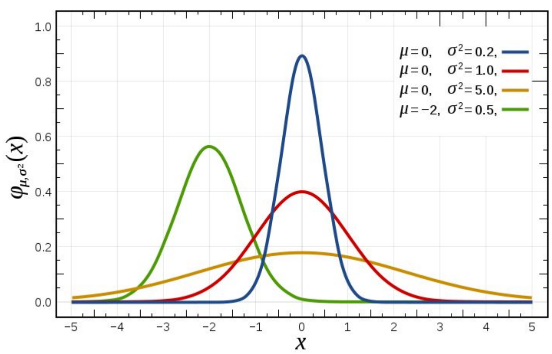
拉普拉斯分布

L2范数：

假设我们让 ω 服从的分布为标准正太分布，即概率密度为 $1/\sqrt{2\pi} * exp(-(x)^2/2)$ ，那么式子（7）多出的项就成了 $C||\omega||_2^2$ ，其中C为常数，重写式子（7）：

$$\min \quad log\prod_{i=1}^{m} p(y_i|x_i;\omega) * p(\omega) = -log\prod_{i=1}^{m} p(y_i|x_i;\omega) + C||\omega||_2^2 \quad 式子（9）$$

熟悉吧，这不就是加了L2范数的优化目标函数么。假设 ω 服从标准正太分布的话，根据图我们可以看出，其实我们就是预先假设了 ω 的最终值可能取到0附近的概率特别大。



高斯分布

因此最后来回答问题：

降低过拟合程度:

正则化之所以能够降低过拟合的原因在于，正则化是结构风险最小化的一种策略实现。

给loss function加上正则化项，能使得新得到的优化目标函数h = f+normal，需要在f和normal中做一个权衡（trade-off），如果还像原来只优化f的情况下，那可能得到一组解比较复杂，使得正则项normal比较大，那么h就不是最优的，因此可以看出加正则项能让解更加简单，符合奥卡姆剃刀理论，同时也比较符合在偏差和方差（方差表示模型的复杂度）分析中，通过降低模型复杂度，得到更小的泛化误差，降低过拟合程度。

L1正则化和L2正则化:

L1正则化就是在loss function后边所加正则项为L1范数，加上L1范数容易得到稀疏解（0比较多）。L2正则化就是loss function后边所加正则项为L2范数，加上L2范数相比于L1范数来说，得到的解比较平滑（不是稀疏），但是同样能够保证解中接近于0（但不是等于0，所以相对平滑）的维度比较多，降低模型的复杂度。