比较全面的L1和L2正则化的解释

原创 石头 机器学习算法那些事 2018-11-07

前言

前段时间写了一篇文章《深入理解线性回归算法（二）：正则项的详细分析》，文章提到L1是通过稀疏参数（减少参数的数量）来降低复杂度，L2是通过减小参数值的大小来降低复杂度。网上关于L1和L2正则化降低复杂度的解释五花八门，易让人混淆，看完各种版本的解释后过几天又全部忘记了。因此，文章的内容总结了网上各种版本的解释，并加上了自己的理解，希望对大家有所帮助。

目录

1、优化角度分析

损失函数：

$$L(\mathrm{w}) = E_D(w) + \lambda E_w(w)$$

模型最优化等价于损失函数的最小化：

$$\min_w L(w) = \min_w (E_D(w) + \lambda E_w(w))$$

1、L2正则化的优化角度分析

L2正则化模型的最优化问题等价于：

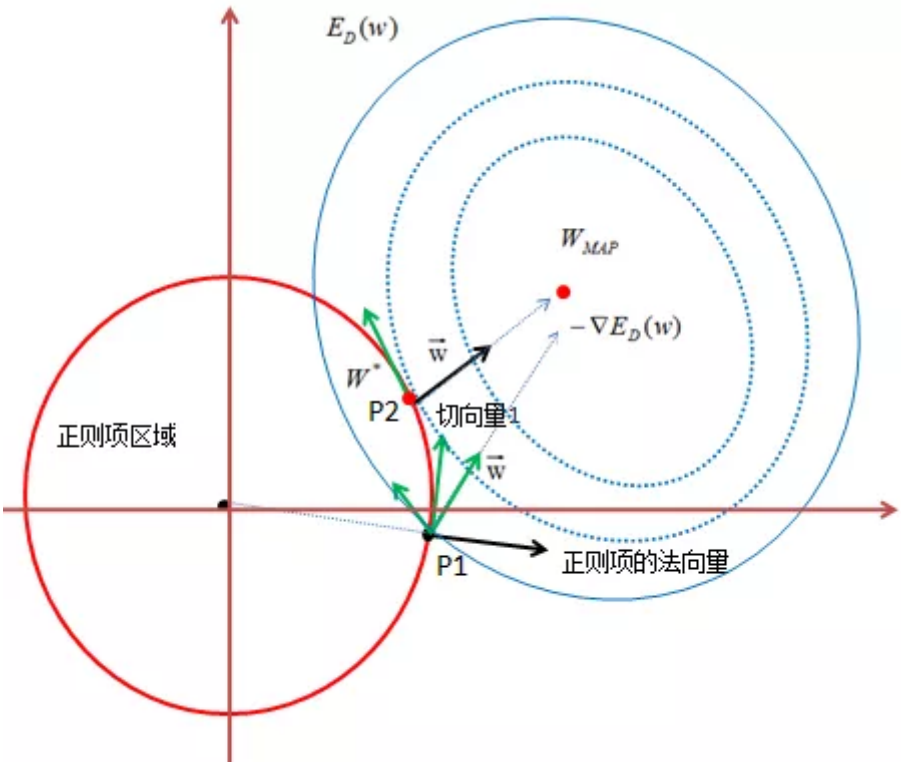
$$\min_w L(w) = \min_w (E_D(w) + \lambda \sum_{i=1}^{n} w_i^2)$$

等价为凸优化问题：

$$\begin{cases} \min\limits_w E_D(w) \\ \sum\limits_{i=1}^{n} w_i^2 \le C, \text{其中} C \text{与正则化参数} \lambda \text{成反比关系} \end{cases}$$

在限定的区域，找到使$E_D(w)$最小的值。

图形表示为：



上图所示，红色实线是正则项区域的边界，蓝色实线是$E_D(w)$的等高线，越靠里的等高圆，$E_D(w)$越小，梯度的反方向是$E_D(w)$减小最大的方向，用$\vec{w}$表示，正则项边界的法向量用实黑色箭头表示。

正则项边界在点P1的切向量有$E_D(w)$负梯度方向的分量，所以该点会有往相邻的等高虚线圆运动的趋势；当P1点移动到P2点，正则项边界在点P2的切向量与$E_D(w)$梯度方向的向量垂直，即该点没有往负梯度方向运动的趋势；所以P2点是$E_D(w)$最小的点。

结论：L2正则化项使$E_D(w)$值最小时对应的参数变小。

2、L1正则化的优化角度分析

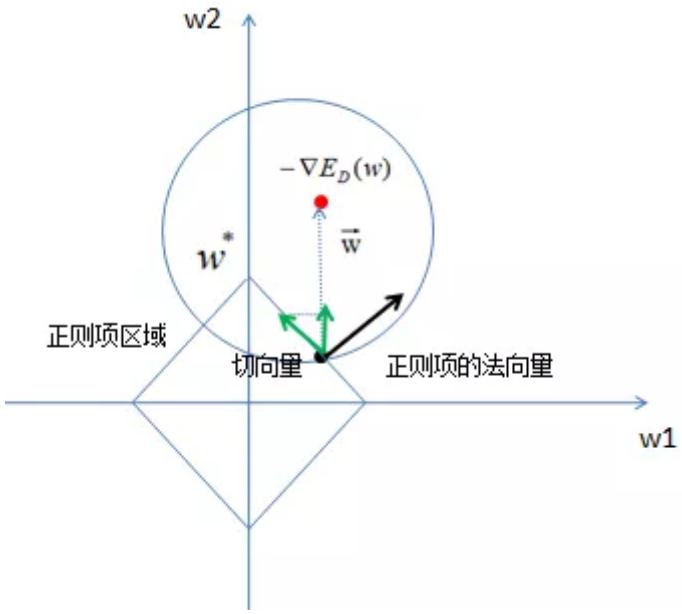
L1正则化模型的最优化问题等价于：

$$\min_w L(w) = \min_w (E_D(w) + \lambda \sum_{i=1}^{n} | w_i |)$$

等价为凸优化问题：

$$\begin{cases} \min\limits_w E_D(w) \\ \sum\limits_{i=1}^{n} | w_i | \le C, \text{其中} C \text{与正则化参数} \lambda \text{成反比关系} \end{cases}$$

在限定的区域，找到使$E_D(w)$最小的值。



结论： 如上图，因为切向量始终指向w2轴，所以L1正则化容易使参数为0，即特征稀疏化。

2、梯度角度分析

1、L1正则化

L1正则化的损失函数为：

$$L(\mathrm{w}) = E_D(w) + \frac{\lambda}{n}\sum_{i=1}^{n}|w_i|$$

求$L(\mathrm{w})$的梯度

$$\frac{\partial L(\mathrm{w})}{\partial w} = \frac{\partial E_D(w)}{\partial w} + \frac{\lambda \operatorname{sgn}(w)}{n}$$

参数w更新：

$$w' = w - \eta\frac{\partial L(\mathrm{w})}{\partial w}$$

$$w' = w - \frac{\eta\lambda \operatorname{sgn}(w)}{n} - \frac{\partial E_D(w)}{\partial w}$$，其中η为学习率

上式可知，当w大于0时，更新的参数w变小；当w小于0时，更新的参数w变大；所以，L1正则化容易使参数变为0，即特征稀疏化。

2、L2正则化

L2正则化的损失函数为：

$$L(w) = E_D(w) + \frac{\lambda}{2\mathrm{n}}\sum_{i=1}^{n}w_i^2$$

求$L(\mathrm{w})$的梯度：

$$\frac{\partial L(w)}{\partial w} = \frac{\partial E_D(w)}{\partial w} + \lambda w$$

参数w更新：

$$\mathrm{w}' = w - \eta\frac{\partial L(w)}{\partial w}$$

$$w' = w - \frac{\eta\lambda w}{n} - \frac{\partial E_D(w)}{\partial w}$$

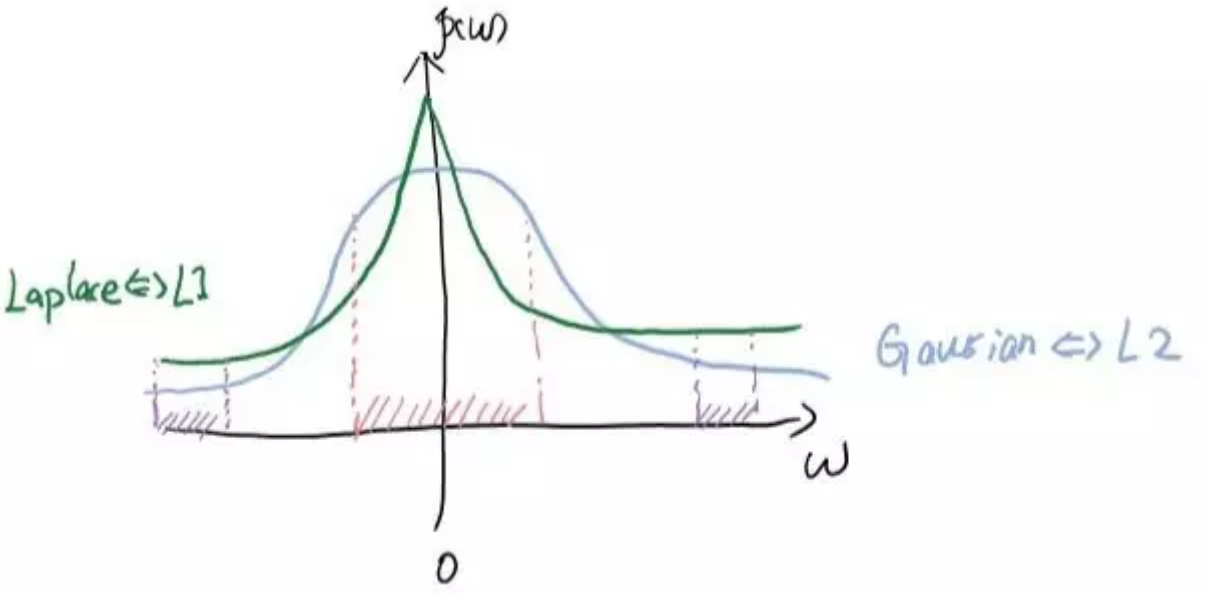
$$w' = (1 - \frac{\eta\lambda}{n})w - \frac{\partial E_D(w)}{\partial w}$$

由上式可知，正则化的更新参数相比于未含正则项的更新参数多了 $\frac{\eta\lambda}{n}w$ 项，当w趋向于0时，参数减小的非常缓慢，因此L2正则化使参数减小到很小的范围，但不为0。

3、先验概率角度分析

文章《深入理解线性回归算法（二）：正则项的详细分析》提到，当先验分布是拉普拉斯分布时，正则化项为L1范数；当先验分布是高斯分布时，正则化项为L2范数。本节通过先验分布来推断L1正则化和L2正则化的性质。

画高斯分布和拉普拉斯分布图（来自知乎某网友）：



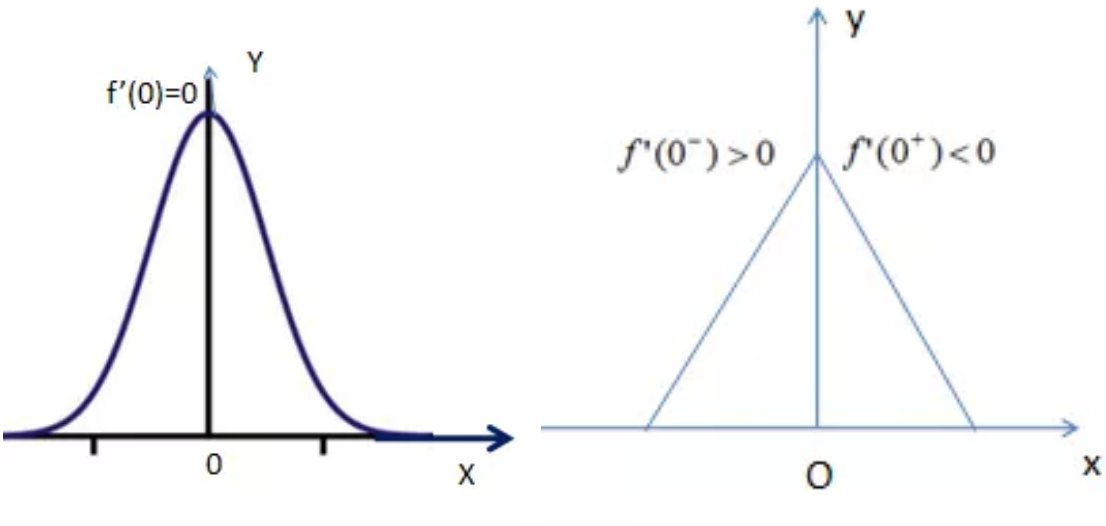
由上图可知，拉普拉斯分布在参数w=0点的概率最高，因此L1正则化相比于L2正则化更容易使参数为0；高斯分布在零附近的概率较大，因此L2正则化相比于L1正则化更容易使参数分布在一个很小的范围内。

4、知乎点赞最多的图形角度分析

函数极值的判断定理：

（1）当该点导数存在，且该导数等于零时，则该点为极值点；
（2）当该点导数不存在，左导数和右导数的符号相异时，则该点为极值点。

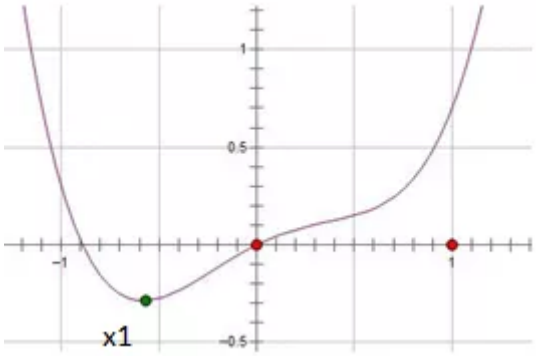
如下面两图：



左图对应第一种情况的极值，右图对应第二种情况的极值。本节的思想就是用了第二种极值的思想，只要证明参数w在0附近的左导数和右导数符合相异，等价于参数w在0取得了极值。

图形角度分析

损失函数L如下：



黑色点为极值点x1，由极值定义：L'(x1)=0；

含L2正则化的损失函数: $f_2(x) = L + Cx^2 (C > 0)$

对$f_2(x)$求导：

$$f_2'(x) = L'(x) + 2Cx$$

令$x = x1$

$$f_2'(x1) = L'(x1) + 2Cx1$$

$$\because L'(x1) = 0$$

$$\therefore f_2'(x1) = 2Cx1$$

$$\because x1 < 0$$

$$\therefore f_2'(x1) < 0$$

$$又 \because f_2'(0) = L'(0)$$

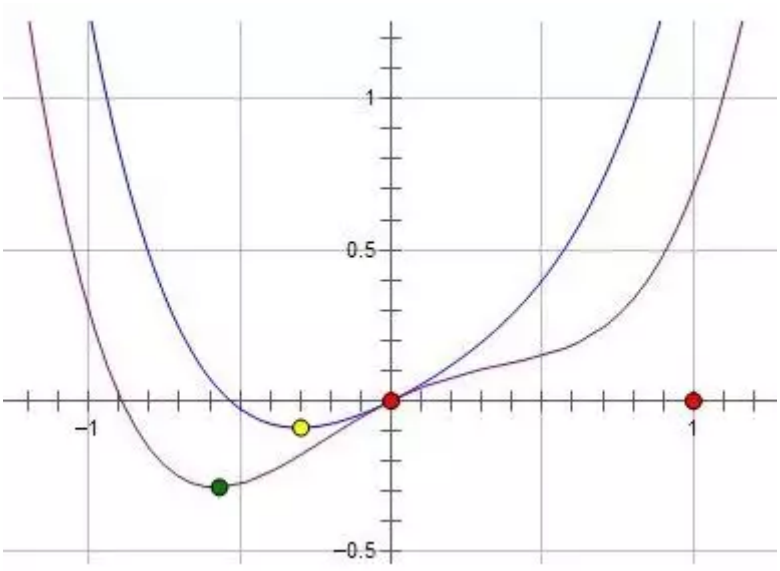
由上张图可知，$L'(0) > 0$

$$\therefore f_2'(0) > 0$$

即$f_2'(x)$在$x1$和0是异号，

$\therefore f_2(x)$在$(x1,0)$取极值

由结论可定性的画含L2正则化的图：



极值点为黄色点，即正则化L2模型的参数变小了。

含L1正则化的损失函数： $f_1(x) = L + C|x|$

对$f_1(x)$求导

$$f_1'(x) = L'(x) + C * \mathrm{sgn}(x)$$

当$x \to 0^+$时

$$f_1'(0^+) = L'(x) + C$$

当$x \to 0^-$时

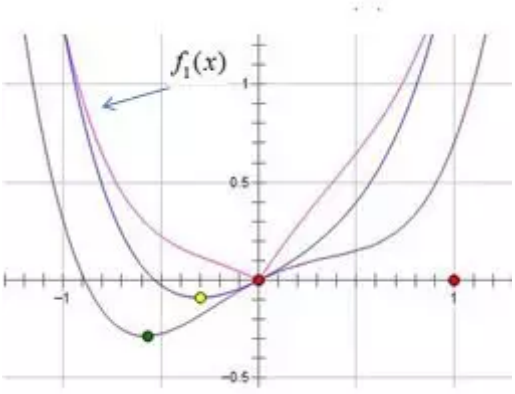
$$f_1'(0^-) = L'(x) - C$$

由第二条定理可知，

$$f_1'(0^+) * f_1'(0^-) < 0$$

得：$C > |L'(x)|$

因此，**只要C满足推论的条件，则损失函数在0点取极值(粉红色曲线），即L1正则化模型参数个数减少了。**



5、限制条件法

这种思想还是来自知乎的，觉得很有趣，所以就记录在这篇文章了，思想用到了凸函数的性质。我就直接粘贴这种推导了，若有不懂的地方请微信我。

考虑标量 w，希望 $\min_w f(w) + \lambda|w|_1$

假设 $f(w)$ 在 $w=0$ 附近为凸，则对充分小的 Δ，我们有

$$f(\Delta) \geq f(0) + f'(0)\Delta$$

这时候 $w=0$ 为局部最小值的充分必要条件就是：对 $\forall\Delta$ 接近于0，

$$f(\Delta) + \lambda|\Delta| \geq f(0)$$

由于有凸性质，$w=0$ 的充分条件为：

$$f'(0)\Delta + \lambda|\Delta| \geq 0$$

分情况考虑 Δ 正负，得到，

$$|f'(0)| \leq \lambda$$

而注意如果是l2:

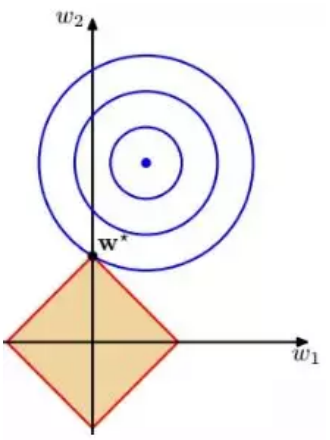
$$\min_w f(w) + \lambda||w||_2^2$$

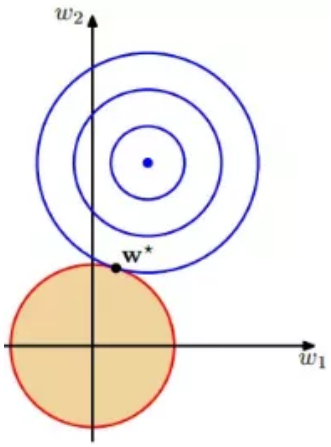
$w=0$ 为局部最小值的充分必要条件为 $f'(0)=0$。

结论：含L1正则化的损失函数在0点取得极值的条件比相应的L2正则化要宽松的多，所以，L1正则化更容易得到稀疏解（w=0）。

6、PRML的图形角度分析

因为L1正则化在零点附近具有很明显的棱角，L2正则化则在零附近比较平缓。所以L1正则化更容易使参数为零，L2正则化则减小参数值，如下图。





（1）L1正则化使参数为零　（2）L2正则化使参数减小

7、总结

本文总结了自己在网上看到的各种角度分析L1正则化和L2正则化降低复杂度的问题，希望这篇文章能够给大家平时在检索相关问题时带来一点帮助。若有更好的想法，期待您的精彩回复，文章若有不足之处，欢迎更正指出。

参考：

https://www.zhihu.com/question/37096933

林轩田老师 《机器学习基石》

推荐阅读文章

深入理解线性回归算法（二）：正则项的详细分析

浅谈频率学派和贝叶斯学派

浅谈先验分布和后验分布

-END-