**L1与L2损失函数和正则化的区别**

在机器学习实践中，你也许需要在神秘的L1和L2中做出选择。通常的两个决策为：1) L1范数 vs L2范数 的损失函数； 2) L1正则化 vs L2正则化。

**作为损失函数**

  L1范数损失函数，也被称为最小绝对值偏差（LAD），最小绝对值误差（LAE）。总的说来，它是把目标值（YiYi）与估计值（f(xi)f(xi)）的绝对差值的总和（SS）最小化：

S=∑i=1n|Yi−f(xi)|.S=∑i=1n|Yi−f(xi)|.

  L2范数损失函数，也被称为最小平方误差（LSE）。总的来说，它是把目标值（YiYi）与估计值（f(xi)f(xi)）的差值的平方和（SS）最小化：

S=∑i=1n(Yi−f(xi))2.S=∑i=1n(Yi−f(xi))2.

  L1范数与L2范数作为损失函数的区别能快速地总结如下：

| **L2损失函数** | **L1损失函数** |
| --- | --- |
| 不是非常的鲁棒（robust） | 鲁棒 |
| 稳定解 | 不稳定解 |
| 总是一个解 | 可能多个解 |

**鲁棒性（robustness）**，根据维基百科，被解释为：

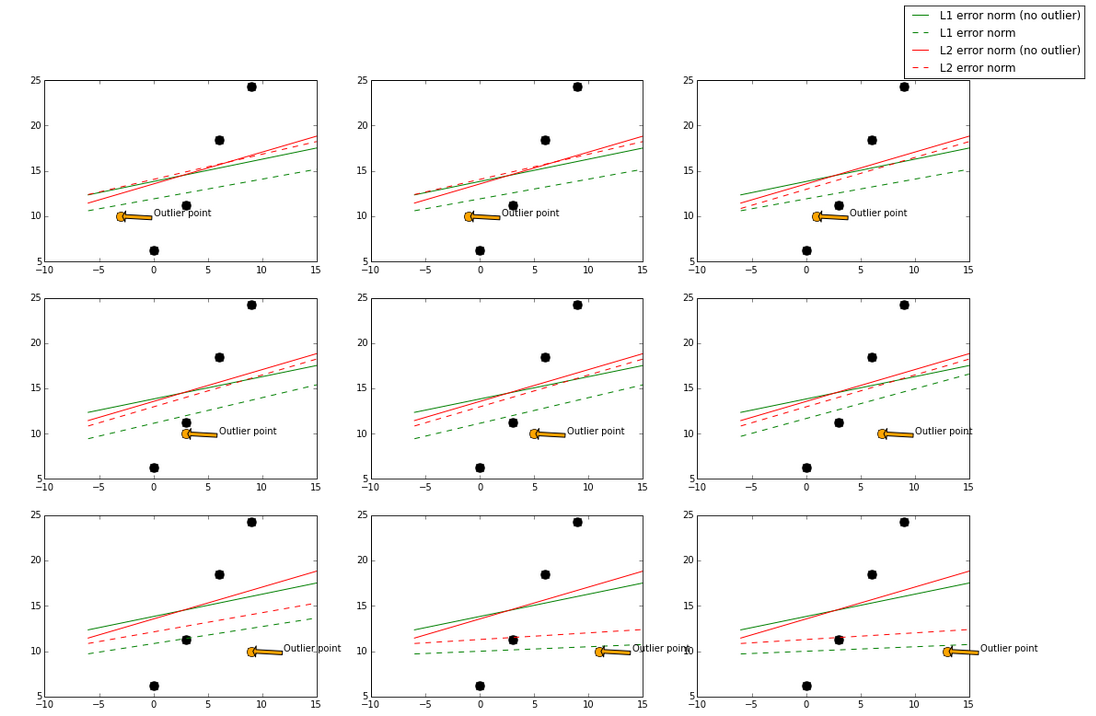
因为与最小平方相比，最小绝对值偏差方法的鲁棒性更好，因此，它在许多场合都有应用。最小绝对值偏差之所以是鲁棒的，是因为它能处理数据中的异常值。这或许在那些异常值可能被安全地和有效地忽略的研究中很有用。如果需要考虑任一或全部的异常值，那么最小绝对值偏差是更好的选择。

从直观上说，因为L2范数将误差平方化（如果误差大于1，则误差会放大很多），模型的误差会比L1范数来得大（ e vs e^2 ），因此模型会对这个样本更加敏感，这就需要调整模型来最小化误差。如果这个样本是一个异常值，模型就需要调整以适应单个的异常值，这会牺牲许多其它正常的样本，因为这些正常样本的误差比这单个的异常值的误差小。

**稳定性**，根据维基百科，被解释为：

最小绝对值偏差方法的不稳定性意味着，对于数据集的一个小的水平方向的波动，回归线也许会跳跃很大。在一些数据结构（data configurations）上，该方法有许多连续解；但是，对数据集的一个微小移动，就会跳过某个数据结构在一定区域内的许多连续解。（The method has continuous solutions for some data configurations; however, by moving a datum a small amount, one could “jump past” a configuration which has multiple solutions that span a region. ）在跳过这个区域内的解后，最小绝对值偏差线可能会比之前的线有更大的倾斜。相反地，最小平方法的解是稳定的，因为对于一个数据点的任何微小波动，回归线总是只会发生轻微移动；也就说，回归参数是数据集的连续函数。

下面的图是用真实数据和真实拟合模型生成的：

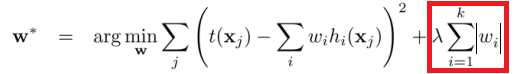


这里使用的基本模型为梯度提升回归（GradientBoostingRegressor），使用L1范数和L2范数作为损失函数。绿线和红色分别代表了模型使用L1范数与L2范数作为损失函数时的情形。实线代表了训练的模型中不含有异常值（橙色）的情形，虚线代表了训练的模型中含有异常值（橙色）的情形。  
  我缓慢地将这个异常值从左向右移动，使得它在中间时不那么异常，而在左右两边时更加异常。当这个异常值不那么异常时（在中间的情形），在拟合直线的时候，L2范数的变动较小，而L1范数的表动较大。  
  当这个异常值更加异常（上左位置，下右位置，它们离左、右两边更加远）时，这两个范数都有大的变动，但是再一次地，L1范数总体上比L2范数变动更大。  
  通过数据可视化，我们能够对这两个损失函数的稳定性有更好的认知。

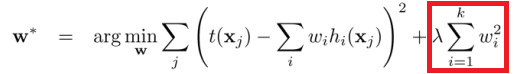
**作为正规化**

  在机器学习中，正规化是防止过拟合的一种重要技巧。从数学上讲，它会增加一个正则项，防止系数拟合得过好以至于过拟合。L1与L2的区别只在于，L2是权重的平方和，而L1就是权重的和。如下：

最小平方损失函数的L1正则化：



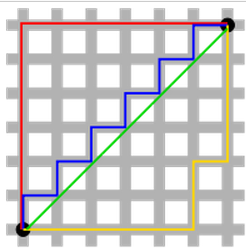
最小平方损失函数的L2正则化：



它们的性质的区别能快速地总结如下：

| **L2正则化** | **L1正则化** |
| --- | --- |
| 计算效率高（因为有解析解） | 在非稀疏情形下计算效率低 |
| 非稀疏输出 | 稀疏输出 |
| 无特征选择 | 内置特征选择 |

**解的唯一性**是一个更简单的性质，但需要一点想象。首先，看下图：



绿色的线（L2范数）是唯一的最短的路径，而红色、蓝色、黄色线条（L1范数）都是同一路径，长度一样（12）。可以将其扩展至n-维的情形。这就是为什么L2范数有唯一解而L1并不是。  
  **内置特征选择**是L1范数被经常提及的有用的性质，而L2范数并不具备。这是L1范数的自然结果，它趋向于产生稀疏的系数（在后面会解释）。假设模型有100个系数，但是仅仅只有其中的10个是非零的，这实际上是说“其余的90个系数在预测目标值时都是无用的”。L2范数产生非稀疏的系数，因此它不具备这个性质。  
  **稀疏性**指的是一个矩阵（或向量）中只有少数的项是非零的。L1范数具备性质：产生许多0或非常小的系数和少量大的系数。  
  **计算效率**。L1范数没有一个解析解，但是L2范数有。这就允许L2范数在计算上能高效地计算。然而，L1范数的解具备稀疏性，这就允许它可以使用稀疏算法，以使得计算更加高效。