论文阅读报告——Regression Shrinkage and Selection via the Lasso

2018E8008661001 孙普

一、摘要

文章针对线性模型的参数估计，提出了 lasso 方法，在最小化残差平方和的基础上，对系数增加了约束条件：系数的绝对值之和小于常数。这种约束会自然地使得一些系数等于零，这样模型就会有较好的解释性。lasso 的思想可以应用于其它多种统计模型。

二、lasso 的定义及示例

传统的最小二乘（OLS）估计有两个显著的缺点：一是估计准确度不高，原因在于，最小二乘估计往往都具有“小偏差，大方差”的特点；二是可解释性差。第一个问题产生的根源在于，OLS 估计为了提高训练时拟合的精确度，会尽量把噪声也拟合上，即所谓的过拟合（Overfitting），反映在系数上就是某些系数的数量级非常大，大系数可以把特征微小的变动放大，通过多个正负项的叠加尽量把每个点都拟合上。第二个问题的原因在于，估计出来的系数太多，不稀疏。

为了解决上述两个问题，两个标准的做法是 subset selection(可解释性好，但在不同训练集上表现不稳定) 和岭回归（稳定，但是只是压缩系数没有将一些系数完全置零，所以可解释性还是不好）。本文的 lasso 既压缩系数，又将一些系数置零，同时保留了两种做法的优点。

文中的 lasso 定义是这种形式（假设输出 y_i 均值为 0）：

$$\hat{\beta} = \arg\min\{\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2\} \text{ , subject to } \sum_j |\beta_j| \le t$$

这种形式和我们熟悉的形式是等价的：

$$\hat{\beta} = \arg\min\{\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|\}$$

两个式子中，分别通过 t 和 λ 控制系数压缩和置零的程度。

文中将 Subset regression、ridge regression、lasso、garotte 四种方法在样本相互正交情况下对系数的压缩情况作了比较，如图 1 所示：

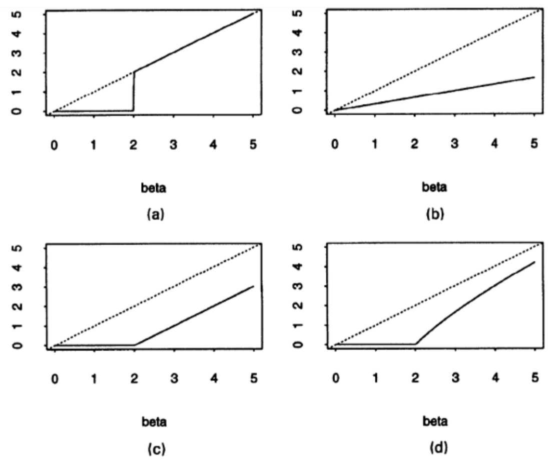

图 1 (a)Subset regression, (b)ridge regression, (c)lasso, (d)garotte 在正交样本下对系数的压缩示意图

如图 1(c)所示，lasso 会对 OLS 所得的系数设一个阈值，小于这个阈值的系数全部置零，大于这个阈值的系数按照比例进行压缩。

对于非正交的样本，作者研究了具有相关性的二维特征样本，发现 lasso 所得的系数不会随着各个特征的相关性大小而变化，但是岭回归所得的系数就会，因为岭回归倾向于把系数压缩成尽量相等的数。作者在文中还比较了 lasso 和岭回归用于贝叶斯估计所得的结果。贝叶斯估计的系数是一个分布，lasso 模型用于贝叶斯估计时，所产生的分布峰值点（在 0 点附近）更高，方差更大，说明 lasso 更倾向于得出 0 和较大的系数，而岭回归则倾向于得出彼此接近、接近 0 但不为 0 的系数。

三、lasso 的求解

作者在文中先讨论了 t 的估计方法。t 的大小会直接影响 lasso 的估计误差。t 如果太大，会造成欠拟合，太小又会造成过拟合。文中作者给出了三种计算 t 的方法：交叉验证（cross-validation）,泛化交叉验证（ generalized cross-validation）和分析性无偏风险估计（analytical unbiased estimate of risk）。理论上，前两种方法用于样本分布未知的情况，后一种方法用于样本分布已知的情况，但实际应用中并不需要如此严格，选择最便捷的方法即可。计算复杂度上，无偏风险估计最优，泛化交叉验证次之，交叉验证最复杂。

有了具体的t之后，作者进一步研究了 lasso 的具体解法。不同于岭回归，lasso 并不是连续可微的函数，所以求解起来较为不便。对于一个固定的$t \geq 0$，lasso 可以被表示为带有2^p个线性约束的最小二乘问题，2^p个线性约束对应于p个系数2^p种不同符号的组合，Lawson 和 Hansen(1974)提出了解决满足线性不等式$G\beta \leq h$的最小二乘问题的方法，这里G是一个$m \times p$的矩阵，对应p维向量β上的m个线性不等式约束。对于 lasso 问题，$m = 2^p$太大了，难以直接应用，故文中通过逐步加入约束条件来解决。

四、模拟实验

作者在文中生成了一些带噪声的样本，控制样本数量和系数大小，比较了子集选择、岭回归、lasso 的表现。对于数量少、系数大的样本，子集选择最好，lasso 性能一般，岭回归表现非常差；对于中等数量、中等大小系数的样本，lasso 的性能最好，岭回归次之，之后是子集选择；对于大数量、小系数的样本，岭回归的表现最好，其次是 lasso，最后是子集选择。

五、应用

lasso 可以被用到更多一般的回归模型上，例如 proportional hazards model（Tibshirani，1994），Logistic Regression 等等。

六、小结

lasso 的本质就是在标准 OLS 方法上加入惩罚的正则项，来使估计得到的系数不会太大，从而在欠拟合和过拟合上得到平衡。由 lasso 估计而得的系数会有部分系数置零和整体系数压缩的效果，所以 lasso 同时具有 subset selection（可解释性）和岭回归（稳定性）的优点。