线性回归, 岭回归以及 Lasso 回归

邵李翔　赵张弛　祖劭康

摘　要

我们采用 Boston 数据集，使用线性回归、lasso、岭回归三种方式进行建模，以数据集中的 medv 变量为响应变量，其余 13 个变量为预测变量，探究之间的线性关系。并比较三种模型求解出来的系数，分析三种方法的优缺点。

1　数据集描述

本次实验采用的数据集为 Boston (波士顿房价) 数据集，它记录了波士顿周围 506 个街区的 medv (房价中位数)。我们将设法用 13 个预测变量如 rm (每栋住宅的平均房间数)，age (平均房龄)，lstat (社会经济地位低的家庭所占比例) 等来预测 medv (房价中位数)。下表是部分数据集展示。

表 1: 数据集中部分数据展示

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.01	18.00	2.31	0.00	0.54	6.58	65.20	4.09	1.00	296.00	15.30	396.90	4.98	24.00
2	0.03	0.00	7.07	0.00	0.47	6.42	78.90	4.97	2.00	242.00	17.80	396.90	9.14	21.60
3	0.03	0.00	7.07	0.00	0.47	7.18	61.10	4.97	2.00	242.00	17.80	392.83	4.03	34.70

2　线性回归

线性模型结构如下：

$$y = \beta^T x + \varepsilon \tag{1}$$

这里 x 是一个 p 维向量，代表 p 个预测变量。在给定 n 个数据后，基于最小二乘法求解模型系数, 此时模型损失函数为：

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 \tag{2}$$

这里我们可以直接得到系数 β 的最小二乘解：$\hat{\beta} = (X^T X)^{-1} XY$

我们对所有预测变量进行多元回归, 得到的结果如表 2。由各个预测变量对应的系数的值可以看出来，有些变量比如 black，age 等系数绝对值小于 0.01，可以说与响应变量——房价中位数关系不大，而且表格第五列是系数显著性检验的结果，比如 indus、age 对应的 p 值显然是表明接受原假设，认为该变量的系数应该等于 0。说明线性回归在某些数据集上还存在一定的局限性。

表 2: 多元线性回归拟合结果

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.4595	5.1035	7.14	0.0000
crim	-0.1080	0.0329	-3.29	0.0011
zn	0.0464	0.0137	3.38	0.0008
indus	0.0206	0.0615	0.33	0.7383
chas	2.6867	0.8616	3.12	0.0019
nox	-17.7666	3.8197	-4.65	0.0000
rm	3.8099	0.4179	9.12	0.0000
age	0.0007	0.0132	0.05	0.9582
dis	-1.4756	0.1995	-7.40	0.0000
rad	0.3060	0.0663	4.61	0.0000
tax	-0.0123	0.0038	-3.28	0.0011
ptratio	-0.9527	0.1308	-7.28	0.0000
black	0.0093	0.0027	3.47	0.0006
lstat	-0.5248	0.0507	-10.35	0.0000

3 Lasso

3.1 模型介绍

Lasso 在基础的线性回归模型的损失函数上，增加了 L1 正则项，假设有 p 个预测变量，此时 Lasso 损失函数如下:

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

这里 λ 是超参数。

3.2 超参数 λ 选择

调用 R 语言 glmnet 包，实现 Lasso Regression，将数据分成十份，计算不同超参数 λ 下交叉验证 (Cross Validation) 的误差，并选择最优的超参数，结果如图 1 所示 由结果可知，我们选取最小的均方误差对应的超参数系数，最优的 λ 取值约为 0.021。

3.3 结果及分析

分别用 glmnet 包和自己编写的使用循环坐标下降法进行优化的代码在最优的 λ 取值下进行 Lasso Regression，拟合得到的模型系数如表 3 所示

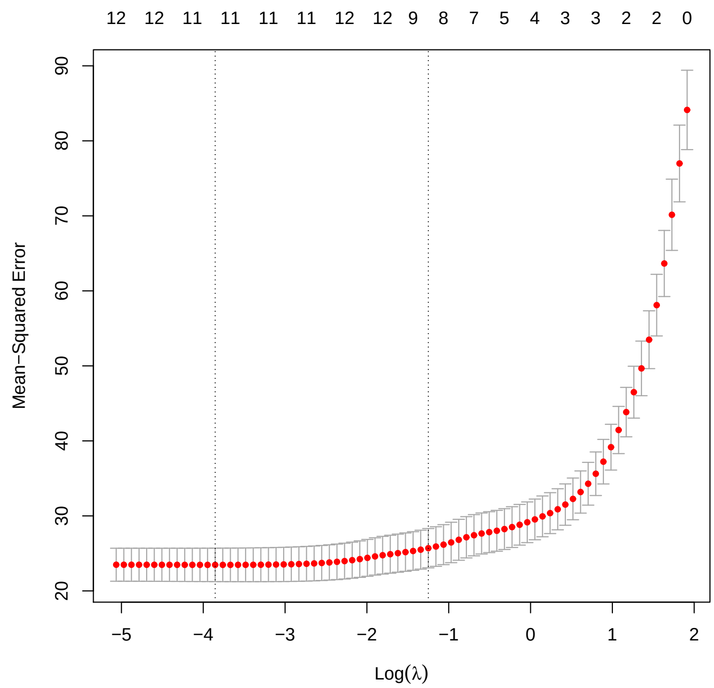

图 1: Lasso Cross Validation

表 3: 自己编写与 glmnet 包用 lasso 估计系数的结果

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	intercept
glmnet	-0.10	0.04	0.00	2.69	-16.57	3.85	0.00	-1.42	0.26	-0.01	-0.93	0.01	-0.52	34.91
ours	-0.10	0.02	-0.02	4.45	-11.34	3.41	0.00	-0.75	0.00	0.00	0.00	0.00	-0.55	17.16

从表中我们可以看出，glmnet 包的结果与我们的结果在各个项与房价的正负相关性上是相同的，但我们的结果将更多的系数压缩到 0。存在这种差别的原因是 lasso 无法得到系数解析解，所以要采用别的计算方法，而我们优化计算系数的方式是循环坐标下降法，而 R 自带包计算系数的方法是最小角回归法，计算方法上最小角回归法更容易得到最优解。但两种方法的结果总体比较还是较为一致的。

在 glmnet 的结果中，房价与 crim 犯罪率、nox 氮氧化物浓度、dis 到就业中心的距离、ptratio 学生与教师比例、lstat 人口数量呈负相关，与 zn 住宅用地比例、chas 查尔斯河虚拟变量、rm 每栋住宅的平均房间数、rad 可达公路数、black 黑人比例呈正相关，与 indus、age 两个变量无关，而这也与之前线性回归模型参数估计中，发现这两个变量的系数显著性检验并未通过的结果相一致。

在我们的结果中。房价与犯罪率、indus 商场面积、氮氧化物浓度、到就业中心的距离、人口数量呈负相关，与住宅用地比例、查尔斯河虚拟变量、每栋住宅的平均房间数呈正相关，与其余变量无关。

以上结果都与常识相符，比如犯罪率与房价的负相关性。接下来我们从 MSE 判断两种模型的优劣性，两种代码的 MSE 如下表所示

表 4: 采用自己编写与 glmnet 包的系数计算的 MSE 比较

	MSE
glmnet	21.92
ours	27.72

由结果可见，glmnet 包训练的模型 MSE 更小，说明采用最小角回归方法优化计算的 lasso 系数估计值效果更好。

4 岭回归

4.1 模型介绍

岭回归是在基础的线性回归模型的损失函数上，增加了 L2 正则项，同样假设有 p 个预测变量，此时 Lasso 损失函数如下:

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

由于岭回归的损失函数可导，所以也可以直接求解出一个回归系数的估计值为 $\hat{\beta} = (X^T X + \lambda I)^{-1} XY$，随着 λ 的增大，$(X^T X + \lambda I)^{-1}$ 就越小，模型的方差就越小；而 λ 越大使得 β 的估计值更加偏离真实值，模型的偏差就越大。所以岭回归的关键是找到一个合理的 λ 值来平衡模型的方差和偏差。

4.2 自己实现的岭回归与自带包实现的岭回归

这里我们先比较自己实现的岭回归的参数估计与自带包实现的参数。首先选取超参数 λ 为 0.01, 使用两种不同的代码实现岭回归, 得到模型中的参数如下表。我们可以看到两者系数相差无几，而之前 lasso 自己编写的代码和自带包估计的系数差距较大，我们分析认为是岭回归同样能得到模型的解析解，可以直接用数据来计算得到每个超参数下系数的最优估计，所以在岭回归方面，我们自己编写的岭回归参数估计与 R 中自带包计算的效果几乎相同。

表 5: 自己编写与 R 包用岭回归方法估计系数的结果

	截距项	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
手写代码	35.96	-0.11	0.05	0.02	2.69	-17.48	3.83	0.00	-1.47	0.30	-0.01	-0.95	0.01	-0.52
自带包	36.46	-0.11	0.05	0.02	2.69	-17.76	3.81	0.00	-1.48	0.31	-0.01	-0.95	0.01	-0.52

4.3 基于交叉验证选择最佳参数

自动选择参数 λ 值的范围进行岭回归, 选择在 $\lambda = 10^{-3}$ 到 $\lambda = 10$ 的范围内进行岭回归, 如图所示可得每个变量的系数随着参数 λ 变化所得到的曲线. 该图上方的 13 是系数个数，而下方 10，15，20，25 是每个 λ 的值对应的各变量参数的绝对值之和。我们可以从图的左端看出，即使各个变量系数的值都接近到 0，但在图像上方对应的值仍为 13，说明岭回归没有变量选择的作用，只会随着 λ 值增加，而压缩各变量系数的值。

将数据分为 10 折, 使用交叉验证法选择调节参数 λ, 得到系数变化曲线如图 3 所示

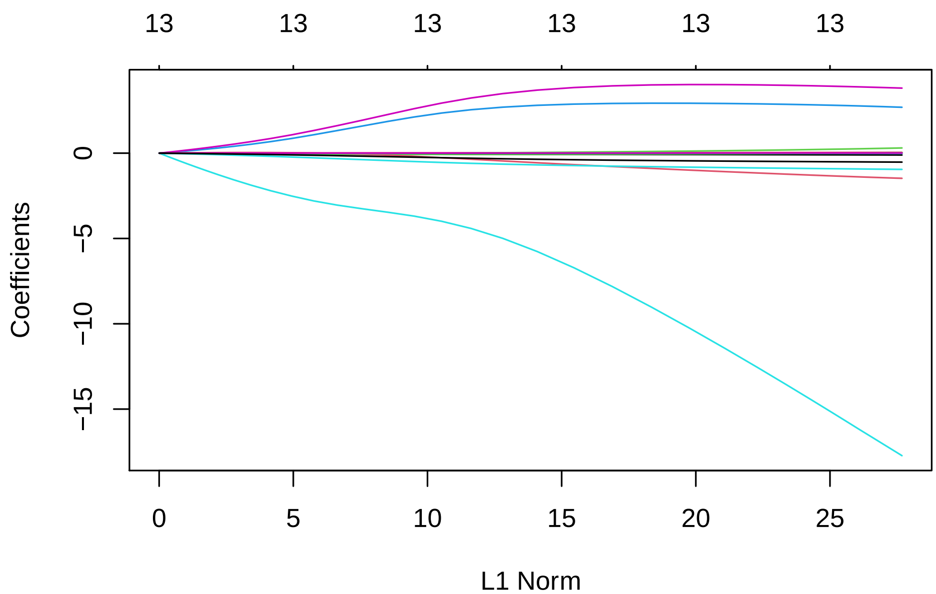

图 2: 系数变化曲线

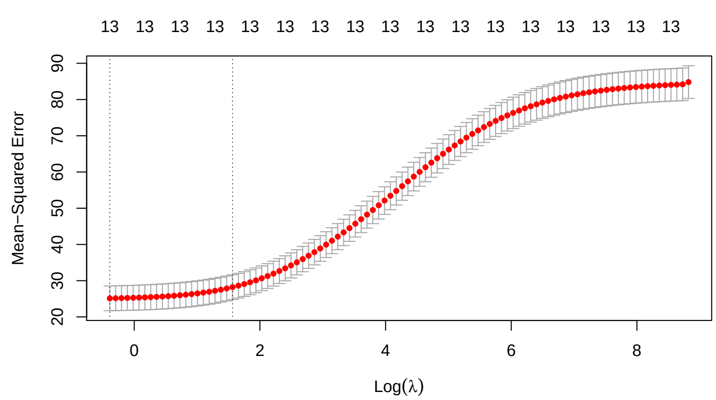

图 3: 交叉验证法的系数变化曲线

我们分别选取 λ 的 LSE 值以及最小的 λ 值下的变量系数, 结果如图 4 所示，数据第一列是 λ 的 LSE 值对应的变量系数，第二列是 CV 准则下最小 MSE 的 λ 值下的变量系数。

	1	2
(Intercept)	28.002622542	20.639034849
crim	-0.087575337	-0.066141464
zn	0.032682858	0.019688710
indus	-0.037995597	-0.070139429
chas	2.899744506	2.691457142
nox	-11.914110425	-4.964448286
rm	4.011259959	3.483521810
age	-0.003730882	-0.008045836
dis	-1.118912658	-0.454261144
rad	0.153749809	0.023165777
tax	-0.005751983	-0.002826955
ptratio	-0.854993947	-0.643103460
black	0.009073718	0.007326400
lstat	-0.472427733	-0.329619357

图 4: 两个 λ 值下的系数

5 总结

对于 Boston 数据集直接采用线性回归进行参数估计效果并不好，我们分别采用 lasso 回归与岭回归方法同样对该数据集进行参数估计，lasso 具有变量选择的效果，而且基于最小角回归算法得到的 lasso 估计参数比我们自己采用循环坐标下降法计算得到的参数效果好。而岭回归具有压缩系数的作用，但无法将系数压缩到 0。