Ridge と Lasso の推定

2019 年 5 月 8 日

大森 夢拓

1 はじめに

複雑な構造を内在する現象の分析では，多項式回帰などの説明変数に関して非線形なモデルを考える必要がある．このような複雑なモデルの推定において生じる過学習を防ぐ手法が正則化法である．本稿では正則化法の中でも，Ridge と Lasso を用いた推定を紹介する．また，それらを用いて実際のデータに対して計算機実験を行った．

2 Ridge と Lasso の推定

i 番目 $(i = 1, 2, \ldots, n)$ のデータの目的変数を y_i，説明変数を $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$，誤差を $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$，および回帰係数を $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$ とした時の線形回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{1}$$

を考える．パラメータベクトルの長さの概念を一般化した L_q ノルムと呼ばれる正則化項を用いた正則最小 2 乗法は

$$S\alpha(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})\big)^2 + \alpha \sum_{j=1}^{p} |\beta_j|^q \tag{2}$$

のよう記述できる．$\alpha \in \mathbb{R}^+$ はモデルの複雑度を設定するハイパーパラメータであり，式 (2) の第 2 項が $q = 1 (L_1$ ノルム) の場合を Lasso，$q = 2$ の場合を Ridge と呼ぶ．

2.1 Ridge

切片を除く回帰係数の 2 乗和を正則化項としてとることで，説明変数間の強い相関によって生じる回帰係数の推定値の不安定性を回避する手法を Ridge という．Ridge 推定量の求め方は次のようになる．

j 番目の説明変数に関するデータの平均値 $\bar{x}_j$ を中心化したデータを $z_{ij} = x_{ij} - \bar{x}_j$ として，式 (1) は次のように変形できる．

$$y_i = \beta_0^* + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + \epsilon_i \tag{3}$$

ただし，$\beta_0^* = \beta_0 + \beta_1 \bar{x}_{i1} + \cdots + \beta_p \bar{x}_p$ である．さらに，$\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)^\top \in \mathbb{R}^p$ および $\mathbf{1} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$ と定義すると式 (3) は以下の行列形式で記述できる．

$$\begin{aligned} \boldsymbol{y} &= \beta_0^* \mathbf{1} + Z\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \\ Z &= \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & z_{ij} & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix} \end{aligned} \tag{4}$$

よって，Ridge の線形回帰モデルは，

$$S\alpha(\beta_0^*, \boldsymbol{\beta}_1) = (\boldsymbol{y} - \beta_0^* \mathbf{1} - Z\boldsymbol{\beta}_1)^\top (\boldsymbol{y} - \beta_0^* \mathbf{1} - Z\boldsymbol{\beta}_1) + \alpha \boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 \tag{5}$$

と記述できる．これを最小化することで，次のような切片と回帰係数ベクトルの推定量が与えられる．

$$\hat{\beta}_0^* = \bar{y}, \quad \hat{\boldsymbol{\beta}}_1 = (Z^\top Z + \alpha I_p)^{-1} Z^\top \boldsymbol{y} \tag{6}$$

式 (6) の結果から，式 (1) の切片は $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \cdots - \hat{\beta}_p \bar{x}_p$ と推定されることが分かる．つまり，中心化したデータから構成された計画行列 Z を用いれば，式 (5) の β_0^* を考えることなく，

$$S\alpha(\boldsymbol{\beta}_1) = (\boldsymbol{y} - Z\boldsymbol{\beta}_1)^\top (\boldsymbol{y} - Z\boldsymbol{\beta}_1) + \alpha \boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 \tag{7}$$

の最小化によって $\hat{\boldsymbol{\beta}}_1$ を得られる．

2.2 Lasso

切片を除く回帰係数の絶対値の和を正則化項としてとることで，パラメータの一部は完全に 0 と推定される．このようなモデルの推定と変数選択を同時に実行できる手法を Lasso という．切片を除く回帰係数はデータを中心化することによって切片と切り離して推定できるため

$$S\alpha(\boldsymbol{\beta}_1) = \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{p} \beta_j z_{ij}\big)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \tag{8}$$

の最小化によって与えられる．しかし，L_1 正則化項が微分不可能であるため解析的に $\boldsymbol{\beta}_1$ を求めることはできない．このため，Fu (1998) による shooting アルゴリズムや Efron et al. (2004) による LARS (Least Angle Regression) と呼ばれる計算手法が用いられる．

また，ラグランジュの未定乗数法を適用すると，次の様な制約条件つきのパラメータベクトルの最小化と同等になる．

$$\min \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{p} \beta_j z_{ij}\big)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq \eta \tag{9}$$

3　計算機実験

実データ実験として，アメリカのボストン州の住宅価格を，線形回帰モデルにより推定する．データ数は 506 で，住宅価格に影響を与えていると思われる以下の 13 個の説明変数を持つ．

- x_1: 犯罪率
- x_2: 宅地の割合
- x_3: 非商用地
- x_4: チャールズ川流域か否か
- x_5: 窒素酸化物
- x_6: 部屋数
- x_7: 築年
- x_8: ビジネス地域への距離
- x_9: ハイウェイアクセス指数
- x_{10}: 固定資産税
- x_{11}: 生徒と教師の比率
- x_{12}: 有色人種の割合
- x_{13}: 低所得者の割合

通常の最小 2 乗法，Ridge および Lasso を用いて回帰係数 $\boldsymbol{\beta}$ の推定を行い，結果を比較した．Ridge および Lasso は $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ で実験を行った．この実験結果として，最小 2 乗法により推定された $\hat{\boldsymbol{\beta}}$ の値を図 1，α の変化による Ridge および Lasso の推定値 $\hat{\boldsymbol{\beta}}$ の解パスをそれぞれ図 2，図 3 に示す．α が 0 に近い場合，Ridge と Lasso の各推定値は最小 2 乗法のそれにかなり近いことが分かる．α が大きくなるにつれて，Ridge も Lasso も全ての説明変数の推定値が 0 に向かって縮小している．特に，Lasso において α=100 の場合は全ての回帰係数の値がほぼ 0 となっている．これらの結果から，α の大きさの違いによって正則化の強さが変化することを確認できる．また，Lasso での変数選択の結果も図 3 から確認できる．実際に $\alpha = 100$ において変数選択されたのは，固定資産税 (x_{10}) と有色人種の割合 (x_{12}) であり，$\alpha = 10$ において変数選択されたのは，先ほどの x_{10} と x_{12} に加えて，宅地の割合 (x_2) と低所得者の割合 (x_{13}) であった．

続いて，先の実データの説明変数の数を 103 に増やしたもので同様の実験を行い，推定手法の精度の指標である決定係数を比較した．

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{10}$$

ただし，$\bar{y} = \sum_{i=1}^{n} y_i / n$ とする．この実験結果として，Ridge と Lasso の訓練データとテストデータにおける，α の変化による決定係数の推移を図 4 と表 1 に示す．訓練データでは，α が大きくなるにつれて，Ridge と Lasso の両方の決定係数の値が小さくなっている．特に，Lasso では $\alpha \geq 10$ の時点で決定係数が 0 となっている．これらの結果から，Ridge と Lasso の両方において，α が 0 に近づくほど最小 2 乗法の決定係数に近づいていることが分かる．また，テストデータでは，適切な α を設定することで，Ridge と Lasso の両方において，最小 2 乗法を上回る精度を出すことが分かった．

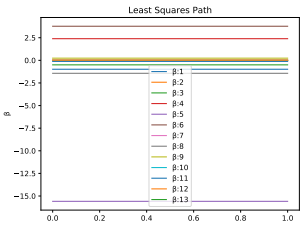

図 1: 最小 2 乗法の推定値

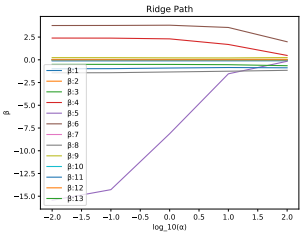

図 2: Ridge の解パス

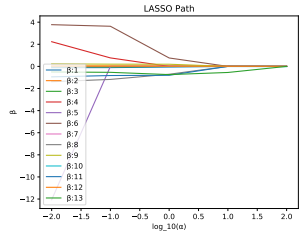

図 3: Lasso の解パス

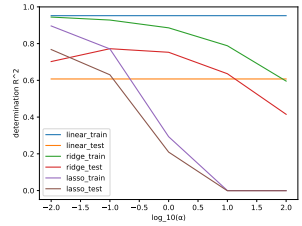

図 4: 決定係数

表 1: 決定係数

α	0.01	0.1	1	10	100
train					
linear	0.95				
ridge	0.94	0.93	0.89	0.79	0.60
lasso	0.90	0.77	0.29	0.00	0.00
test					
linear	0.61				
ridge	0.70	0.77	0.75	0.64	0.42
lasso	0.77	0.63	0.21	0.00	0.00

4　まとめ

Ridge と Lasso を中心に，正則化法や線形回帰モデルについて学ぶことができた．数式だけでなく，Python のパッケージを用いた計算機実験による可視化を通して，自分の中でさらにイメージを具体化させることができた．

今後は，Lasso の計算に用いるアルゴリズムを理解して，パッケージを使わずに実装することが課題である．

参考文献

[1] 小西貞則，『多変量解析入門』. 岩波書店, 2009.