Ridge と Lasso の推定

B4 Y.Omori

Nagoya Institute of Technology
Takeuchi & Karasuyama Lab

May 7, 2019

Outline

正則化法

正則化法

複雑なモデルの推定において生じる過学習を防ぐ手法

今回は以下の 2 つの正則化法を紹介する

- Ridge
- Lasso

線形回帰モデル

次のような線形回帰モデルを考える

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

- y_i: i 番目 $(i = 1, 2, \dots, n)$ のデータの目的変数
- $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^p$: 説明変数
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$: 誤差
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$: 回帰係数

正則化最小 2 乗法

最小 2 乗法の損失関数に次の正則化項を加える

$$S_\alpha(\boldsymbol{\beta}) = \underbrace{\sum_{i=1}^{n}\Big(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij})\Big)^2}_{\text{2 乗誤差の和}} + \alpha \underbrace{\sum_{j=1}^{p}\|\beta_j\|_q^q}_{\text{正則化項}} \tag{2}$$

- $\alpha \in \mathbb{R}^+$: モデルの複雑度を設定するハイパーパラメータ

第 2 項の q の値によって次のように呼ばれる

1. $q = 1 \rightarrow$ Lasso
2. $q = 2 \rightarrow$ Ridge

Next Section

Ridge

説明変数間の強い相関から生じる多重共線性を回避する手法

- 多重共線性: $(X^\top X)^{-1}$ の計算の部分で分母が 0 になる問題

Ridge 推定量を導出するため，線形回帰モデルの中心化を考える

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

- $\bar{x}_j$: j 番目の説明変数に関するデータの平均値
- $z_{ij} = x_{ij} - \bar{x}_j$: $\bar{x}_j$ を中心化したデータ
- $\beta_0^* = \beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p$

Ridge 回帰モデル

中心化した線形回帰モデルは次の行列形式で記述できる

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

- $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)^\top \in \mathbb{R}^p$
- $\mathbf{1} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$

Ridge 回帰モデル

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

この式の最小化により切片と回帰係数ベクトルの推定量が求められる

Ridge の推定量

中心化した Ridge 回帰モデルの推定量

- $\hat{\beta}_0^* = \bar{y} = \sum_{i=1}^n y_i/n$: 切片
- $\hat{\boldsymbol{\beta}}_1 = (Z^\top Z + \alpha I_p)^{-1} Z^\top \boldsymbol{y}$: 回帰係数ベクトル

この結果から式 (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

$\rightarrow$ β_0^* を考えることなく

$$S_\alpha(\boldsymbol{\beta}_1) = (\boldsymbol{y} - Z\boldsymbol{\beta}_1)^\top (\boldsymbol{y} - Z\boldsymbol{\beta}_1) + \alpha \boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 \tag{6}$$

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

Lasso とは

Lasso

パラメータの一部を完全に 0 と推定することから，モデルの推定と変数選択を同時に実行できる手法

切片を除く回帰係数はデータの中心化によって切片と切り離して推定できる
$\rightarrow$ 次の式の最小化によって与えられる

$$S_\alpha(\boldsymbol{\beta}_1) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j z_{ij} \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \tag{7}$$

Lasso の計算手法

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
$\rightarrow$ 次のような計算手法が用いられる

- shooting アルゴリズム (Fu 1998)
- LARS (Efron et al. 2004)

また，ラグランジュの未定乗数法を用いて制約つき最小化問題に変形できる

$$\min_{\boldsymbol{\beta}_1} (\boldsymbol{y} - Z\boldsymbol{\beta}_1)^\top (\boldsymbol{y} - Z\boldsymbol{\beta}_1) \quad \text{subject to} \quad \|\beta_j\|_q^q \leq \eta \tag{8}$$

- η: α によって定まる定数

Ridge と Lasso の違い

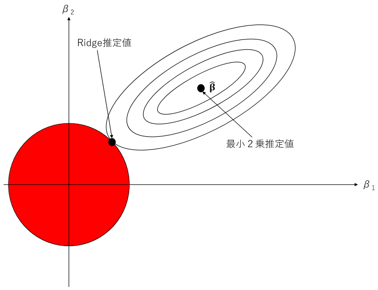

Figure: Ridge の推定

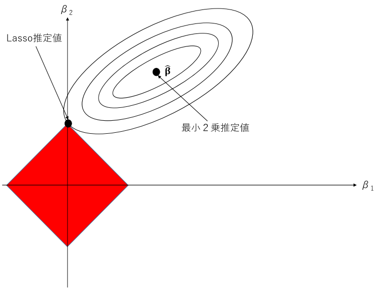

Figure: Lasso の推定

説明変数を x_1 と x_2 の 2 つの場合を考える

- Ridge では $\beta_1^2 + \beta_2^2 \leq \eta$ を満たす解を求める
  - 最小 2 乗推定量 $\hat{\boldsymbol{\beta}}$ よりも β_1，β_2 ともに 0 に近い値となる
- Lasso では $|\beta_1| + |\beta_2| \leq \eta$ を満たす解を求める
  - $\beta_1 = 0$ のようにスパースな解が得られやすい

Next Section

実験の設定

通常の最小 2 乗法，Ridge および Lasso を用いて回帰係数の推定を行った

実験設定

- 目的変数: アメリカのボストン州の住宅価格
- 説明変数: 住宅価格に影響を与えていると思われる以下の 13 個

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	

- データ数: 506

正則化の確認

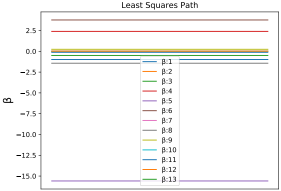

Figure: 最小 2 乗法の推定値

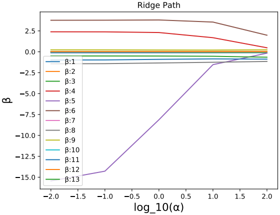

Figure: Ridge の解パス

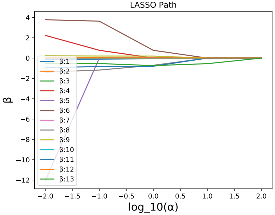

Figure: Lasso の解パス

- α が 0 に近いほど，Ridge と Lasso の各推定値は最小 2 乗法のそれに近い
- α の増化によって，Ridge も Lasso も全説明変数の推定値が 0 に向かって縮小
  - 特に，Lasso において $\alpha = 100$ の場合は全ての回帰係数の値がほぼ 0

$\rightarrow$ 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

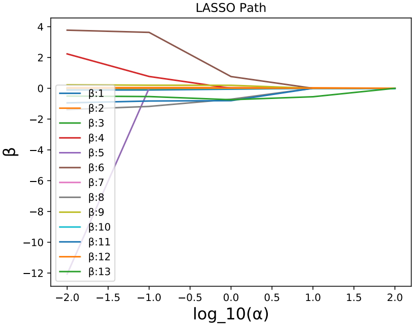

Figure: Lasso の解パス

左図から変数選択の結果が確認できる

- $\alpha = 100$ において選択された変数
  - 固定資産税 (x_{10})
  - 有色人種の割合 (x_{12})
- $\alpha = 10$ において選択された変数
  - x_{10} と x_{12}
  - 宅地の割合 (x_2)
  - 低所得者の割合 (x_{13})

決定係数の実験設定

推定手法の精度を比較するため，説明変数の数を 103 に増やして実験

実験設定

- 目的変数: アメリカのボストン州の住宅価格
- 説明変数: 住宅価格に影響を与えていると思われる 103 個
- データ数: 506

推定手法の精度の指標として決定係数 R^2 を用いた

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{9}$$

決定係数の実験結果

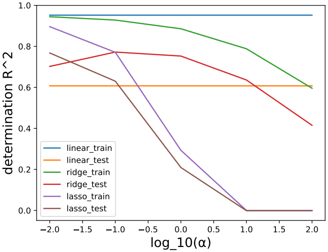

Figure: 決定係数の推移

Table: 決定係数の結果表

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

- 訓練データ
  - α が大きくなるにつれて，両方の決定係数が減少
- テストデータ
  - 適切な α を設定することで，両方において最小 2 乗法を上回る精度を出す

Next Section

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習できた
- 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 (References)

1. 小西貞則, 『多変量解析入門』. 岩波書店, 2009.