

データ解析入門 17 <Lasso・Ridge・Elastic Net>

キーワード：正則化法、Lasso、Ridge、Elastic Net、変数選択、スパースモデリング、グループ効果

はじめに

前稿¹⁾において、変数の数がサンプル数を上回る劣決定系における重回帰分析 (MLR) の問題点を紹介しました。本稿では、劣決定系の問題においてしばしば用いられる正則化法と呼ばれるテクニックについて解説します。具体的には、有名なスパース推定法である Lasso 回帰、Ridge 回帰、および Elastic Net 回帰を紹介します。

正則化法

MLR では、以下の損失関数 L を最小にする回帰係数ベクトル β を求めます。 N 、 y 、 X は、それぞれデータ数、目的変数ベクトル、データ行列です。

$$L(\beta) = \frac{1}{2N} \|y - X\beta\|_2^2 \quad (1)$$

式(1)右辺の数学的表現は L2 ノルムの 2 乗を表します。ノルムとはベクトルの大きさを定める量ののようなものです²⁾。優決定系では、式(1)を最小にする β は最小二乗法により解析的に求められますが、劣決定系では解が一意に定まりません。

そこで正則化法では、式(1)の損失関数に罰則項を導入し、回帰係数が極端に大きな値にならないように制約を課します。たとえば、式(2)の損失関数などのように罰則項(右辺第 2 項)を追加します。

$$L_\lambda(\beta) = \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (2)$$

式(2)に示す罰則項は、回帰係数ベクトルの L1 ノルムであり、 $\|\beta\|_1 = \sum_i |\beta_i|$ です。この最小化問題を解くと、いくつかの回帰係数が 0 に推定されます。このような最小化問題を解く回帰分析手法を Lasso 回帰といいます。ただし、MLR と異なり、解析的に解が得られないため、座標降下法などのアルゴリズムにより回帰係数を推定します。なお、式(2)の λ は罰則の大きさを制御する正則化パラメータであり、Lasso 回帰のハイパーパラメータです。

また、式(3)に示すように、L2 ノルム正則化項を用いる方法も有名です。

$$L_\lambda(\beta) = \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad (3)$$

式(3)の最小化問題は解析的に解くことができ、回帰係数は式(4)から求められます。

$$\beta = (X^T X + \lambda I)^{-1} X^T y \quad (4)$$

このような回帰手法を Ridge 回帰と呼びます。なお、 $\|\beta\|_2 = \sqrt{\sum_i \beta_i^2}$ であり、 I は単位行列です。Lasso 回帰と同様に、正則化パラメータ λ は Ridge 回帰のハイパーパラメータです。

Lasso 回帰および Ridge 回帰について

本稿では、有機化合物の水溶解度データセット³⁾を用い、正則化法の特徴を確認します。ここでは訓練データは 200 件とし、MLR、Lasso 回帰および Ridge 回帰モデルを構築しました。図 1 にテストデータの予測結果を示します。Lasso 回帰および Ridge 回帰では、MLR よりも高い決定係数 R^2 が確認されました。ただし、より多くの訓練データを用いた場合、MLR の R^2 が Lasso 回帰および Ridge 回帰を上回るケースも見られました。

次に、各回帰モデルの標準回帰係数を図 2 に示します。まず、MLR の標準回帰係数のスケールが比較的大きいことがわかります。また、Lasso 回

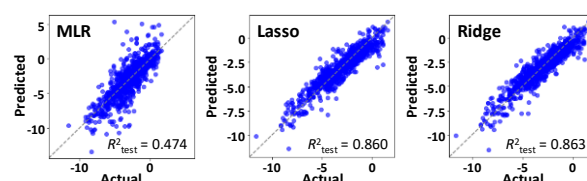


図 1 予測結果 (訓練データ 200 件)

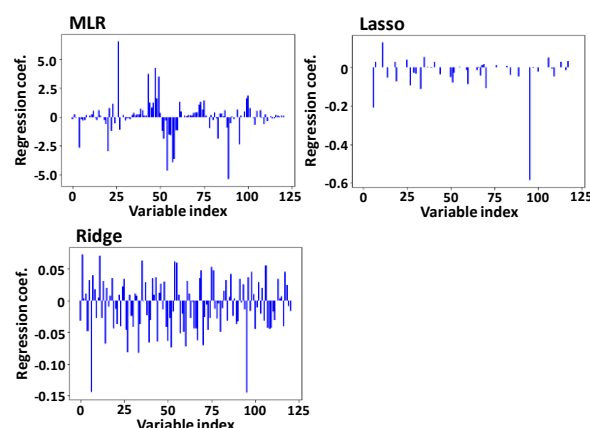


図 2 各予測モデルの標準回帰係数

帰では、くしの歯のような棒グラフが得られました。これは、多くの標準回帰係数が 0 であることを示します。一方、Ridge 回帰では、厳密に 0 である標準回帰係数は得られません。

Lasso 回帰によるスパースモデリング

上記において、Lasso 回帰の標準回帰係数の多くが 0 に推定されることを確認しました。つまり、Lasso 回帰では疎(スパース)な解が得られました。このような挙動は、実用上重宝されることがあります。たとえば、MLR で過剰適合が見られるような変数の多い状況下においても、Lasso 回帰では説明変数が削減されるため、過剰適合が抑制されることがあります。また、予測モデル構築と同時に変数選択が実行されるため、線形回帰モデルの解釈性も向上します。このようにスパースなデータ構造を仮定する解析手法はスパースモデリングなどと呼ばれ、様々な場面で適用されています。

なお、Ridge 回帰ではスパースな解は得られません。両手法の違いを幾何学的に説明します。図 3 では 2 つの説明変数を考えており、赤い領域が正則化による制約の範囲です。青い等高線は式 (1) の損失の大きさを示します。図 3 に示すように、L1 ノルム正則化では、制約条件の形状から一部の解が 0 になりやすいですが、Ridge 回帰はそのような性質は持ちません。この違いが標準回帰係数に現れていました。

また、正則化の強さと標準回帰係数の関係(解パス図)を図 4 に示します。Lasso 回帰では、正則化が強くなるに従い、回帰係数が 0 である変数が増加していることがわかります。一方、Ridge 回帰

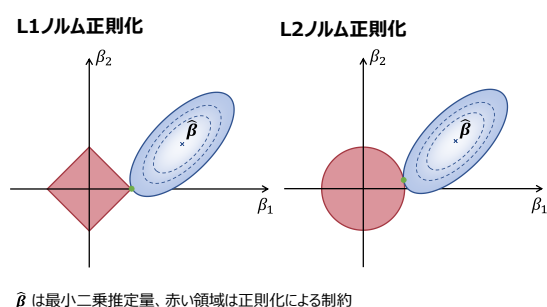


図 3 L1 および L2 ノルム正則化について

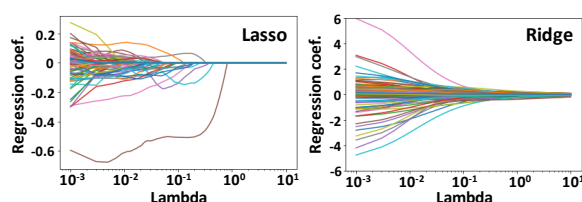


図 4 解パス図

では、正則化パラメータの増大に伴い、回帰係数は 0 に近づくものの、厳密には 0 になりません。また、相関する変数の回帰係数は近い値に推定される傾向にあります。Ridge 回帰のこのような性質はグループ効果などと呼ばれます。一方、Lasso 回帰では相関する変数のうち、任意の変数の回帰係数が 0 に推定される傾向にあり、グループ効果を持ちません。そのため、Lasso 回帰と Ridge 回帰では、異なる解パス図が得られます(図 4)。

Elastic Net 回帰

Lasso 回帰のスパース性と Ridge 回帰のグループ効果を併せ持つ Elastic Net 回帰という手法もよく知られています。式(5)に示すように、L1 および L2 ノルムの重み付き和を罰則項とします。ここで、ハイパーパラメータ α が重みです。

$$L_{\lambda}(\beta) = \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \quad (5)$$

図 5 に Elastic Net 回帰の予測結果および標準回帰係数を示します。今回は Lasso 回帰と同程度にスパースな解が得られました。なお、Elastic Net 回帰ではハイパーパラメータが 1 つ増加しているため、一般により多くの計算時間が必要になります。

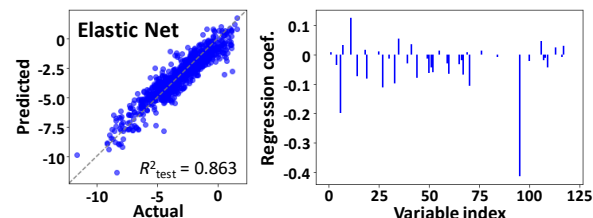


図 5 Elastic Net 回帰の予測結果
および標準回帰係数

おわりに

上述の正則化法では回帰係数に罰則を課します。その結果、スケールの大きな説明変数の回帰係数は強く縮小されることになるため、一般的にはデータ前処理としてスケーリングを行うことが推奨されます。次稿では、本稿で紹介できなかったスパースモデリングの適用例について紹介します。

参考文献

- 1) 永廣卓哉: データ解析入門 16 <部分的最小二乗法>、ORIST テクニカルシート、No. 23-02 (2023)
- 2) <https://ja.wikipedia.org/wiki/ノルム> (accessed on October 7th, 2022)
- 3) J. S. Delaney, *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004)