重回帰分析と L2, L1 正則化

清家大嗣

2020 年 5 月 25 日

1　重回帰分析について

重回帰分析とは, 一つの目的変数を複数の変数によって説明する多変量解析の一つである. 入力変数ベクトル $\boldsymbol{x}$ に対し, 出力変数 y が与えられるとし, 重みベクトル $\boldsymbol{w}$ を用いて, $y \approx \boldsymbol{x}^T\boldsymbol{w}$ で近似する. その誤差を評価するために, 誤差関数 $E[\boldsymbol{x}_1, y_1, ..., \boldsymbol{x}_n, y_n]$ を定義する (n は学習データ数を表す). 例えば, 誤差関数として二乗和誤差 $E_{\boldsymbol{w}}$ を次式で定義すれば, 誤差を最小化する $\boldsymbol{w}$ を解析的に計算できる (行列 X の各行成分は入力変数ベクトル転置 $\boldsymbol{x}_i^T$ である).

$$E_{\boldsymbol{w}} = \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w}) \tag{1}$$

ベクトル $\boldsymbol{w}$ の各成分 w_i に対して, スカラー関数である $E_{\boldsymbol{w}}$ は 凹 関数である. 従って, $E_{\boldsymbol{w}}$ をベクトル $\boldsymbol{w}$ により微分した結果が, 零ベクトルになる $\boldsymbol{w}$ が $E_{\boldsymbol{w}}$ の停留点であり最小値となる. 依って, 次式が零ベクトルとなる点が最小値となる (式変形に, 二次形式のベクトル微分公式 $\partial(\boldsymbol{w}^T A\boldsymbol{w})/\partial\boldsymbol{w} = (A + A^T)\boldsymbol{w}$, スカラー $(1 \times 1$ 行列) の転置は常に等号が成立すること, $\partial(\boldsymbol{a}^T\boldsymbol{w})/\partial\boldsymbol{w} = \partial(\boldsymbol{w}^T\boldsymbol{a})/\partial\boldsymbol{w} = \boldsymbol{a}$ を用いた).

$$\begin{aligned}\frac{\partial E_{\boldsymbol{w}}}{\partial\boldsymbol{w}} &= \frac{1}{2}\frac{\partial}{\partial\boldsymbol{w}}\left(-\boldsymbol{w}^T X^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{w} + 2X^T X\boldsymbol{w}\right) \\ &= -X^T\boldsymbol{y} + X^T X\boldsymbol{w}\end{aligned} \tag{2}$$

つまり, $\boldsymbol{w} = (X^T X)^{-1}X^T\boldsymbol{y}$ が誤差関数を最小化する最適な解となる. ここで, 逆行列が安定して求められる必要がある. 例えば, 行列 X が正則でない場合, 行列式の積に関する公式 $|AB| = |A||B|$ より $X^T X$ も正則でなくなり逆行列を安定に求められない. これは, 多重共線性 (マルチコリニアリティ: multicollinearity) [1] などの, 説明変数間に相関係数が高い組み合わせがある場合などに生じる. 言い換えれば, 連立方程式を解く式が不足するという状況である.

また, 説明変数が多すぎる場合に過学習と呼ばれる問題も発生する. 図 1 にあるようにデータ数とほとんど同等な数の説明変数がある (図は PRML [2] の多項式フィッティングの項と同様な設定で作成した) と, フィッティング対象である sin と似ても似つかないような関数がモデルとして生成されてしまう. このような問題を防ぐために, 誤差関数に正則化項を加える手法が存在する.

1.1　L2 正則化 (Ridge 回帰)

誤差関数に Lq 正則化項を加えることで, 上記の過学習を防ぐことができる. Lq 正則化項は, d 次元の重みベクトル $\boldsymbol{w}$ に対し, 適当な正の実数 λ を用いて $\lambda\{(|w_1|^q + |w_2|^q + ... + |w_d|)^{1/q}\}^q$ により定義される. 例えば, L2 正則化では, 誤差関数は次式のように定義できる.

$$E_{\boldsymbol{w},\mathrm{L2}} = \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{3}$$

(2) 式と同様に, d 次元重みベクトル $\boldsymbol{w}$ で微分すると次式が得られる (I は $d \times d$ の単位行列).

$$\frac{\partial E_{\boldsymbol{w},\mathrm{L2}}}{\partial \boldsymbol{w}} = -X^T\boldsymbol{y} + X^TX\boldsymbol{w} + \lambda I\boldsymbol{w} \tag{4}$$

つまり, $\boldsymbol{w} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$ が L2 正則化項を加えた誤差関数を最小化する最適な解となる. 図 1 を見ると, より sin 曲線に近い形となっていることが分かる ($\ln\lambda = -18$ とした).

1.2 L1 正則化 (Lasso 回帰)

L2 正則化と同様にして, L1 正則化項を加えた誤差関数は次式で定義できる.

$$E_{\boldsymbol{w},\mathrm{L1}} = \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w}) + \lambda\sum_{i=1}^{d}|w_i| \tag{5}$$

この式は, $w_i = 0$ で微分不可能なことや, w_i の正負によって誤差関数の挙動が変わるため, 一般的な解析解を得ることはできない. そのため, 本稿では劣微分 (微分不可能な点 $w_i = 0$ の微分値を $[-1, 1]$ のどちらか決め打ちで再定義した微分の概念を拡張したもの) を用いた勾配降下法を上記の誤差関数に適応した上で, L1 正則化が持つスパース性 ($w_i = 0$ となる解を得られるケースが多いこと) について理解したい.

問題の単純化のため, 誤差関数の正則化項でない項が各 w_i に対して下に凸な関数であると仮定する ((5) 式もその仮定を満たす). 勾配降下法を実行するため $w_i = w_s > 0$ における勾配を考えると, 下に凸な関数 $f(\cdot)$ を用いて次式のように書ける.

$$\frac{\partial E_{\boldsymbol{w},\mathrm{L1}}}{\partial w_i}|_{w_i=w_s} = f'(w_s) + \lambda \cdot 1 \tag{6}$$

(6) 式が 0 となるような点を w_e と定義して, $w_e < 0$ となる場合, その点は実際には極小値とはならない可能性がある. なぜならば, (6) 式は誤差関数を $E_{\boldsymbol{w},\mathrm{L1}} = \frac{1}{2}(\boldsymbol{y}-X\boldsymbol{w})^T(\boldsymbol{y}-X\boldsymbol{w}) + \lambda\sum_{i=1}^{d}w_i$ と定義した場合の極小値を求めたに過ぎないからである. 実際の w_i から w_e までの増分は次式で定義される ($w_i = 0$ に微分不可能な点が存在するが, 劣微分の概念を導入したことに注意する).

$$\Delta E_{\boldsymbol{w},\mathrm{L1}} = \int_{w_s}^{0}(f'(w_i) + \lambda)dw_i + \int_{0}^{w_e}(f'(w_i) - \lambda)dw_i \tag{7}$$

$f'(w_i)$ が単調増加関数であること, $f'(w_e) = -\lambda$ であることから, 第一項目は必ず負となる. 第二項目は, $f'(w_i) > \lambda$ で負となるが, 十分に大きな λ を用いれば, 正になる. つまり, $w_i = 0$ で極小値となる. これが, L1 正則化による解がスパース性を持つ理由である. $w_s < 0$ の場合にも, 同様に考えることができる.

一方, L2 正則化の場合, (7) 式のように w_e の値によって, 0 が極小点となる可能性はない. L1 正則化の (7) 式と同様に考えると, L2 正則化が現在の点から極小点まで移動するまでの変化量は次式でかける ($f'(w_e) + \lambda w_e = 0$ と $f''(w_i) + \lambda > 0$ に注意する ($f(\cdot)$ は下に凸な関数)).

$$\Delta E_{\boldsymbol{w},\mathrm{L2}} = \int_{w_s}^{w_e}(f'(w_i) + \lambda w_i)dw_i \tag{8}$$

参考文献

[1] 「線形回帰の過学習を抑えよう ～Ridge 回帰と Lasso 回帰～」 https://www.n-insight.co.jp/niblog/20190917-1351/.

[2] C.M. Bishop. 元田浩, 栗田多喜夫他訳, 吉識知明訳. 2006. 『パターン認識と機械学習』 発行: 丸善出版株式会社, 編集: 主プリンガー・ジャパン株式会社.

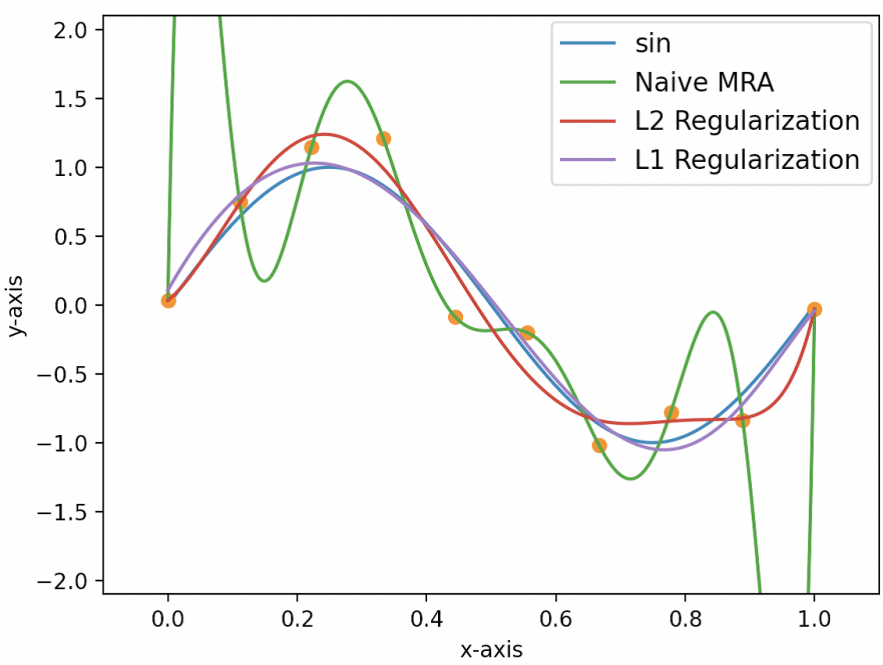

図 1: 重回帰分析 (Multiple Regression Analysis), L2, L1 正則化による多項式曲線フィッティングの例 (PRML [2] を参考に作図). L2 正則化では, $\ln \lambda = -18$ とし, L1 正則化では, $\ln \lambda = -6$ とした. L1 正則化では解析解が求まらないので適当な初期値に対し勾配降下法を用いて重みベクトル $\boldsymbol{w}$ を計算した.