Technical Sheet
No. 23-11

データ解析入門 19 <サポートベクター回帰>

キーワード：SVR、カーネルトリック、カーネル関数

はじめに

本稿では、サポートベクター回帰（SVR; Support vector regression）と呼ばれる回帰モデルを紹介します。SVR は、カーネルトリックと呼ばれる技術を用いることで、非線形回帰モデルへと拡張することができます。カーネルトリックにより、表現力の高い予測モデルの構築が可能になります。

SVR の誤差関数

一般に、回帰モデルの構築では、何らかの損失関数を準備し、損失を最小にするパラメータを求めます。例えば、重回帰分析（MLR）を行う場合、損失関数 $L(\boldsymbol{\beta})$に残差平方和を用い、損失関数が最小になる回帰係数$\boldsymbol{\beta}$を求めます［式(1)］[1]。式(1)の N、$\boldsymbol{y}$、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。

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

ただし、MLR では多重共線性などに注意する必要があります。そのため、回帰係数の大きさにペナルティを課す Ridge 回帰などが提案されています。Ridge 回帰の誤差関数は、残差平方和に回帰係数ベクトルの L2 ノルムの 2 乗を加えた形をしており［式(2)］、予測モデルの当てはまりを大きく損なうことなく、絶対値が極端に大きな回帰係数が算出されることを防ぎます。なお、λは正則化パラメータです。

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

一方、SVR では、式(2)を少し変形した損失関数を用います［式(3)］。ここで、C および ε は SVR のハイパーパラメータ、添え字 i はサンプルインデックス、$f(\boldsymbol{x}_i)$は i 番目の予測値です。

$$L_{C,\varepsilon}(\boldsymbol{\beta}) = C\sum_{i=1}^{N}\max\{0, |y_i - f(\boldsymbol{x}_i)| - \varepsilon\} + \|\boldsymbol{\beta}\|_2^2 \quad (3)$$

式(2), (3)の右辺第 1 項はいずれも予測誤差に関する項ですが、関数の形が異なります（図 1）。SVR の損失関数第 1 項（誤差関数）には、原点付近に平らな領域があり、原点から一定以上離れると直線的に誤差が増加します。一般的に、このような特徴を有する誤差関数は、残差平方和よりもノイズや外れ値に頑健であるといえます。また、この予測誤差をゼロとする範囲は ε チューブなどと呼ばれ、ε はクロスバリデーションなどで決定します。

なお、SVR は、基本的に MLR や Ridge 回帰と同様の線形回帰モデルの仕組みを持っています。線形 SVR の回帰係数は制約付き最適化問題を解くことで算出可能です。SVR の数理的な導出などについては文献[2]が参考になります。

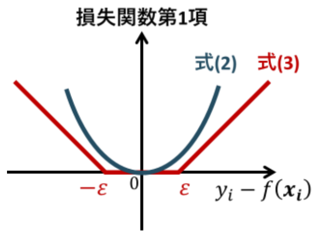

図 1　損失関数第 1 項（誤差関数）

カーネルトリック

目的変数と説明変数は必ずしも線形な相関を有するとは限らず、非線形性の強いデータを解析する場面も想定されます。線形回帰モデルは仕組みがシンプルであり、予測モデルを解釈しやすいといったメリットを持ちます。一方で、非線形回帰モデルの解釈性は必ずしも高くないですが、高い予測性能を期待できます。

基本的に SVR は線形回帰モデルですが、非線形回帰モデルに拡張することが可能です。予測に非線形性を反映するため、まず何らかの関数 $\Phi(\cdot)$

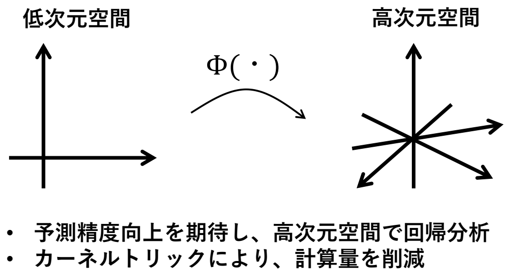

図 2　カーネルトリックについて

地方独立行政法人
大阪産業技術研究所　本部・和泉センター
https://orist.jp/
〒594-1157　和泉市あゆみ野 2 丁目 7 番 1 号
Phone: 0725-51-2525（総合受付）

で説明変数を高次元空間に写像し、非線形に写像した説明変数を用いて予測モデルを構築します。ただし、写像先の高次元空間の次元が高すぎると、計算量が膨大になってしまいます。次に紹介するカーネルトリックという手法を用いると、写像先が無限次元であっても計算できることがあります(図 2)。

本来、非線形 SVR のモデル構築のためには、高次元ベクトルの内積計算をしなければならないですが、カーネルトリックにより内積計算を、ある関数への元の説明変数の代入に置き換えることができます。つまり、カーネルトリックにより大幅な計算量削減が実現します。ただし、カーネルトリックを適用できる関数(カーネル関数)は次の条件を満たす関数に限られます。

1. 任意の $\boldsymbol{x_i}, \boldsymbol{x_j}$ に関して$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = k(\boldsymbol{x_j}, \boldsymbol{x_i})$である。
2. カーネル関数$k(\boldsymbol{x_i}, \boldsymbol{x_j})$を成分とするグラム行列 K が半正定値行列[3]である。

$$K = \begin{bmatrix} k(\boldsymbol{x_1}, \boldsymbol{x_1}) & \cdots & k(\boldsymbol{x_1}, \boldsymbol{x_N}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x_N}, \boldsymbol{x_1}) & \cdots & k(\boldsymbol{x_N}, \boldsymbol{x_N}) \end{bmatrix}$$

なお、機械学習モデルのユーザーの立場では、数学的背景をあまり気にすることなく、上記条件を満たすカーネル関数を利用することができます。よく用いられるカーネル関数として、RBF(Radial basis function)カーネルや多項式カーネルなどがあります[式(4), (5)]。なお、i および j はサンプルインデックス、γ および d は各カーネル関数のハイパーパラメータです。カーネル関数のハイパーパラメータはクロスバリデーションで調整できます。

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\gamma \|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2\right) \tag{4}$$

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(1 + \boldsymbol{x_i^T} \boldsymbol{x_j}\right)^d \tag{5}$$

SVR の適用例

図 3 に線形 SVR および非線形 SVR の適用例を示します。ここでは、sin カーブをもとに生成した人工データ(図中青点)を回帰分析します。図 3 に示すように、線形 SVR では上手く予測できていませんが、RBF カーネルを用いた非線形 SVR では良好に予測できています(図中赤線)。このようにカーネル関数を用いることで表現力の高い回帰モデルを構築できます。ただし、表現力の高い非線形回帰モデルは過剰適合しやすいため、予測モデルの適用範囲には注意しなければなりません。

図 4 に、コンクリートのスランプ試験に関するデータセット[4]に各種回帰手法を適用した結果を示します。本データセットは 7 つの説明変数および 3 つの目的変数から構成されますが、今回はコンクリートの圧縮強度のみを目的変数として予測しました。検討した回帰モデルの中では、非線形 SVR の予測精度が良好でした。

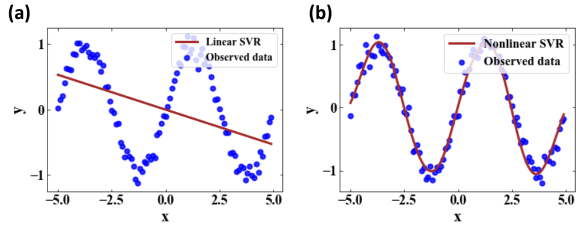

図 3 (a) 線形、(b) 非線形 SVR の適用例

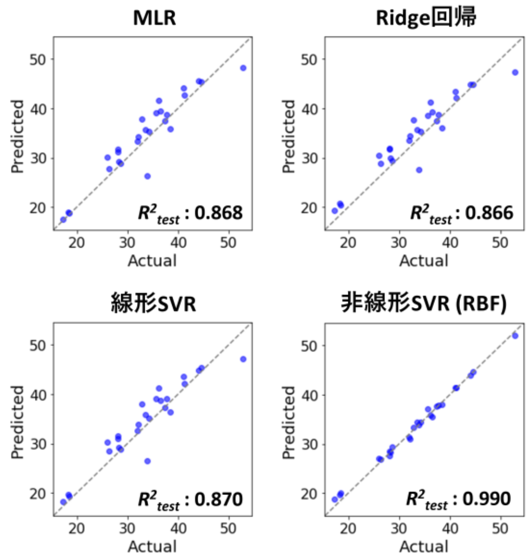

図 4 各種回帰手法の適用結果

おわりに

本稿で紹介した SVR は、非常に有用な回帰手法であり、便利な機械学習ライブラリも公開されています。また、SVR のハイパーパラメータを高速に最適化する方法も提案されています[2]。

参考文献

1) 永廣卓哉:データ解析入門 17 <Lasso・Ridge・Elastic Net>、ORIST テクニカルシート、No. 23-06 (2023)

2) 金子弘昌、化学のための Python によるデータ解析・機械学習入門、オーム社 (2019)

3) https://ja.wikipedia.org/wiki/行列の定値性 (accessed on October 23rd, 2023)

4) http://archive.ics.uci.edu/ml/machine-learning-databases/concrete/slump/slump_test.data (accessed on March 20th, 2023)

発行日　2024 年 1 月 1 日

作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@orist.jp