4M-03

正則化最小二乗法を用いた線形基底関数モデルに対する予測アルゴリズム

倉持　七海†　　須子統太‡

早稲田大学社会科学部†　　早稲田大学社会科学総合学術院‡

1　はじめに

現在,機械学習における予測モデルでは，ニューラルネットワークのように表現力の高い複雑な関数系を仮定し,データから学習するアプローチが主流である．しかし，予測精度の高いモデルを学習するには大量のサンプルが必要となり，パラメータのチューニングが難しいなどの課題がある．

本研究では，正則化最小二乗法を利用することで，様々な基底関数を動的に選択しながら予想精度の高いモデルを構築するアルゴリズムを提案する．そのもとで，いくつかの実データを用いて提案アルゴリズムの評価を行う．

2　多項式回帰モデルに対する予測アルゴリズム

井上らは多項式回帰における最大次数が未知の状況で,次数選択を行いながらモデルを学習するアルゴリズムを提案した.[1]

はじめに,各変数X, Y, βの定義を行う．i番目のデータにおける説明変数ベクトルを$\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \cdots, x_{ip})^T$とし,$X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n})^T$とする.$i$番目のデータにおける目的変数を$\boldsymbol{y_i}$とし,$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$とする.また,パラメータベクトルを,$\boldsymbol{\beta} = (\beta, \cdots \beta_j, \cdots \beta_J)$で表す.

$\Omega(\mathrm{A}, \mathrm{S})$を元の行列$A$から,インデックス集合$S$に含まれる列のみ抽出した行列を抽出する関数と定義する.次に,$expan(A', A)$は,後ろの行列の各列を前の行列の各列に掛けて,列ベクトルを得た後,それらを結合して得られる新しい行列と定義する.例えば,

$$A' = \begin{bmatrix} a'_{11} & a'_{12} \\ \vdots & \vdots \\ a'_{n1} & a'_{n2} \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} \\ \vdots & \vdots \\ a_{n1} & a_{n2} \end{bmatrix}, \quad (1)$$

とした時,

$$expan(A', A) = \begin{bmatrix} a'_{11}a_{11} & a'_{11}a_{12} & a'_{12}a_{11} & a'_{12}a_{12} \\ \vdots & \vdots & \vdots & \vdots \\ a'_{n1}a_{n1} & a'_{11}a_{n2} & a'_{n2}a_{n1} & a'_{n2}a_{n2} \end{bmatrix}, \quad (2)$$

となる.

以上の関数を用いて,井上らのアルゴリズムを図１で示す.

Step1: $\widehat{\boldsymbol{\beta}} = argmin||\boldsymbol{y} - \boldsymbol{\beta}X||_2^2 + \lambda \sum_j |\beta_j|$
Step2: $S = \{j : \widehat{\boldsymbol{\beta}} \neq 0\}$
Step3: $X_{new} = X$
Step4: $X' = \Omega(X_{new}, \mathrm{S})$
Step5: $X_{new} = expan(X', X)$
Step6: $\widehat{\boldsymbol{\beta}} = argmin||y - \boldsymbol{\beta}X_{new}||_2^2 + \lambda \sum_j |\beta_j|$
Step7: $S = \{j : \widehat{\boldsymbol{\beta}} \neq 0\}$
Step8:終了条件を満たすまで,Step4 に戻る
Step9:予測モデル,$\hat{y} = \widehat{\boldsymbol{\beta}} \boldsymbol{x''}_{new}$を出力する

図１　井上アルゴリズム

ここで$\boldsymbol{x''}_{new}$は予測対象の説明変数に対するX_{new}と同じ基底関数を用いた基底関数ベクトルとする.

3　基底関数の拡張

本研究では，井上らのアルゴリズムを拡張することで,線形基底関数モデルにおける基底選択を行うアルゴリズムを作成する.多項式の項を基底関数の項ととらえ,さまざまな基底関数を動的に生成し,繰り返し選択することで,予測モデルを構築する.今回,基底関数のパターンは以下の３種類を考える.

3.1　*TypeA*

井上らのアルゴリズムにおける,Step1 のX と Step5 の$expan(X', X)$のXを,$\tilde{X}$で置き換える.ここで$\tilde{X}$は以下で定義する.

$$\tilde{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} & \log X_{11} & \cdots & \log X_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} & \log X_{n1} & \cdots & \log X_{np} \end{bmatrix} \quad (3)$$

これにより,説明変数に対数をとった基底が追加される.

3.2　*Type*B

A prediction algorithm for linear basis function models using regularized least squares
†Nanami Kuramochi, School of Social Sciences, Waseda University
‡Tota Suko, Faculty of Social Sciences,Waseda University

井上らのアルゴリズムの Step5 の$expan(X', X)$を以下の$expan'(X', X)$で置き換える.

$$expan'(A', A) = \begin{bmatrix} a'_{11}a_{11} & a'_{11}a_{12} & a'_{12}a_{11} & a'_{12}a_{12} & \frac{a'_{11}}{a_{11}} & \frac{a'_{11}}{a_{12}} & \frac{a'_{12}}{a_{11}} & \frac{a'_{12}}{a_{12}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a'_{n1}a_{n1} & a'_{11}a_{n2} & a'_{n2}a_{n1} & a'_{n2}a_{n2} & \frac{a'_{n1}}{a_{n1}} & \frac{a'_{n1}}{a_{n2}} & \frac{a'_{n2}}{a_{n1}} & \frac{a'_{n2}}{a_{n2}} \end{bmatrix}. \quad (4)$$

これによりマイナスの次数を持つ基底を表現できる.

3.3 *TypeC*

井上らのアルゴリズムにおける, Step1 のX を$\tilde{X}$で置き換え, Step5 の$expan(X', X)$を$expan'(X', \tilde{X})$で置き換える. これにより, $TypeA$の対数の基底と, $TypeB$のマイナスの次数を持つ基底の両方を表現できる.

4 実データによる実験

4.1 実験方法と結果

新たに作成したアルゴリズムの有効性を比較するために, UCI Machine Learning Repository[2]のベンチマークデータに対し, 実験を行った. 各データセットに対し, 5 分割のクロスバリデーションを行い, 平均二乗誤差を評価した. 実験結果を表 1 に示す.

4.2 考察

提案アルゴリズムは, 表現できるモデルのクラスが LASSO や井上のアルゴリズムよりも大きい. そのため, 各クラスで予測誤差を最小にする最適なモデルが選択できていれば, 提案アルゴリズムの方が予測誤差を小さくできると考えられる. しかし, LASSO を用いた変数選択を繰り返しているため, 必ずしも最適なモデルの選択はできていない. そのため, 予測誤差にばらつきが生じている. ただし, 井上アルゴリズムや LASSO に比べ, 予測誤差を小さくするデータも存在しており, 比較的良い変数の選択ができていると考えられる.

5 おわりに

本研究では, 従来アルゴリズムを拡張することで, 様々な基底関数を動的に選択しながら予想を構築するアルゴリズムを提案した. ベンチマークデータによる評価実験で提案アルゴリズムの有効性を示した.

表 1 各データセットの平均二乗誤差

	Yacht Hydrodynamics	Auto MPG
サンプルサイズ	308	392
変数の数	6	5
LASSO	79.74	23.25
RandomForest	20.10	17.67
井上アルゴリズム	12.71	20.54
提案アルゴリズム *TypeA*	9.59	20.26
提案アルゴリズム *TypeB*	16.98	19.15
提案アルゴリズム *TypeC*	19.91	19.36

	ConcreteSlump Test(Slump)	ConcreteSlump Test(Flow)	ConcreteSlump Test(Strength)
サンプルサイズ	103	103	103
変数の数	7	7	7
LASSO	64.57	175.87	8.24
RandomForest	60.63	212.15	26.14
井上アルゴリズム	46.17	183.79	0.96
提案アルゴリズム *TypeA*	79.19	179.40	6.35
提案アルゴリズム *TypeB*	73.71	204.05	66.91
提案アルゴリズム *TypeC*	64.91	180.72	5.78

謝辞

本研究の一部は, 日本学術振興会科学研究費基盤研究(C)一般(No. 21K11796)により行われた.

参考文献

[1]井上一磨, 清水良太郎, 須子統太, 後藤正幸 ”最大次数が未知の多項式回帰モデルに対するスパース推定に関する一考察” 信学技報, IBISML2018-87, pp. 321-328, 2018

[2] “UC Irvine Machine Learning Repository” UCI Machine Learning Repository (参照 2023-01-11)