多項ロジットモデルの変数選択問題に対する整数最適化手法の構築

神谷　俊介

東京農工大学大学院工学府情報工学専攻（現：株式会社 NTT データ数理システム）
指導教員：宮代隆平　東京農工大学　准教授

1. はじめに

多項ロジット (multinomial logit; MNL) モデルは離散選択モデルの一つであり，人が複数のカテゴリから一つを選択する行為をモデル化したものである．このモデルは重み行列 $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m) \in \mathbb{R}^{m \times p}$ をもち，それを用いて各カテゴリに対する選択確率を予測する．ここでは，教師あり学習の標準的状況を仮定し，人 i は説明変数ベクトル $\boldsymbol{x}_i \in \mathbb{R}^p$ で表され，カテゴリ $y_i \in \{1, 2, \ldots, m\}$ を選択したことが観測済みであるとする．重み行列の推定にはたとえば観測データを用いた最尤推定が用いられるが，過適合によりモデルの予測精度が低下する問題点が指摘されている．そのため，重み行列のいくつかの要素を 0 にしてモデルの表現能力を削減することが重要となる．

変数選択はモデルの表現能力を削減するための一つの方法であり，負の対数尤度関数として次式で定義される損失関数を，説明変数集合の基数制約下で最小化する変数選択問題により実現される．

$$\ell^{\mathrm{MNL}}(y, \boldsymbol{\eta}) = -\eta_y + \log \sum_{s=1}^{m} \exp(\eta_s).$$

ここで，$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m) \in \mathbb{R}^m$ は各カテゴリに対する好ましさを表す．観測済みの人 i については，好ましさの予測値は $\eta_r := \boldsymbol{w}_r^\top \boldsymbol{x}_i$ となり，損失は $\ell^{\mathrm{MNL}}(y_i, \boldsymbol{\eta})$ と算出される．ただし，この損失関数の非線形性と基数制約の整数性から，MNL モデルの変数選択問題に対する効率的な厳密解法は知られていない．代わりに，これまで L_1 正則化手法をはじめとした多くの凸近似手法が提案されてきた．これらの手法は求解時間が小さいが，得られる推定量の質が悪い場合が存在することが知られている．

その一方で近年，MNL モデルの特殊ケースの一つである 2 項ロジット (binomial logit; BNL) モデルについて，整数最適化を用いた厳密解法に近い手法が提案されてきている．たとえば Sato et al. [1] は目的関数を区分線形関数で近似することで，元問題を混合整数線形計画問題に近似した．彼らの手法は説明変数の個数 p が 33 以下のインスタンスに対し現実的な実行時間で近似モデルに対する最適解を発見し，ヒューリスティクスの一つであるステップワイズ法に比べ高い予測精度を示す．また Bertsimas et al. [2] は目的関数に L_2 正則化項を加えることで，変数選択問題のより大規模なインスタンスが外部近似法を用いて効率的に求解可能であることを示した．ここで外部近似法とは，一般の混合整数非線形計画問題に対する汎用解法の一つである．

Bertsimas et al. の手法は既存手法である L_1 正則化手法に比べよい推定量を十分高速に出力するが，いくつかの問題点を有している．まず損失関数に対して凸性だけを仮定しているが，たとえば損失関数の共役関数が複数の停留点をもつ場合にアルゴリズムが動作しない．次に，実用的に興味のある L_2 正則化パラメータと予測精度や実行時間の関係が不明である．そこで本研究では，この手法に対し多クラス分類モデルの変数選択問題への拡張を施したうえで，上述の問題点を解決することを目指した．

2. 外部近似法を用いた整数最適化

多クラス分類モデルの変数選択問題について，目的関数に L_2 正則化項を加えることを考える．この最適化問題は，最適な変数集合を求める最適化問題と L_2 正則化最尤推定の 2 段階計画問題に等価変形できる．提案手法は，このうち前者の最適化問題に対して外部近似法を適用することで実現される．具体的には，L_2 正則化最尤推定の最適値を関数値とする関数 c を用いて，変数選択問題を次式で定式化した．

$$\text{(BS)} \quad \left| \quad \underset{\boldsymbol{z} \in S_k^p}{\text{minimize}} \quad c(\boldsymbol{z}). \right.$$

ここで $\boldsymbol{z}$ は説明変数の取捨を規定する 0–1 変数であり，$S_k^p = \{\boldsymbol{z} \in \{0,1\}^p \mid \sum_{j=1}^{p} z_j \leq k\}$ は説明変数集合の基数制約を表す．外部近似法は混合整数線形計画問題を繰り返し解く手法であり，非線形関数の各整

数点での 1 次近似である外部近似を各反復で逐次的に追加していき元問題の最適解を得る．最適化問題 (BS) に対しては，第 t 反復目に以下の混合整数線形問題を解く手法として定義できる．

$$
(\widetilde{\mathrm{BS}}_t) \left| \begin{array}{ll} \underset{\boldsymbol{z}\in S_k^p,\, \zeta\in\mathbb{R}}{\text{minimize}} & \zeta \\ \text{subject to} & \zeta \geq c(\boldsymbol{z}_i) + \nabla c(\boldsymbol{z}_i)^\top (\boldsymbol{z} - \boldsymbol{z}_i) \\ & (i = 1,\, 2, \ldots,\, t). \end{array} \right.
$$

数値実験では，現実問題を反映したものとランダム生成したものから成るインスタンスセットを用い，L_1 正則化手法との比較を行った．その結果を受け，上述の拡張が予測精度や実行時間の意味で有効であることを示した．

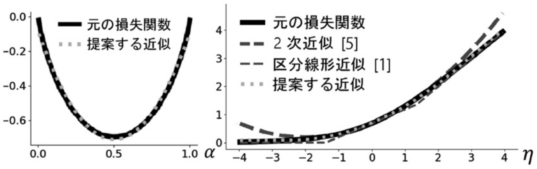
3. 収束性の解析

提案手法が最適解を得るための十分条件は，Bonami et al. [3] が述べる外部近似法の収束に必要な仮定を利用して整理した．外部近似法が KKT 条件を用いて収束性が示される都合から，関数 c の凸性や連続的微分可能性が要求される．これらの条件は凸解析の枠組みで議論でき，特に損失関数が実数値関数の場合は，損失関数とその共役関数に関する条件で記述し直せる．たとえば勾配 ∇c の連続性は，L_2 正則化最尤推定の双対問題に対する最適解写像の連続性が本質的となる．そこで凸解析における結果を用いると，この条件は損失関数の共役関数に対する狭義凸性と連続性に言い換えられる．これらの解析結果は，理論的に不完全であった Bertsimas et al. の結果を補完している．

本論文では特に MNL モデルに対して，損失関数 ℓ^{MNL} とその共役関数が上記の条件を満たし，提案手法で最適解が得られることを証明した．

4. アルゴリズムの高速化

提案手法を高速化する手段として，損失関数 ℓ^{MNL} に対する近似手法を提案する．本稿 2 節で述べたとおり提案手法は混合整数線形計画問題の反復解法であるが，外部近似を追加するたびに勾配 ∇c の計算が必要となるため，非線形計画問題を繰り返し解く必要がある．そして，この非線形性は損失関数の共役関数 $\hat{\ell}^{\mathrm{MNL}} : \{1, 2, \ldots, m\} \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ が対数関数を含むことに起因する．そこで，この関数を次式のとおり凸 2 次関数で近似し，最適化問題を凸 2 次計画問題に変形させることで計算効率の向上を図った．



(a) 共役関数　(b) 損失関数

図 1　提案する近似手法（BNL モデル）

$$
\begin{aligned}
\hat{\ell}^{\mathrm{MNL}}(y, \boldsymbol{\alpha}) &= \sum_{s \neq y} \alpha_s \log \alpha_s + (1 + \alpha_y) \log(1 + \alpha_y) \\
&\approx \sum_{s \neq y} \left(\frac{5}{2}\alpha_s^2 - \frac{5}{2}\alpha_s - \frac{1}{12} \right).
\end{aligned}
$$

ここで，実効定義域に関する記述は省略している．

提案する近似手法は，MNL モデルの損失関数を BNL モデルの損失関数の和で近似する Titsias [4] による既存手法と，BNL モデルの損失関数の共役関数に対する多項式回帰（図 1 (a)）で構成される．ここで後者の BNL モデルの損失関数に対する近似は，二つの既存手法の利点を併せもつ方法であると解釈できる（図 1 (b)）．これらの結果から，提案する近似がアルゴリズムの高速化に寄与し，かつ高い近似精度をもつと予想した．

数値実験では，本稿 2 節の実験に関して損失関数に近似を施しても同等の予測精度が実現されることを確認し，上述の予想が正しいことを実証した．

参考文献

[1] T. Sato, Y. Takano, R. Miyashiro and A. Yoshise, “Feature subset selection for logistic regression via mixed integer optimization,” *Computational Optimization and Applications*, **64**, pp. 865–880, 2016.

[2] D. Bertsimas, J. Pauphilet and B. V. Parys, “Sparse classification and phase transitions: A discrete optimization perspective,” arXiv: 1710.01352, 2017.

[3] P. Bonami, L. Biegler, A. Conn, G. Cornuéjols, I. Grossmann, C. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya and A. Wächter. “An algorithmic framework for convex mixed integer nonlinear programs,” *Discrete Optimization*, **5**, pp. 186–204, 2008.

[4] M. K. Titsias, “One-vs-each approximation to softmax for scalable estimation of probabilities,” *Neural Information Processing Systems 2016*, pp. 4161–4169, 2016.

[5] K. Tanaka and H. Nakagawa, “A method of corporate credit rating classification based on support vector machine and its validation in comparsison of sequential logit model,” *Transactions of the Operations Research of Japan*, **57**, pp. 92–111, 2014.