3B-03

店舗選択モデルの構築と整数計画法による変数選択

佐藤俊樹 † 高野祐一 ‡ 中原孝信 ♮
† 筑波大学システム情報工学研究科 ‡ 専修大学ネットワーク情報学部 ♮ 専修大学商学部

1 はじめに

現代の消費者は，最寄り品の購買に限ってもスーパーマーケット，総合スーパー，百貨店，コンビニエンスストア，そしてドラッグストアなど，様々な小売業態において購買機会を有する．また，その中でも同じカテゴリーの商品を取り扱う業態が多数存在するため，消費者にとって店舗のスイッチング・コストは低くなっている．このような状況の中で，小売店が消費者の店舗選択行動をはじめとする購買行動を把握することは，他店との差別化などの戦略を検討する上で重要な指針となるであろう．

これまでに店舗選択行動に関する研究は数多く行われてきているが，Pan and Zinkhan [1] は消費者の店舗選択行動の要因を商品レベル，店舗レベル，個人レベルの 3 つに分類した．商品レベルは，商品の品質，価格，品揃え，そしてチラシなど各商品が持つ要因で，店舗レベルは，規模，利便性，レジの待ち時間，サービスの品質などの要因である．そして個人レベルは，買い物時間や買い物の目的，デモグラフィックからなる要因である．

消費者が店舗を選択する際には，このような複数レベルの要因が相互に関係していると考えられるが，本研究では，消費者の店舗選択における商品レベルの要因を明らかにする．具体的には，消費者の複数店舗における購買状況を把握できるスキャンパネルデータを対象に，整数計画法による変数選択手法を利用して店舗選択モデルを構築する．本研究で扱う店舗選択モデルは，特定の店舗における購買行動と他店における購買行動の違いをモデル化しており，各店舗を選択する際に想起される商品群を明らかできる点が特徴である．

Variable Selection via Integer Programming for Developing Store Choice Model
Toshiki SATO†, Yuichi TAKANO‡ and Takanobu NAKAHARA♮
‡ Graduate School of Systems and Information Engineering, University of Tsukuba
‡ School of Network and Information, Senshu University
♮ School of Commerce, Senshu University

2 ロジスティック回帰モデル

本研究では佐藤ら [2] によって提案された，整数計画法によるロジスティック回帰モデルの変数選択手法を利用する．

スキャンパネルデータの各レコードに与えられた店舗ラベルのうち，分析対象とする 2 店舗を選択する．選択した 2 店舗のラベルを持つレコード $i = 1, 2, \ldots, n$ をサンプルとし，購買店舗の 2 値ラベル $y_i \in \{-1, 1\}$ を，購買商品の商品分類からなる p 次元ベクトル $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ によって予測する問題を考える．ロジスティック回帰モデルでは，以下のロジスティック関数を使用してラベルの生起確率をモデル化する：

$$\Pr(y \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-y(\boldsymbol{w}^\top \boldsymbol{x} + b))}. \quad (1)$$

偏回帰係数 $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)^\top$ と切片 b は以下の対数尤度関数が最大になるように求める：

$$\begin{aligned} \ell(b, \boldsymbol{w}) &= \log\left(\prod_{i=1}^{n} \Pr(y_i \mid \boldsymbol{x}_i)\right) \\ &= -\sum_{i=1}^{n} \log(1 + \exp((-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b))) \\ &= -\sum_{i=1}^{n} f(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)). \end{aligned} \quad (2)$$

ただし，$f(v) = \log(1 + \exp(-v))$ とし，この関数はロジスティック損失関数と呼ばれる．

3 変数選択問題の定式化

多くの情報量規準は，説明変数の部分集合を $S \subseteq \{1, 2, \ldots, p\}$ として

$$-2 \max\{\ell(b, \boldsymbol{w}) \mid w_j = 0 \ (j \notin S)\} + g(n)(|S| + 1) \quad (3)$$

で表される．$g(n) = 2$ の場合は AIC に，$g(n) = \log(n)$ の場合は BIC に対応する．これらの情報量規準の値が小さい S が望ましいとされている．

0-1 決定変数 $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)^\top$ を導入し，$j \in S$ のとき $z_j = 1$，それ以外のとき $z_j = 0$ とする．これに

より，情報量規準 (3) を最小化する変数集合 S を決定する問題は次のように書ける：

$$\min_{b,\boldsymbol{w},\boldsymbol{z}}.\ 2\sum_{i=1}^{n} f(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)) + g(n)\left(\sum_{j=1}^{p} z_j + 1\right) \tag{4}$$

$$\text{s. t. } z_j = 0 \Rightarrow w_j = 0 \quad (j = 1, 2, \ldots, p), \tag{5}$$

$$z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p). \tag{6}$$

制約 (5) は，タイプ 1 の特殊順序集合 (SOS1) 制約による表現が可能である．SOS1 制約は，集合の要素のうち非ゼロの値をとるのは高々1 つであることを意味する．よって，$j = 1, 2, \ldots, p$ に対して集合 $\{1 - z_j, w_j\}$ に SOS1 制約を課せばよい．SOS1 制約は標準的な整数計画ソルバーに実装されている．

問題 (4)–(6) の目的関数は非線形であるためこのまま解くことは難しい．よって，本研究では目的関数を区分線形近似することで混合整数線形計画問題に変形する．

4 区分線形近似

点集合 $V = \{v_1, v_2, \ldots, v_m\}$ を与え，点 v_k におけるロジスティック損失関数の接線を次で表す：

$$h(v; v_k) = f'(v_k)(v - v_k) + f(v_k) \quad (k = 1, 2, \ldots, m).$$

するとロジスティック損失関数は次のように近似できる：

$$\begin{aligned} f(v) &\approx \max\{h(v; v_k) \mid k = 1, 2, \ldots, m\} \\ &= \min\{t \mid t \geq h(v; v_k)\ (k = 1, 2, \ldots, m)\}. \end{aligned}$$

たとえば $v_1 = -\infty, v_2 = -2, v_3 = 0, v_4 = 2, v_5 = \infty$ としたときの接線集合は図 1 のようになる．ただし，v_1, v_5 について以下が成立することに注意されたい．

$$h(v; v_1) = -v,\ h(v; v_5) = 0.$$

よって決定変数 $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^\top$ を導入し，ロジスティック損失関数を区分線形近似すると，問題 (4)–(6) は次のように変形できる：

$$\min_{b,\boldsymbol{t},\boldsymbol{w},\boldsymbol{z}}.\ 2\sum_{i=1}^{n} t_i + g(n)\left(\sum_{j=1}^{p} z_j + 1\right) \tag{7}$$

$$\begin{aligned} \text{s. t. }\ & t_i \geq f'(v_k)(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - v_k) + f(v_k) \\ & \qquad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m), \end{aligned} \tag{8}$$

$$z_j = 0 \Rightarrow w_j = 0 \quad (j = 1, 2, \ldots, p), \tag{9}$$

$$z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p). \tag{10}$$

これは混合整数線形計画問題であり，標準的な整数計画ソルバーで扱うことができる．

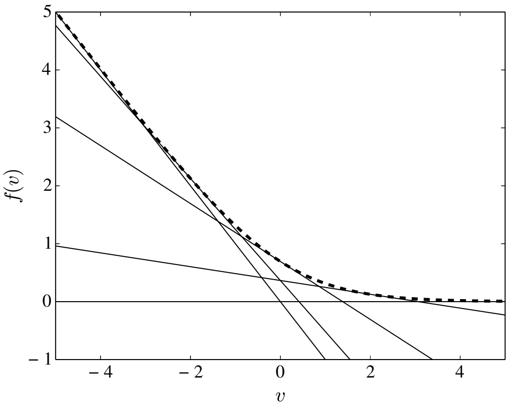

図 1: ロジスティック損失関数の区分線形近似

5 数値実験

本研究で利用したデータは，株式会社マクロミルより提供されたホームスキャン方式のスキャンパネルデータである．ホームスキャン方式では，各モニタの複数店舗における購買行動を記録することができ，このデータは，約 6,500 人のモニタによる 2012 年の 1 年間の食料品に関する購買履歴が記録されている．

このデータを使用して 5 分割交差検証を行った．ROC 曲線下の面積である AUC を予測精度指標とした．問題 (7)–(10) の求解には整数計画ソルバー Gurobi Optimizer 6.0.0 を使用し，統計解析ソフト R 3.1.2 の `step` 関数によるステップワイズ法と性能を比較した．実験結果の詳細については当日報告する．

謝辞

本研究の一部は，専修大学情報科学研究所の共同研究助成を受けたものである．

参考文献

[1] Pan, Y., and G.M. Zinkhan (2006), “Determinants of retail patronage: A meta-analytical perspective,” Journal of Retailing, Vol. 82, No. 3, pp. 229–243.

[2] 佐藤俊樹，高野祐一，宮代隆平，吉瀬章子 (2015),「混合整数最適化によるロジスティック回帰モデルの変数選択」，日本オペレーションズ・リサーチ学会 2015 年春季研究発表会アブストラクト集（掲載予定）．