ノンパラメトリック・ロジスティック回帰の統計理論

大阪大学大学院基礎工学研究科 屋良淳朝
大阪大学大学院基礎工学研究科, 理研 AIP 寺田吉壱

はじめに

深層学習は, 様々な実世界の複雑なタスクに対して極めて高い性能を発揮しており, 注目を浴びている. なぜ深層学習による推定量が他の推定量を優越するかの解明は, 我々が深層学習をどのように利用するのが適切かについての指針を与えるため, 重要な課題である. この課題を解決するために, 様々な観点から深層学習の理論研究が行われている. 特に, 回帰問題や分類問題を中心に, 深層学習を用いたノンパラメトリック推定量の理論的性質の解明が進んでいる.

回帰問題においては, 多くの研究が, 深層学習の優れた適応能力を明らかにしており, 深層学習がカーネル法などの他のノンパラメトリック推定量を優越する理由が部分的に解明されている. さらに, いくつかの研究では, 深層学習を用いた推定量は次元の呪いを回避できる可能性があることを示唆している（e.g., Schmidt-Hieber, 2020）.

分類問題では, 誤判別率の収束など判別性能に関する研究が進んでいる. 一般に, 分類問題では, 経験的な誤判別率の最小化を実施することは難しいため, 分類器（推定量）はヒンジ損失やロジスティック損失などの凸な代理損失を最小化することによって得られる. 近年, 様々な代理損失に対して, 深層学習を用いた分類法における誤判別率の収束レートが明らかとなり, 分類問題においても深層学習の利点が解明されている（e.g., Kim et al., 2021）.

一方で, ロジスティック回帰分析のように, 分類だけでなくクラス所属確率（条件付き確率）の推定を同時に行う場合も多い. このような場合, 推定された分類器の分類性能だけでなく, 条件付き確率の推定精度も分類問題における重要な話題の一つである. 通常, 深層学習による分類器は, 出力層にロジスティックシグモイド関数やソフトマックス関数を用いることで各クラスの所属確率を出力する. 条件付き確率の推定は, 単にクラスの所属を予測するよりも多くの情報を与えるので, 現実世界のデータ分析においても有用である.

本研究では, ロジスティック回帰による条件付き確率のノンパラメトリック推定を考える. 具体的には, 条件付き確率のノンパラメトリック最尤推定量（nonparametric maximum likelihood estimator, NPMLE）の理論的性質について解析する. NPMLE の一致性を示すには, 単に推定量と真の条件付き確率のロジスティック損失に関する差の期待値の収束を考えればよく, ロジスティック回帰においては推定量と真の条件付き確率の間の Kullback-Leibler (KL) 情報量の期待値を考えれば良い. しかし, 条件付き確率のノンパラメトリック推定では, 真の条件付き確率とその推定量の KL 情報量は容易に発散することが知られている. 具体的には, 推定量の台が真の条件付き確率の台を覆っていなければ発散する. 従って, ノンパラメトリックロジスティック回帰において KL 情報量に関する収束を導くためには,「真の条件付き確率が bounded away from zero である」といった仮定が必要となる（e.g., Ohn and Kim, 2022）.

そこで, 本研究では van de Geer (2000) のアプローチを用いて真の条件付き確率とその推定量の間の Hellinger 距離を直接評価し, 多クラスのロジスティック回帰分析におけるノンパラメトリック最尤推定量のオラクル不等式を導出する. Hellinger 距離の収束を考えることで, KL 情報量の評価に比べて, 現実的な仮定の下で推定量の性能を評価することができる. また, オラクル不等式の重要な応用として, 真のクラス条件付き確率が composition structured function（Schmidt-Hieber, 2020）であるときに深層学習による推定量の Hellinger 距離に関する収束レートを導出する. さらに, その収束レートがほとんどミニマックス最適であることを示す.

主結果

K クラスの分類問題を考える. $\mathcal{X} \subset \mathbb{R}^d$ を入力空間, $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^K$ をラベルの集合とする. ここで, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ は K 次元の標準基底ベクトルである. データ $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{X} \times \mathcal{Y}$ は以下のモデルから生成されると仮定する.

$$Y_k \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Bernoulli}(\eta_k(\boldsymbol{x})), \quad \boldsymbol{X} \sim P_{\boldsymbol{X}}, \quad k = 1, \ldots, K. \tag{1}$$

ここで, $\eta_k(\boldsymbol{x}) \coloneqq \mathbb{P}(\boldsymbol{Y} = \boldsymbol{e}_k \mid \boldsymbol{X} = \boldsymbol{x})$ は真のクラス条件付き確率であり, $P_{\boldsymbol{X}}$ は入力空間 $\boldsymbol{X}$ 上の未知の分布である．$\boldsymbol{X}$ と $\boldsymbol{Y}$ の同時分布を P と書く．$\mathcal{D}_n = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$ を母集団分布 P からのサイズ n の i.i.d. サンプルとする．本研究では, 条件付き確率 $\boldsymbol{\eta}$ のノンパラメトリック推定に焦点を当てる．

データ $\mathcal{D}_n$ が与えられたとき, 条件付き確率 $\boldsymbol{p} = (p_1(\boldsymbol{x}), \ldots, p_K(\boldsymbol{x}))^\top$ の負の対数尤度関数は

$$L_n(\boldsymbol{p}) \coloneqq -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik} \log p_k(\boldsymbol{X}_i) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i^\top \log \boldsymbol{p}(\boldsymbol{X}_i) \tag{2}$$

で与えられ, 最尤推定量 (maximum likelihood estimator, MLE) は $\hat{\boldsymbol{p}}_n \in \arg\min_{\boldsymbol{p}\in\mathcal{F}_n} L_n(\boldsymbol{p})$ で定義される．ここで, $\mathcal{F}_n$ は推定量の候補となる関数の集合 (仮説空間) である．損失関数 (2) に超過リスクは

$$\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{\eta}(\boldsymbol{X})^\top \log \frac{\boldsymbol{\eta}(\boldsymbol{X})}{\hat{\boldsymbol{p}}(\boldsymbol{X})}\right] = \mathbb{E}_{\boldsymbol{X}}[\mathrm{KL}(\boldsymbol{\eta}(\boldsymbol{X}) \,\|\, \hat{\boldsymbol{p}}(\boldsymbol{X}))] \tag{3}$$

で表される．ここで, $\mathrm{KL}(\cdot \,\|\, \cdot)$ は KL 情報量を表す．MLE $\hat{\boldsymbol{p}}_n$ の一致性を示すには, 超過リスク (3) を評価するのが自然に思えるが, モデル $\mathcal{F}_n$ として “小さい” クラスを考えたときでさえ超過リスク (3) が発散してしまう例が存在する．そのため，本研究では元々の超過リスク (3) の代わりに以下の量を評価する：

$$R(\boldsymbol{\eta}(\boldsymbol{X}), \hat{\boldsymbol{p}}(\boldsymbol{X})) \coloneqq \mathbb{E}_{\boldsymbol{X}}[H^2(\boldsymbol{\eta}(\boldsymbol{X}), \hat{\boldsymbol{p}}(\boldsymbol{X}))]. \tag{4}$$

ここで, H^2 は Hellinger 距離を表す.

上記の設定の下で，以下の不等式は真の条件付き確率と NPMLE との Hellinger 距離に関する評価を与える．詳細は Yara and Terada (2025) を参照されたい.

定理 1 (オラクル不等式)**.** K クラスの分類問題 (1) を考える．適当な条件の下で，普遍定数 c と $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ なる任意の $\delta \geq \delta_n$ に対して

$$\mathbb{E}_{\mathcal{D}_n}[R(\hat{\boldsymbol{p}}_n, \boldsymbol{\eta})] \leq 514(1 + c_0^2)(\delta^2 + R(\tilde{\boldsymbol{p}}_n, \boldsymbol{\eta})) + \frac{c^3}{n} \tag{5}$$

が成り立つ．ここで, $\tilde{\boldsymbol{p}}_n$ は $\mathcal{F}_n$ の任意の元であり，$\mathbb{E}_{\mathcal{D}_n}$ は学習データ $\mathcal{D}_n$ についての期待値を表す.

このオラクル不等式によって，様々な真の条件付き確率が属する関数クラスと推定モデルの設定のもとで NPMLE の収束レートを導出することができる．例えば，推定モデルとして適当な大きさの DNN を用いた場合，$\boldsymbol{\eta}$ が合成関数の構造を持つ場合（Schmidt-Hieber (2020) を見よ）は

$$\mathbb{E}_{\mathcal{D}_n}[R(\hat{\boldsymbol{p}}_n, \boldsymbol{\eta})] \lesssim \phi_n \log(n)^3, \quad \phi_n \coloneqq \max_{i=0,\ldots,q} n^{-\frac{\beta_i^*}{\beta_i^* + t_i}}, \quad \beta_i^* \coloneqq \beta_i \prod_{l=i+1}^{q} (\beta_l \wedge 1)$$

が成立する．この収束レートは入力の次元に依存しておらず，次元の呪いを回避していることがわかる．

紙面の都合上省略するが，当日は上記の収束レートのミニマックス最適性や，真の条件付き確率が非等方 Besov 空間に属する場合の収束レートについても紹介する．

参考文献

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

Ilsang Ohn and Yongdai Kim. Nonconvex sparse regularization for deep neural networks and its optimality. *Neural computation*, 34:476–517, 2022.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48:1875 – 1897, 2020.

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge university press, 2000.

Atsutomo Yara and Yoshikazu Terada. Nonparametric logistic regression with deep learning. *Bernoulli*, 2025. page in press.