

< 修 士 論 文 >

観測にバイアスを持つ状況下における 二値分類問題

滋賀大学大学院
データサイエンス研究科
データサイエンス専攻

修了年度：2022年度

学籍番号：6021142

氏 名：渡邊 翔太郎

指導教員：松井 秀俊

提出年月日：2023年1月10日

目次

1	はじめに	1
2	問題設定と既存手法	3
2.1	PN 分類	3
2.2	PU 分類	3
2.3	PUBN 分類	4
2.4	Pconf 分類	5
2.4.1	Pconf 分類に対する理論解析	5
3	提案手法	7
3.1	PbN 分類	7
3.2	Adjusted PbN 分類	8
3.3	PbN 分類に対する理論解析	9
4	数値実験	11
4.1	人工データ実験 1	11
4.1.1	設定	11
4.1.2	4つの状況下における実験	12
4.1.3	ϕ の推定にずれが生じたときの影響	13
4.2	人工データ実験 2	15
4.3	ベンチマークデータ実験	17
4.3.1	無線による屋内測定データセット	17
4.3.2	MNIST	19
5	結論	19
A	付録	24
A.1	PUBN 分類リスクの導出	24

A.2	Pconf 分類リスクの導出	24
A.3	補題 1 の導出	25
A.4	定理 1 の導出	26
A.5	PbN 分類リスクの導出	26
A.6	補題 2 の導出	27
A.7	定理 2 の導出	29

1 はじめに

二値分類の目的は、サンプルが正のクラスか負のクラスのどちらに属するかを識別することである。二値分類問題において、一般的な教師あり分類では、完全な正の (Positive; P) データと完全な負の (Negative; N) データを用いて分類器を学習していた。ここで、完全なデータとは、手元にあるデータから、それらを生成している分布の全体を表現できるものとする。この完全な P と N のデータを必要としている一般的な教師あり分類 (PN 分類) は、強い教師からの学習方法といえ、実際には、このような強い教師を用意することが難しい場面が多い。これに対し、弱い教師からの機械学習が近年盛んに研究されており、P データとラベルのない (Unlabeled; U) データから分類器を学習する PU 分類 (du Plessis et al. 2014; du Plessis et al. 2015; Elkan & Noto 2008; Kiryo et al. 2017), 半教師あり分類 (Chapelle et al. 2006; Sakai et al. 2017), ラベルノイズ学習 (Natarajan et al. 2013; Shi et al. 2018; Zhang et al. 2019) などがある。

本論文では、N データが多様であり包括的に特徴付けることができない状況や、それ以前に、N データの大多数が得られない状況を考える。この状況において、完全な N データを集めることは困難だが、N データの偏ったごく一部にラベルを付けるだけなら比較的容易である場合がある。UCI Machine Learning Repository (Bache & Lichman, 2013) から入手可能な無線屋内測位データセットを通して、実データ状況の一例を説明する。このデータセットは、4つの部屋で計測された7つの WiFi 信号の強度からなり、ある部屋 (P データ) の信号を他の3つの部屋 (N データ) の信号から識別したいとする。部屋ごとに信号の強さは異なるが、N データである3部屋のうち1部屋は観測されなかったり、3部屋の観測に偏りがあったりする可能性がある。このような場合でも、P の部屋と N の部屋を正しく分類したい。このような観測に偏りのある N (biased Negative; bN) データに対するアプローチとして、Li et al. (2010) と Fei & Liu (2015) は、ともにテキスト分類の文脈で問題解決に取り組んだ。Li et al. (2010) は、bN データが分類に悪影響を及ぼす可能性に言及し、bN データを完全に切り捨て、P データと U データを用いて分類を行った。一方で Fei & Liu (2015) は、偏りのない U データを集めることさえも難しいと考え、P と bN データの

みから分類器を学習した。ただ、この手法は、文書間の類似性の尺度に依存しているため、テキスト分類に特化している。これら2つの研究に対し、Hsieh et al. (2019) は、特定のドメイン知識を必要とせず、P, U, bN データに対して任意の分類器 (線形モデルから深層モデル) を当てはめることで、学習を行った (PUbN 分類)。ただ、Hsieh et al. (2019) では、P, U, bN データからの学習を P, U データからの学習に拡張しているが、P, bN データからの学習は想定していない。U データが得られていない場合、特徴量全体の分布を推定することが困難なため、分類器の構築が難しくなる。本論文では、PbN 分類として、PUbN 分類を、P, bN データからの学習に拡張し、U データが得られない場合でも分類器を学習できるようにする。

しかし、PbN 分類では、観測されているデータである事後確率を正確に求められない問題がある。この問題に対するアプローチとして、Pconf (Positive confidence) 分類における、観測されたデータが P である事後確率を表す、信頼度を陽に求められないことにより生じる、信頼度の歪みによる悪影響を修正する方策 (Shinoda et al., 2020) を組み込む。ここで Pconf 分類とは、N データや U データを必要とせず、信頼度を備えた P データのみから二値分類器を学習する学習方法である (Ishida et al., 2018)。本論文では、Shinoda et al. (2020) の考えを適用し、信頼度などがバイアスによって変形することを「歪み」と称する。Pconf 分類では、信頼度を求めるにあたり、専門家の知識が必要であった。これに対して PbN 分類は、専門家の知識を必要とせずデータのみから観測されているデータである確率を求めることができる。人工データとベンチマークデータへの適用を通して、手法の有効性を検証する。

本論文の構成は、以下のとおりである。第2節は、PN, PU, PUbN, Pconf 分類の問題設定を確認する。第3節は、まず、PbN 分類の問題設定を提示し、次に歪んだ観測されているデータである事後確率を調整する方法を提示する。第4節では、PN, PbN 分類と調整された PbN 分類の性能比較を、人工データとベンチマークデータを用いて実験を行い、手法の有効性を検証する。第5節では結論を述べる。付録 A では、各分類手法のリスク表現と Pconf 分類、提案手法の理論解析に必要な補題や定理の導出を示す。

2 問題設定と既存手法

本節では、PN 分類と様々な弱教師あり分類の問題設定について説明する

2.1 PN 分類

いま、 $x \in \mathbb{R}^d$, $y \in \{+1, -1\}$ をそれぞれ d 次元特徴量ベクトル, およびラベルとし, これらは, 未知の確率密度関数 $p(x, y)$ に従うものとする. また, $p_P(x) = p(x|y = +1)$, $p_N(x) = p(x|y = -1)$ をそれぞれ正, 負のラベルを持つ特徴量の分布とし, $\pi = p(y = +1)$ をクラス事前確率とする. $g: \mathbb{R}^d \rightarrow \mathbb{R}$ を二値分類器とし, $\ell: \mathbb{R} \rightarrow \mathbb{R}_+$ を損失関数とする. 偽陰性率と偽陽性率をそれぞれ $R_P^+(g) = E_{x \sim p_P(x)}[\ell(g(x))]$, $R_N^-(g) = E_{x \sim p_N(x)}[\ell(-g(x))]$ とすると, PN 分類リスクは次のように表される.

$$R(g) = \pi R_P^+(g) + (1 - \pi) R_N^-(g). \quad (2.1)$$

この分類リスク (2.1) を最小化することで, g を決定する.

実際には, 損失関数に対する期待値が含まれているため, 分類リスク (2.1) の最小化を直接実行することはできない. そこで, $p_P(x)$ と $p_N(x)$ から独立して P データセット $\mathcal{X}_P = \{x_i^P\}_{i=1}^{n_P}$ と N データセット $\mathcal{X}_N = \{x_i^N\}_{i=1}^{n_N}$ をサンプリングし, これらを用いて, 分類リスク (2.1) を次のように近似する.

$$\hat{R}(g) = \pi \hat{R}_P^+(g) + (1 - \pi) \hat{R}_N^-(g).$$

ここで, $\hat{R}_P^+(g) = \frac{1}{n_P} \sum_{i=1}^{n_P} \ell(g(x_i^P))$, $\hat{R}_N^-(g) = \frac{1}{n_N} \sum_{i=1}^{n_N} \ell(-g(x_i^N))$ は, それぞれ偽陰性率と偽陽性率の経験損失である.

2.2 PU 分類

PU 分類 (du Plessis et al. 2014; du Plessis et al. 2015; Elkan & Noto 2008; Kiryo et al. 2017) とは, 正のデータとラベルのないデータから二値分類を行うものである. いま, $R_P^-(g) = E_{x \sim p_P(X)}[\ell(-g(x))]$, $R_U^-(g) = E_{x \sim p(x)}[\ell(-g(x))]$ とすると, $(1 - \pi) R_N^-(g) =$

$R_U^-(g) - \pi R_P^-(g)$ より, PU 分類のリスクは次のように表される.

$$R(g) = \pi R_P^+(g) - \pi R_P^-(g) + R_U^-(g).$$

$p(x)$ から独立して得られた, ラベルのないデータセットを $\mathcal{X}_U = \{x_i^U\}_{i=1}^{n_U}$ とする. このとき, PU 分類は, \mathcal{X}_N は利用できないが, $\mathcal{X}_P, \mathcal{X}_U$ より $R_N^-(g)$ を間接的に近似できるため, PU 分類リスクは次のように近似できる.

$$\hat{R}(g) = \pi \hat{R}_P^+(g) - \pi \hat{R}_P^-(g) + \hat{R}_U^-(g).$$

2.3 PUbN 分類

PUbN 分類は, P, U, bN データから二値分類するためのアプローチである. PUbN 分類に対する分類器を得るために, Hsieh et al. (2019) は, 観測された P, bN, U データに対して $+1$ 、その他のデータに対して -1 を返す潜在変数 s を導入した. いま, $\rho = p(y = -1, s = +1)$, $p_{bN}(x) = p(x|y = -1, s = +1)$, $p_{s=-1}(x) = p(x|s = -1)$ とする. さらに $R_{bN}^-(g) = E_{x \sim p_{bN}(x)}[\ell(-g(x))]$, $R_{s=-1}^-(g) = E_{x \sim p_{s=-1}(x)}[\ell(-g(x))]$ とすると, PUbN 分類リスクは次のように表される.

$$R(g) = \pi R_P^+(g) + \rho R_{bN}^-(g) + (1 - \pi - \rho) R_{s=-1}^-(g). \quad (2.2)$$

導出は付録 A.1 にある. bN データを $\mathcal{X}_{bN} = \{x_i^{bN}\}_{i=1}^{n_{bN}}$ とすると, (2.2) は, 次のように近似できる.

$$\hat{R}(g) = \pi \hat{R}_P^+(g) + \rho \hat{R}_{bN}^-(g) + (1 - \pi - \rho) \hat{R}_{s=-1}^-(g),$$

ここで, $\hat{R}_{bN}^-(g)$ と $\hat{R}_{s=-1}^-(g)$ は, それぞれ $R_{bN}^-(g)$ と $R_{s=-1}^-(g)$ の経験リスクである. Hsieh et al. (2019) は, P と bN データに対して U データがどれだけ重要であるかに着目し, U データにどれだけ重きを置くかを考慮した学習を行うため, (2.2) の $R_{s=-1}^-(g)$ を 3 項に分解している.

2.4 Pconf分類

Pconf分類 (Ishida et al., 2018) とは、負のデータもラベルのないデータもなく、正のデータのみから二値分類を試みるものである。 $R_P^-(g) = E_{x \sim p_P(X)}[\ell(-g(x))]$ とすると、Pconf分類リスクは次のように表される。

$$R(g) = \pi \left[R_P^+(g) + R_P^- \left(\frac{1-r}{r} g \right) \right]. \quad (2.3)$$

ここで、 $r = p(y = +1|x)$ は観測されたデータがPであることに対する信頼度で、 $r > 0$ と仮定している。この分類リスク (2.3) の導出は付録A.2にある。Pconf分類では、 $\mathcal{X}_P = \{x_i^P\}_{i=1}^{n_P}$ のみから分類器を学習することができ、分類リスクは次のように近似できる。

$$\hat{R}(g) = \pi \left[\hat{R}_P^+(g) + \hat{R}_P^- \left(\frac{1-r}{r} g \right) \right].$$

ここで、 $\hat{R}_P^-(g) = \frac{1}{n_P} \sum_{i=1}^{n_P} \ell(-g(x_i^P))$ である。

しかし、信頼度 r を求めるには、 $p(x)$ もしくは $p(x|y = -1)$ を直接用いるか、完全なPNデータからロジスティック回帰を分類器として利用することが必要で、実際の場面でこれらが既知である状況は考えにくい。そこで、信頼度 r は、ドメイン知識を借りるなどして概算する必要があるが、それによりバイアスが生じ信頼度に歪みが生じる可能性がある (Shinoda et al., 2020)。そこで、Shinoda et al. (2020) は、事前知識として偽陰性率を仮定し、偽陰性率と経験的偽陰性率の二乗誤差を最小化することで信頼度を修正した。

2.4.1 Pconf分類に対する理論解析

ここでは、Pconf分類の推定誤差境界を確認する。ただし、この小節の理論解析は、あくまでも信頼度に歪みが生じない場合に限ることに注意されたい。

$\sup_{g \in \mathcal{G}} \|g\|_\infty \leq C_g$ のような $C_g > 0$ と、 $\sup_{|z| \leq C_g} \ell(z) \leq C_\ell$ のような $C_\ell > 0$ が存在すると仮定する。Mohri et al. (2012) のように、 $\ell(z)$ はすべての $|z| \leq C_g$ に対してリプシッツ連続であり、リプシッツ定数 L_ℓ を持つと仮定する。 $g^* = \operatorname{argmin}_{g \in \mathcal{G}} R(g)$ を、真のリスクを最小化する関数、 $\hat{g} = \operatorname{argmin}_{g \in \mathcal{G}} \hat{R}(g)$ を経験リスクを最小化する関数とする。このとき、推定誤差 $R(\hat{g}) - R(g^*)$ を上から抑える問題を考える。 $r = p(y = +1|x) > 0$ と仮定している

ため、統計的学習理論に基づく誤差境界を導くためには、 r がゼロに近づきすぎることがないようにしなければならない。そのために、 $r \geq C_r$ となるような $C_r > 0$ が存在すると仮定する。

補題 1. 任意の $\delta > 0$ に対して、 $\hat{R}(g)$ を評価するためのデータの繰り返しサンプリングに対して、少なくとも $1 - \delta$ の確率で、次の一様偏差境界が成立する。

$$\sup_{g \in \mathcal{G}} |\hat{R}(g) - R(g)| \leq 2\pi \left(L_\ell + \frac{L_\ell}{C_r} \right) \mathfrak{R}_{n_P}(\mathcal{G}) + \pi \left(C_\ell + \frac{C_\ell}{C_r} \right) \sqrt{\frac{\ln 2/\delta}{2n_P}}.$$

ここで、 $\mathfrak{R}_{n_P}(\mathcal{G})$ は $p(x|y = +1)$ から得られたサンプルサイズ n_P の \mathcal{X}_P に対するラデマッハ複雑度である。

証明は付録 A.3 にある。補題 1 は、すべての $g \in \mathcal{G}$ に対して、高確率で $\hat{R}(g)$ が $R(g)$ の周りに集中し、その集中の度合いが $\mathfrak{R}_{n_P}(\mathcal{G})$ により制御されることを保証するものである。この補題に基づき、以下のように推定誤差の境界を設定することができる。

定理 1. 任意の $\delta > 0$ に対して、 \hat{g} を学習するためのデータの繰り返しサンプリングに対して、少なくとも $1 - \delta$ の確率で次式が成立する。

$$R(\hat{g}) - R(g^*) \leq 4\pi \left(L_\ell + \frac{L_\ell}{C_r} \right) \mathfrak{R}_{n_P}(\mathcal{G}) + 2\pi \left(C_\ell + \frac{C_\ell}{C_r} \right) \sqrt{\frac{\log 2/\delta}{2n_P}}.$$

証明は付録 A.4 にある。定理 1 は、 $n \rightarrow \infty$ のとき、常に $R(\hat{g}) \rightarrow R(g^*)$ となることを意味しており、 $\hat{R}(g)$ を用いた学習が一致性を持つことを保証している (Ledoux & Talagrand, 1991)。

また、次のように定義される線形パラメータモデルを考える

$$\mathcal{G} = \{g(x) = \langle w, \varphi(x) \rangle_{\mathcal{H}} \mid \|w\|_{\mathcal{H}} \leq C_w, \|\varphi(x)\|_{\mathcal{H}} \leq C_\varphi\}, \quad (2.4)$$

ここで、 \mathcal{H} はヒルベルト空間、 $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ は \mathcal{H} の内積、 $w \in \mathcal{H}$ はパラメータベクトル、 $\varphi : \mathbb{R}^d \rightarrow \mathcal{H}$ は特徴写像であり、 $C_w > 0$ と $C_\varphi > 0$ は定数である (Schölkopf & Smola, 2001)。 $\mathfrak{R}_{n_P}(\mathcal{G}) \leq C_w C_\varphi / \sqrt{n_P}$ であることが知られているため (Mohri et al, 2012)、 $O_p(1/\sqrt{n_P})$ で $R(\hat{g}) \rightarrow R(g^*)$ となる。さらに、 ℓ が狭義凸であれば $\hat{g} \rightarrow g^*$ となり、(2.4) の \mathcal{G} を用いれば $O_p(1/\sqrt{n_P})$ で $\hat{g} \rightarrow g^*$ となる (Boyd & Vandenberghe, 2004)。

3 提案手法

本節では、まず、PbN 分類のための分類リスクを提案する。次に、観測されているデータである事後確率の歪みに対する修正方法を提案する。

3.1 PbN 分類

PN 分類とは異なり、完全な負のデータではなく、一部の偏った負のデータである bN データしか観測されていないという状況で、二値分類を試みるものを PbN 学習と呼ぶ。この問題に対するアプローチのために、観測されている P と bN のデータであれば +1、未観測の N のデータであれば -1 を返す潜在変数 s を導入する。クラス事前確率 $\pi = p(y = +1)$ は本論文では既知であると仮定する。なお、Hsieh et al.(2019) でも同様の仮定が検討されており、Chen et al. (2020) や Kiryo et al. (2017) は、 π が既知であると仮定して PU 分類にアプローチしている。一方、du Plessis & Sugiyama (2014), du Plessis et al. (2016), Yao et al. (2021) は、PU 分類のために π を推定する方法を考えた。

いま、 $\rho = p(y = -1, s = +1)$, $p_{bN}(x) = p(x|y = -1, s = +1)$, $p_{s=+1}(x) = p(x|s = +1)$ とし、 $\sigma = p(s = +1|x)$ を観測されているデータである確率とする。 π は既知であると仮定しているので、 ρ と $p(s = +1)$ はそれぞれ次式で推定される。

$$\rho = \frac{n_{bN}}{n_P} \pi, \quad p(s = +1) = \frac{n_P + n_{bN}}{n_P} \pi.$$

$R_{bN}^-(g) = E_{x \sim p_{bN}(x)}[\ell(-g(x))]$, $R_{s=+1}^-(g) = E_{x \sim p_{s=+1}(x)}[\ell(-g(x))]$ とすると、 $\sigma \neq 0$ であれば、PbN 分類リスクは次のように表される。

$$R(g) = \pi R_P^+(g) + \rho R_{bN}^-(g) + (\pi + \rho) R_{s=+1}^- \left(\frac{1 - \sigma}{\sigma} g \right). \quad (3.1)$$

導出は、付録 A.5 にある。 $\mathcal{X}_{bN} = \{x_i^{bN}\}_{i=1}^{n_{bN}}$, $\mathcal{X}_{s=+1} = \{x_i^{s=+1}\}_{i=1}^{n_P + n_{bN}}$ をそれぞれ $p_{bN}(x)$, $p_{s=+1}(x)$ から独立して引き出されるサンプルとすると、分類リスク (3.1) は次のように近似できる。

$$\hat{R}(g) = \pi \hat{R}_P^+(g) + \rho \hat{R}_{bN}^-(g) + (\pi + \rho) \hat{R}_{s=+1}^- \left(\frac{1 - \sigma}{\sigma} g \right).$$

ここで $\hat{R}_{bN}^-(g) = \frac{1}{n_{bN}} \sum_{i=1}^{n_{bN}} \ell(-g(x_i^{bN}))$, $\hat{R}_{s=+1}^- = \frac{1}{n_p+n_{bN}} \sum_{i=1}^{n_p+n_{bN}} \ell(-g(x_i^{s=+1}))$ である。この場合、Pconf 分類のときと同様に、 σ は $p(x)$ に依存しているため陽に求められない。しかし、bN データを用いると、 $p_{\text{bias}}(x) = \pi p(x|y = +1) + (1 - \pi)p(x|y = -1, s = +1)$ として、擬似的な $p(x)$ なら求めることはできる。これを用いて、次のように歪んだ σ を直接求めることができる。

$$\tilde{\sigma} = \frac{p(s = +1)p(x|s = +1)}{p_{\text{bias}}(x)}. \quad (3.2)$$

Pconf 分類では、(2.3) の信頼度 r をドメイン知識を利用して推定する必要があった。一方、(3.1) の σ は、確率密度 $p(x|s = +1), p(x|y = +1), p(x|y = -1, s = +1)$ をカーネル密度推定などにより推定することにより、(3.2) のように推定することが可能である。ただし、PUbN 分類 (Hsieh et al., 2019) とは異なり、高次元データに対して密度推定による学習を提供することが困難な場合があることに注意されたい。PUbN 分類は、P データと bN データを P データとして扱い、PU 分類問題を解き、出力層にシグモイド関数を用いたディープニューラルネットワークを適用してデータが観測される確率を推定した。一方、提案手法では U データを想定していないため、適用が困難である。

3.2 Adjusted PbN 分類

(3.2) で与えられる確率 $\tilde{\sigma}$ は、bN データの偏りにより歪んでおり、予測精度に悪い影響を与えることが分かっている。本節では、 $\tilde{\sigma}$ の歪みを軽減するために、Pconf 分類における信頼度の歪みを軽減する方策 (Shinoda et al., 2020) を組み込むことを提案する。

いま、歪んだ $\tilde{\sigma}$ を修正するために、ハイパーパラメータ k ($0 < k < \infty$) を用いた、以下の分類リスクを考える。

$$\hat{R}_{\text{bias}}(g) = \pi \hat{R}_P^+(g) + \rho \hat{R}_{bN}^-(g) + (\pi + \rho) \hat{R}_{s=+1}^- \left(\frac{1 - \tilde{\sigma}^k}{\tilde{\sigma}^k} g \right). \quad (3.3)$$

ハイパーパラメータ k は、P データと N データを含む検証セットがあれば、クロスバリデーションによって選択できるが、ここでは、P データと bN データしかないので、 k を選択することは困難である。そこで、この問題を解決するために、事前知識として次の偽陰性率

が既知であると仮定する.

$$\phi = \int_{\{x:g(x)<0\}} p(x|y=+1)dx.$$

この仮定の下では, 以下の二乗誤差を最小化することで, 既知である偽陰性率と, 経験的偽陰性率の差が最も小さくなる, 最適なハイパーパラメータ k^* を選択することができる.

$$k^* = \operatorname{argmin}_k \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \ell_{01}(g(x_i)) - \phi \right)^2.$$

ここで, $\ell_{01}(z) = (1 - \operatorname{sign}(z))/2$ は 0-1 損失である. 選択された k^* を用いて, 調整された PbN 分類の代替的なリスクは, 次のように表される (Watanabe & Matsui, 2023).

$$\hat{R}_{\text{bias}}(g) = \pi \hat{R}_P^+(g) + \rho \hat{R}_{bN}^-(g) + (\pi + \rho) \hat{R}_{s=+1}^- \left(\frac{1 - \tilde{\sigma}^{k^*}}{\tilde{\sigma}^{k^*}} g \right).$$

実際には, 偽陰性率 ϕ が既知であるという仮定は現実的ではない. しかし, 4.1.3 節の数値実験で示すように, 提案手法の分類精度は ϕ の選択に敏感でない.

3.3 PbN 分類に対する理論解析

ここでは, Pconf 分類のときと同様に, PbN 分類の推定誤差境界を確認する. ただし, この小節の理論解析は, あくまでも観測されているデータである確率 $\sigma = p(s=+1)$ に歪みが生じない場合に限ることに注意されたい.

いま, $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq C_g$ のような $C_g > 0$ と, $\sup_{|z| \leq C_g} \ell(z) \leq C_\ell$ のような $C_\ell > 0$ が存在すると仮定する. Mohri et al. (2012) と同様に, $\ell(z)$ はすべての $|z| \leq C_g$ に対してリプシッツ連続であり, リプシッツ定数 L_ℓ を持つと仮定する. $g^* = \operatorname{argmin}_{g \in \mathcal{G}} R(g)$ を真のリスクを最小化する関数, $\hat{g} = \operatorname{argmin}_{g \in \mathcal{G}} \hat{R}(g)$ を経験リスクを最小化する関数とする. 推定誤差 $R(\hat{g}) - R(g^*)$ を上から抑える問題を考える. $\sigma = p(s=+1|x) > 0$ と仮定しているため, 統計的学習理論に基づく誤差境界を導くためには, σ がゼロに近づきすぎることがないようにしなければならない. そのために, $\sigma \geq C_\sigma$ となるような $C_\sigma > 0$ が存在すると仮定する.

補題 2. 任意の $\delta > 0$ に対して, $\hat{R}(g)$ を評価するためのデータの繰り返しサンプリングに対して, 次の一様偏差境界が, 少なくとも $1 - \delta$ の確率で成立する.

$$\begin{aligned} \sup_{g \in \mathcal{G}} |\hat{R}(g) - R(g)| &\leq 2\pi L_\ell \mathfrak{R}_{n_P}(\mathcal{G}) + 2\rho L_\ell \mathfrak{R}_{n_{bN}} + 2(\pi + \rho)(L_\ell/C_\sigma) \mathfrak{R}_{s=+1}(\mathcal{G}) \\ &\quad + \pi C_\ell \sqrt{\frac{\log 6/\delta}{2n_P}} + \rho C_\ell \sqrt{\frac{\log 6/\delta}{2n_{bN}}} + (\pi + \rho)(C_\ell/C_\sigma) \sqrt{\frac{\log 6/\delta}{2(n_P + n_{bN})}}. \end{aligned}$$

ここで, $\mathfrak{R}_{n_P}(\mathcal{G})$, $\mathfrak{R}_{n_{bN}}(\mathcal{G})$, $\mathfrak{R}_{s=+1}(\mathcal{G})$ は, それぞれ \mathcal{X}_P , \mathcal{X}_{bN} , $\mathcal{X}_{s=+1}$ に対するラデマッハ複雑度である.

証明は付録 A.6 にある. 補題 2 は, すべての $g \in \mathcal{G}$ に対して, 高確率で $\hat{R}(g)$ が $R(g)$ の周りに集中し, その集中の度合いが $\mathfrak{R}_{n_P}(\mathcal{G})$, $\mathfrak{R}_{n_{bN}}(\mathcal{G})$, $\mathfrak{R}_{s=+1}(\mathcal{G})$ により制御されることを保証するものである. この補題に基づき, 以下のように推定誤差の境界を導出することができる.

定理 2. 任意の $\delta > 0$ に対して, \hat{g} を学習するためのデータの繰り返しサンプリングに対して, 少なくとも $1 - \delta$ の確率で次式が成立する.

$$\begin{aligned} R(\hat{g}) - R(g^*) &\leq 4\pi L_\ell \mathfrak{R}_{n_P}(\mathcal{G}) + 4\rho L_\ell \mathfrak{R}_{n_{bN}}(\mathcal{G}) + 4(\pi + \rho)(L_\ell/C_\sigma) \mathfrak{R}_{s=+1}(\mathcal{G}) \\ &\quad + 2\pi C_\ell \sqrt{\frac{\log 6/\delta}{2n_P}} + 2\rho C_\ell \sqrt{\frac{\log 6/\delta}{2n_{bN}}} + 2(\pi + \rho)(C_\ell/C_\sigma) \sqrt{\frac{\log 6/\delta}{2(n_P + n_{bN})}}. \end{aligned}$$

証明は付録 A.7 にある. 定理 2 は, $n \rightarrow \infty$ のとき, 常に $R(\hat{g}) \rightarrow R(g^*)$ となることを意味しており, $\hat{R}(g)$ を用いた学習が一致性を持つことを保証している (Ledoux & Talagrand, 1991).

また, 次のように定義される線形パラメータモデルを考える

$$\mathcal{G} = \{g(x) = \langle w, \varphi(x) \rangle_{\mathcal{H}} \mid \|w\|_{\mathcal{H}} \leq C_w, \|\varphi(x)\|_{\mathcal{H}} \leq C_\varphi\}, \quad (3.4)$$

ここで, \mathcal{H} はヒルベルト空間, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ は \mathcal{H} の内積, $w \in \mathcal{H}$ はパラメータベクトル, $\varphi : \mathbb{R}^d \rightarrow \mathcal{H}$ は特徴写像であり, $C_w > 0$ と $C_\varphi > 0$ は定数である (Schölkopf & Smola, 2001). このとき, $\mathfrak{R}_{n_P}(\mathcal{G}) \leq C_w C_\varphi / \sqrt{n_P}$ であることが知られており (Mohri et al, 2012), 同様に, $\mathfrak{R}_{n_{bN}}(\mathcal{G}) \leq C_w C_\varphi / \sqrt{n_{bN}}$, $\mathfrak{R}_{s=+1}(\mathcal{G}) \leq C_w C_\varphi / \sqrt{n_P + n_{bN}}$ となる. したがって, $O_p(1/\sqrt{n_P} +$

$1/\sqrt{n_{bN}} + 1/\sqrt{n_P + n_{bN}}$ で $R(\hat{g}) \rightarrow R(g^*)$ となる. さらに, ℓ が狭義凸であれば $\hat{g} \rightarrow g^*$ となり, (3.4) の \mathcal{G} を用いれば $O_p(1/\sqrt{n_P} + 1/\sqrt{n_{bN}} + 1/\sqrt{n_P + n_{bN}})$ で $\hat{g} \rightarrow g^*$ となる (Boyd & Vandenberghe, 2004).

4 数値実験

本節では, 提案手法の性能を人工データおよびベンチマークデータを用いての実験で検証する.

4.1 人工データ実験 1

4.1.1 設定

今回の実験では, 独立して抽出した P と N のデータのサンプルから, トレーニングデータセット, 検証データセット, テストデータセット, 偽陰性率推定用データセットを生成する. トレーニングデータセットには 500 個の P のサンプルと 100 個の bN のサンプル, 検証データセットには 500 個の P のサンプル, テストデータセットと偽陰性率推定用データセットには 500 個の P のサンプル及び 500 個の N のサンプルが含まれているとした. 調整された PbN 分類による正解率を, ナイーブ PbN 分類, 通常の PN 分類と比較した. ここで, ナイーブ PbN 分類とは, ハイパーパラメータ k を用いて σ の歪みを軽減していない PbN 分類である. ナイーブ PbN 分類と PN 分類では, トレーニングデータセットと検証データセットを合わせてトレーニングデータセットとした.

全ての手法に, 二値分類器として線形モデル $g(x) = a^T x + \beta$ を, 損失関数としてロジスティック損失 $\ell_L(z) = \log(1 + e^{-z})$ を用い, 最適化には, 確率的勾配降下法を適用することで分類器を構築した. ハイパーパラメータ k の候補は, $k \in \{0.3, 0.5, 0.7, 1.0, 1.5, 2.0, 4.0\}$ とした. $p(x|s = +1)$, $p(x|y = +1)$, $p(x|y = -1, s = +1)$ がガウス分布に従うものとし, それらの平均ベクトルと分散共分散行列をそれぞれ標本平均ベクトルと標本分散共分散行列で推定した. 推定された $p(x|s = +1)$, $p(x|y = +1)$, $p(x|y = -1, s = +1)$ を用いて σ は (3.2) のように算出される. また, 最適化安定のため, 0.01 未満の $\bar{\sigma}$ は 0.01 に切り上げた.

P データは, 平均 $[0, 0]^T$, 分散共分散行列が単位行列のガウス分布から発生させた. 一方

で, N データは, 平均ベクトルが互いに異なり, 分散共分散行列が全て単位行列である 4 つの混合ガウス分布から発生されているとした. 具体的には, N データの分布は, $[1.0, 1.0]^T$, $[1.5, 1.5]^T$, $[2.0, 2.0]^T$, $[2.5, 2.5]^T$ の 4 つを平均ベクトルとした, P データと N データとの重複度が大きいグループに分布するパターンと, $[2.0, 2.0]^T$, $[3.0, 3.0]^T$, $[4.0, 4.0]^T$, $[5.0, 5.0]^T$ の 4 つを平均ベクトルとした重複度が小さいグループに分布するパターンの 2 通りを考えた. これにより, クラスの重複度による精度の違いを検証する. また, N データのうち, 実際にどのデータが偏って観測されているかによって, 2 つのケースを検討した. 具体的には, N データを構成する 4 つの成分のうち 1 つを bN データとする場合と, N データを構成する 4 つの成分それぞれから, 確率 $[0.25, 0.25, 0.25, 0.25]/[0.40, 0.10, 0.35, 0.15]/[0.15, 0.40, 0.10, 0.35]/[0.35, 0.15, 0.40, 0.10]$ で部分的に得られたものを bN データとする場合を考えた.

以上をまとめて, 次の 4 つの状況下で数値実験を行う.

- **状況 1.** クラスの重複度が大きく, N データを構成する 4 つの成分のうち 1 つを bN データとする場合
- **状況 2.** クラスの重複度が小さく, N データを構成する 4 つの成分のうち 1 つを bN データとする場合
- **状況 3.** クラスの重複度が大きく, N データを構成する 4 つの成分それぞれから部分的に得られたものを bN データとする場合
- **状況 4.** クラスの重複度が小さく, N データを構成する 4 つの成分それぞれから部分的に得られたものを bN データとする場合

これらの設定それぞれにおいて, ナイーブ PbN 分類, 調整された PbN 分類, PN 分類分類による正解率を算出した.

4.1.2 4 つの状況下における実験

状況 1 と状況 2 の 10 回の試行における正解率の平均値と標準偏差を, 表 1 に示す. μ_{bn} は bN データの平均であり, $\hat{\phi}$ は推定された偽陰性率の, 10 回の試行における平均値と標準偏差である. A.PbN は調整された PbN 分類を, N.PbN はナイーブ PbN 分類を表す. 調整

表 1: 人工データ実験 1 における, 状況 1, 2 の 10 回の試行による正解率の平均値と標準偏差. 有意水準 5% の t 検定に基づく最良の手法と同等の手法を太字で示した.

状況 1				
μ_{bn}	A.PbN	N.PbN	PN	$\hat{\phi}$
$[1.0, 1.0]^{\top}$	85.31 ± 0.82	83.39 ± 1.15	83.04 ± 0.97	10.50 ± 1.20
$[1.5, 1.5]^{\top}$	86.09 ± 0.54	85.91 ± 0.79	83.35 ± 1.15	
$[2.0, 2.0]^{\top}$	85.70 ± 0.65	86.01 ± 1.27	82.19 ± 0.79	
$[2.5, 2.5]^{\top}$	85.73 ± 1.02	85.78 ± 0.79	79.26 ± 1.10	
状況 2				
μ_{bn}	A.PbN	N.PbN	PN	$\hat{\phi}$
$[2.0, 2.0]^{\top}$	95.81 ± 0.57	93.72 ± 0.38	96.14 ± 0.40	2.55 ± 0.71
$[3.0, 3.0]^{\top}$	96.10 ± 0.65	92.18 ± 1.10	94.40 ± 0.77	
$[4.0, 4.0]^{\top}$	95.88 ± 0.74	84.20 ± 2.19	91.26 ± 1.07	
$[5.0, 5.0]^{\top}$	95.37 ± 1.02	76.86 ± 1.74	87.10 ± 0.75	

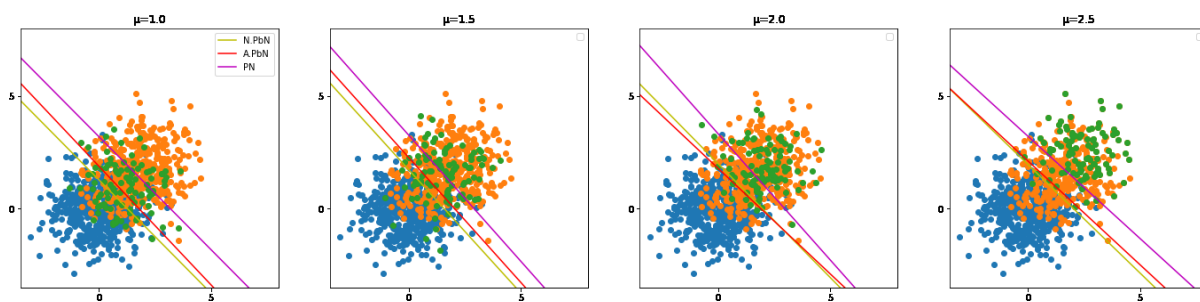
された PbN 分類は, クラスの重複度が大きい場合は, PN 分類より有意に優れており, クラスの重複度が小さい場合は, ナイーブ PbN 分類より有意に優れている.

次に, 状況 3 と状況 4 での, 10 回の試行における正解率の平均値と標準偏差を, 表 2 に示す. 調整された PbN 分類は, クラスの重複度が大きい場合は, PN 分類より有意に優れており, クラスの重複度が小さい場合は, ナイーブ PbN 分類より有意に優れている. 以上のことにより, 提案手法は, クラスの重複度, bN データの形状に関係なく精度が安定していることがわかる.

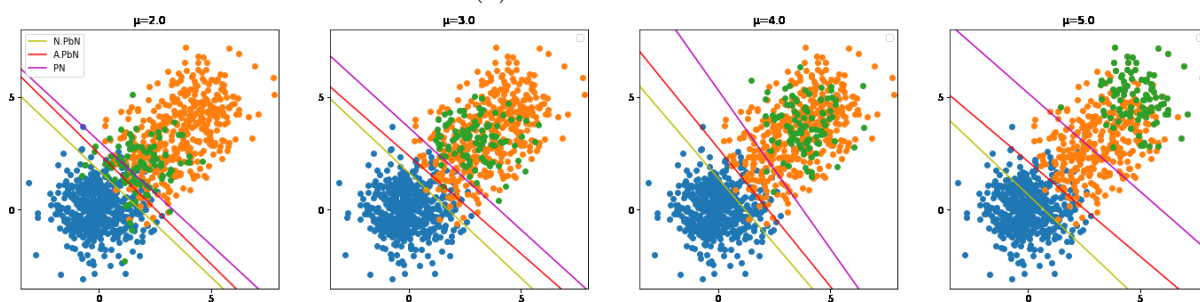
また, 状況 1~4 における, 分類境界を図 1(a)-(d) に示す. 青色が P データ, 緑色が bN データ, オレンジ色が未観測の N データを表す. 偏った PbN 分類, 調整された PbN 分類, PN 分類の分類境界は, それぞれ, 黄色, 赤色, 紫色で示されている. 図より, 提案手法に比べて, ナイーブ PbN 分類は P データの影響を強く受けており, PN 分類は bN データの影響を強く受けていることがわかる.

4.1.3 ϕ の推定にずれが生じたときの影響

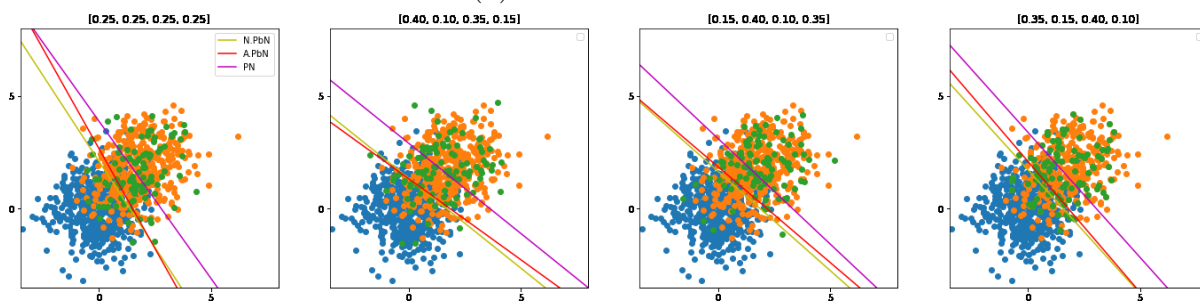
4.1.2 項の実験では, 偽陰性率 ϕ を正確に推定できると仮定したが, 実際には, その仮定は必ずしも成り立つとはいえない. そこで, ϕ の推定値 $\hat{\phi}$ にずれを生じさせた状態で実験を行い, ϕ の推定誤差に対して, PbN 分類の精度がどの程度影響を受けるかを検証した.



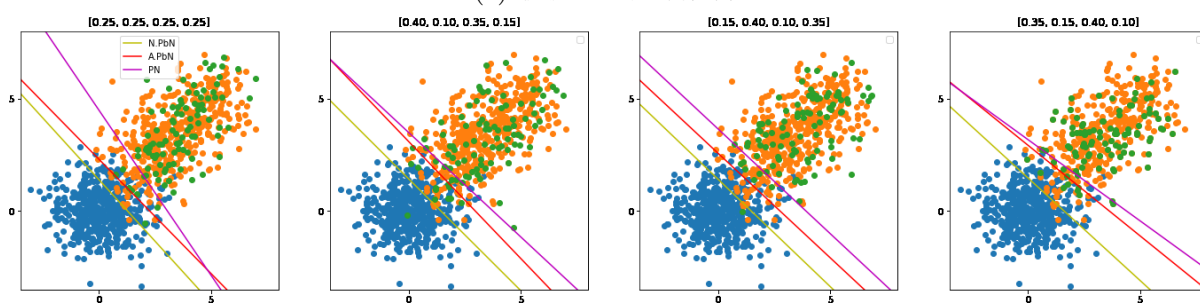
(a) 状況 1 の分類境界



(b) 状況 2 の分類境界



(c) 状況 3 の分類境界



(d) 状況 4 の分類境界

図 1: 4つの状況下の分類境界

表 2: 人工データ実験 1 における, 状況 3, 4 の 10 回の試行による正解率の平均値と標準偏差. 有意水準 5% の t 検定に基づく最良の手法と同等の手法を太字で示した.

状況 3				
bN	A.PbN	N.PbN	PN	$\hat{\phi}$
[0.25, 0.25, 0.25, 0.25]	86.18 \pm 1.05	85.92 \pm 0.85	82.50 \pm 1.01	10.58 \pm 1.34
[0.40, 0.10, 0.35, 0.15]	86.08 \pm 0.99	86.08 \pm 1.11	82.79 \pm 1.35	
[0.15, 0.40, 0.10, 0.35]	86.64 \pm 1.05	86.24 \pm 1.30	82.49 \pm 1.21	
[0.35, 0.15, 0.40, 0.10]	86.04 \pm 1.03	85.94 \pm 1.11	83.26 \pm 1.50	
状況 4				
bN	A.PbN	N.PbN	PN	$\hat{\phi}$
[0.25, 0.25, 0.25, 0.25]	96.21 \pm 0.79	86.93 \pm 1.45	95.19 \pm 0.44	2.52 \pm 0.78
[0.40, 0.10, 0.35, 0.15]	96.38 \pm 0.38	88.80 \pm 1.12	95.74 \pm 0.55	
[0.15, 0.40, 0.10, 0.35]	95.46 \pm 0.55	86.89 \pm 1.96	94.13 \pm 0.72	
[0.35, 0.15, 0.40, 0.10]	95.82 \pm 0.65	89.66 \pm 1.83	95.06 \pm 0.52	

ここでは, 4.1.2 節の状況 1 と状況 2 と同じ設定で, 正確に推定できた $\hat{\phi}$ の値に $c \in \{0.5, 0.7, 1.3, 1.5\}$ を乗じることで, 推定にずれを生じさせた状態で, 調整された PbN 分類を適用した. 正解率の 10 回の平均値と標準偏差を, 表 3 に示す. クラスの重複度が大きい場合, $c = 0.5, 0.7$ のときは, 推定誤差の影響が大きかったが, クラスの重複度が小さい場合は, $\hat{\phi}$ を過小評価していなければ, 推定誤差の影響はほとんどなかった. この結果から, 提案手法は, 推定誤差を大きめに評価するかクラスの重なり度合いが小さい場合であれば, ϕ を正確に推定できない実用的な状況であっても有効であることがわかる.

4.2 人工データ実験 2

4.1 節では, 提案手法の有効性を数値的, 視覚的に検証するために 2 次元のデータについて実験を行った. ここでは, 高次元に拡張したときの PbN 分類の精度について検証する.

ここでは, P データは平均ベクトル $[0, 0, \dots, 0]^T$, 分散共分散行列が単位行列の多次元ガウス分布から生成させた. 一方で, N データは, 平均ベクトルが互いに異なり, 分散共分散行列が全て単位行列である 4 つの多次元混合ガウス分布から生成されているとした. 具体

表 3: 人工データ実験 1 における, 状況 1, 2 の推定誤差のある $\hat{\phi}$ を用いた, 調整された PbN 分類の 10 回の試行における正解率の平均値と標準偏差. 有意水準 5% の t 検定に基づき, $c = 1.0$ のときと同等の結果を太字で示した. μ_{bn} は bN データの平均である.

状況 1				
μ_{bn}	$c = 0.5$	$c = 0.7$	$c = 1.3$	$c = 1.5$
[1.0, 1.0]	79.23 \pm 2.65	83.24 \pm 2.19	85.53 \pm 0.89	85.66 \pm 0.78
[1.5, 1.5]	81.44 \pm 1.67	83.68 \pm 1.29	85.21 \pm 0.52	85.14 \pm 0.47
[2.0, 2.0]	81.02 \pm 2.32	84.04 \pm 1.57	85.80 \pm 1.74	85.00 \pm 1.82
[2.5, 2.5]	79.74 \pm 2.76	83.14 \pm 2.27	84.27 \pm 1.53	82.74 \pm 1.69
状況 2				
μ_{bn}	$c = 0.5$	$c = 0.7$	$c = 1.3$	$c = 1.5$
[2.0, 2.0]	89.67 \pm 1.79	94.32 \pm 1.04	96.24 \pm 0.53	96.24 \pm 0.53
[3.0, 3.0]	90.61 \pm 1.38	95.14 \pm 1.48	95.76 \pm 0.45	95.76 \pm 0.45
[4.0, 4.0]	91.60 \pm 3.95	94.74 \pm 1.60	95.44 \pm 0.64	95.30 \pm 0.77
[5.0, 5.0]	89.91 \pm 2.71	95.67 \pm 0.70	95.44 \pm 0.48	95.44 \pm 0.48

的には, N データの分布は, 次の 4 つを平均ベクトルとした.

$$\begin{aligned}\mu_{\text{bN}}^1 &= (3/\sqrt{d})\mathbf{1}_d, \quad \mu_{\text{bN}}^2 = \mu_{\text{bN}}^1 + u_1, \\ \mu_{\text{bN}}^3 &= \mu_{\text{bN}}^2 + u_2, \quad \mu_{\text{bN}}^4 = \mu_{\text{bN}}^3 + u_3,\end{aligned}$$

ここで, $\mathbf{1}_d$ は全ての要素が 1 である d 次元ベクトル, $u_k = (u_{k1}, \dots, u_{kd})^T$ ($k = 1, 2, 3$) は要素が一様分布 $U(-10/\sqrt{d}, 10/\sqrt{d})$ に従う d 次元ベクトルである. 高次元のとき, ガウス分布の指数部分が発散してしまうため, σ の推定に必要な $p(x|s = +1)$, $p(x|y = +1)$, $p(x|y = -1, s = +1)$ の推定には, バンド幅が 0.1 のガウスカーネルを用いたカーネル密度推定法を利用した. 本シミュレーションでは, 4.1 節の状況 1, 2 と同様に, N 個のデータの 4 つのガウス成分のうち 1 つが観測されたものとし, N 個のデータに対して, 観測された成分を置き換えるシミュレーションを行った. モデル $g(x)$, 損失関数 $\ell(z)$ など, その他の設定は 4.1 節と同じものとした.

10 次元, 50 次元, 100 次元での 10 回の試行における平均と標準偏差を表 4 に示す. 提案手法は, bN データが P データから離れている場合, 他の手法より有意に優れており, bN データが P データと近い場合でも, 十分な精度を得ることができる. この結果から, 提案手法は, 高次元の場合においても精度が安定していることがわかる.

表 4: 人工データ実験 2 における 10 回の試行による正解率の平均値と標準偏差. 有意水準 5% の t 検定に基づく最良の手法と同等の手法を太字で示した.

$d = 10$				
μ_{bn}	A.PbN	N.PbN	PN	$\hat{\phi}$
μ_{bn}^1	95.64 ± 0.90	95.22 ± 0.88	95.67 ± 1.00	2.58 ± 0.98
μ_{bn}^2	90.86 ± 1.08	91.10 ± 0.69	89.83 ± 0.58	
μ_{bn}^3	87.30 ± 2.61	86.78 ± 0.95	84.92 ± 1.10	
μ_{bn}^4	80.80 ± 1.26	81.23 ± 1.14	79.70 ± 0.81	
$d = 50$				
μ_{bn}	A.PbN	N.PbN	PN	$\hat{\phi}$
μ_{bn}^1	89.43 ± 2.68	89.10 ± 3.46	89.40 ± 2.79	3.61 ± 0.97
μ_{bn}^2	89.73 ± 0.96	89.24 ± 0.26	88.98 ± 0.13	
μ_{bn}^3	88.48 ± 0.81	88.17 ± 0.47	87.20 ± 0.83	
μ_{bn}^4	87.39 ± 2.05	85.76 ± 1.19	84.00 ± 1.34	
$d = 100$				
μ_{bn}	A.PbN	N.PbN	PN	$\hat{\phi}$
μ_{bn}^1	88.69 ± 3.65	88.90 ± 3.23	88.74 ± 3.30	6.25 ± 1.33
μ_{bn}^2	88.52 ± 0.82	88.71 ± 0.93	87.79 ± 1.33	
μ_{bn}^3	87.54 ± 1.24	86.83 ± 1.05	85.97 ± 1.19	
μ_{bn}^4	84.37 ± 3.28	79.33 ± 1.30	77.63 ± 1.06	

4.3 ベンチマークデータ実験

4.3.1 無線による屋内測定データセット

UCI Machine Learning Repository から引用した, Wireless Indoor Localization Data Set (Bhatt, 2017) を使用した. このデータは, 無線 LAN の信号強度を利用して屋内の位置を特定する実験のために収集されたもので, 1~4 で番号付けされた 4 つの部屋において, それぞれ 7 次元の特徴量ベクトルからなるサンプルサイズ 500 のデータを計測している. ここでは, 2 番の部屋を正のクラス, それ以外の部屋を負のクラスとし, 二値分類用のデータセットを作成した. そして, このデータセットを, トレーニングデータセット, 検証データセット, テストデータセット, 偽陰性率推定用データセットに分けた. トレーニングデータセットには, 200 個の P のサンプルと 100 個の bN のサンプル, 検証データセットには 100 個の P のサンプル, テストデータセットと偽陰性率推定用データセットには, 100 個の P のサンプル及び 300 個の N のサンプルが含まれているとした.

表 5: Wireless Indoor Localization Data Set で, bN データを, 1, 3, 4 番の各部屋からのみ得られた場合と, 各部屋からランダムに得られた場合としたときの, 100 回の試行による正解率と F スコアの平均値と標準偏差. 有意水準 5% の t 検定に基づく最良の手法と同等の手法を太字で示した.

μ_{bn}	A.PbN	N.PbN	PN	$\hat{\phi}$
Accuracy				
Room 1.	86.28 \pm 4.76	83.76 \pm 3.42	82.45 \pm 3.09	3.34 \pm 2.17
Room 3.	97.87 \pm 1.72	96.60 \pm 4.99	94.61 \pm 5.61	
Room 4.	89.96 \pm 6.39	87.83 \pm 6.76	77.18 \pm 8.23	
Random	98.05 \pm 0.87	97.97 \pm 0.98	96.25 \pm 3.32	
F-score ($\times 10^2$)				
Room 1.	78.88 \pm 5.98	75.67 \pm 3.85	74.16 \pm 3.33	
Room 3.	95.85 \pm 2.61	90.97 \pm 18.60	86.40 \pm 17.84	
Room 4.	84.05 \pm 7.62	81.22 \pm 7.59	69.58 \pm 8.20	
Random	96.13 \pm 1.70	95.98 \pm 1.90	93.21 \pm 4.92	

調整された PbN 分類による正解率と F スコアを, ナイーブ PbN 分類, 通常の PN 分類と比較した. 全ての手法に, 二値分類器としては線形モデル $g(x) = a^T x + \beta$ を, 損失関数としてはロジスティック損失 $\ell_L(z) = \log(1 + e^{-z})$ を用い, 最適化には, 確率的勾配降下法を適用することで学習器を構築した. 観測されているデータである確率 σ の推定に必要な, $p(x|s = +1)$, $p(x|y = +1)$, $p(x|y = -1, s = +1)$ の推定には, ガウスカーネルおよびバンド幅 0.1 によるカーネル密度推定を用いた. (3.3) のハイパーパラメータ k の候補は, $k \in \{0.5, 0.7, 0.9, 1.0, 1.2, 1.5, 2.0\}$ とし, 最適化安定のため, 0.01 未満の $\hat{\sigma}$ は 0.01 に切り上げた.

bN データは, 1, 3, 4 番の各部屋からのみ得られた場合と, 各部屋からランダムに得られた場合の 2 パターンで分類を行った. 100 回の試行における正解率と F スコアの平均値と標準偏差を表 5 に示す. $\hat{\phi}$ は推定された偽陰性率の, 100 回の試行における平均値と標準偏差である. 提案手法は, 全ての場合で, 既存手法と同等か, より好ましい性能を示している.

4.3.2 MNIST

MNIST のデータセット (LeCun & Cortes, 1998) は, 0 から 9 までの手書き数字からなる 70000 枚のグレースケール画像で構成される. 各画像のピクセルサイズは 28×28 , すなわち 784 次元であり, 0 から 9 のラベルと関連付けられている. この実験では, 偶数を正のクラス, 奇数を負のクラスとして扱い, 二値分類を行った. そして, データセットをトレーニング, 検証, テストデータセットと偽陰性率を推定するためのデータセットに分けた. サンプルサイズが大きいと提案手法の計算が困難になるため, データセットの全サンプルの 10 分の 1 を使用した. そして, 2000 個の P のサンプルと 400 個の bN のサンプルをトレーニングデータセット, 500 個の P のサンプルを検証データセット, 500 個の P のサンプルを偽陰性率推定のためのデータセットとする. さらに, 1, 3, 5, 7, 9 の数字からそれぞれ部分的に得られたデータを bN データとして扱った.

調整された PbN 分類による正解率を, 偏った PbN 分類, 通常の PN 分類と比較した. 全ての手法に, 二値分類器としては隠れ層 3 層の完全連結型ニューラルネットワーク (784-300-300-300-1) を, 損失関数としてはロジスティック損失 $\ell_L(z) = \log(1 + e^{-z})$ を用い, 最適化には, Adaptive moment (ADAM) を適用することで学習器を構築した. さらに, Backurs et al.(2019) が提案したカーネル密度推定法を適用し, 観測されているデータである確率 σ の推定に必要な, $p(x|s = +1)$, $p(x|y = +1)$, $p(x|y = -1, s = +1)$ の推定を行った. (3.3) のハイパーパラメータ k の候補は, $k \in \{0.5, 0.7, 0.9, 1.0, 1.2, 1.5, 2.0\}$ とし, 最適化安定のため, 0.01 未満の $\hat{\sigma}$ は, 0.01 に切り上げた.

10 回の試行における正解率の平均値と標準偏差を表 6 に示す. $\hat{\phi}$ は推定された偽陰性率の, 10 回の試行における平均値と標準偏差である. 提案手法は, 全ての場合で, 既存手法と同等か, より好ましい性能を示している.

5 結論

本論文では, 正のデータと, 偏った一部の負のデータのみが観測された状態での分類問題である PbN 問題に対する新しいアプローチを提案した. 提案手法では, 各インスタンス

表 6: MNIST で, bN データを, 1, 3, 5, 7, 9 それぞれから部分的に得られた場合の, 10 回の試行による正解率の平均値と標準偏差. 有意水準 5% の t 検定に基づく最良の手法と同等の手法を太字で示した.

bN データの構成率	A.PbN	N.PbN	PN	$\hat{\phi}$
[0.40, 0.30, 0.20, 0.05, 0.05]	76.78 \pm 3.77	69.53 \pm 5.77	75.25 \pm 2.92	6.47 \pm 1.38
[0.05, 0.40, 0.30, 0.20, 0.05]	77.41 \pm 4.24	71.08 \pm 3.50	74.71 \pm 2.64	
[0.05, 0.05, 0.40, 0.30, 0.20]	77.51 \pm 2.80	72.55 \pm 7.10	73.62 \pm 2.99	
[0.20, 0.05, 0.05, 0.40, 0.30]	77.30 \pm 4.13	69.90 \pm 6.40	75.14 \pm 2.47	
[0.30, 0.20, 0.05, 0.05, 0.40]	77.20 \pm 5.25	72.59 \pm 1.93	75.71 \pm 3.25	

が実際に観測されたかどうかを示す確率の傾きに歪みが生じるために, リスクを適切に推定できない. そこで, 偽陰性率と経験的偽陰性率の二乗誤差を最小化することで, 歪みを軽減した. その結果, 人工データセットとベンチマークデータセットの解析により, 既存の分類手法よりも優れていることが確認できた. また, 特定の状況下では, 偽陰性率の推定にずれがあったとしても分類精度への影響が小さいことが分かった.

提案手法は, U データを用いず, P データと bN データのみで分類を行うことができる. しかし, 4.1 節で述べたように, 特に高次元データでは U データがない場合, 確率 σ の推定が困難である. 5.3.2 節では, 高次元に適用可能なカーネル密度推定を適用したが, 不安定な結果が得られることがある. 今後の課題として, 高次元のデータに対して安定した推定方法を検討する. また, 5.3.2 節では, 計算量が多いため, 全サンプルの一部のみを用いて分析を行った. 計算負荷の改善も今後の課題の一つである. また, 偽陰性率が既知であるという仮定を必要としないハイパーパラメータ選択の別の方法を模索する必要がある.

謝辞

本研究の遂行にあたり, 指導教員として終始多大なご指導を賜った, 滋賀大学データサイエンス学部データサイエンス学科准教授松井秀俊先生に深謝致します. 同学科教授椎名洋先生, 並びに同学科准教授岩山幸治先生には, 本論文の作成にあたり, 副査として適切なお助言を賜りました. ここに深謝の意を表します. 最後に, 松井研究室の皆様に, 本研究の遂行にあたり多大なご助言, ご協力頂きました. ここに感謝の意を表します.

参考文献

- Bache, K. and Lichman, M. UCI Machine Learning Repository. University of California, School of Information and Computer Scienc, 2013. <http://archive.ics.uci.edu/ml>
- Backurs, A., Indyk, P., and Wagner, T. Space and Time Efficient Kernel Density Estimation in High Dimensions. In *NeurIPS*, pp. 15773—15782, 2019.
- Bhatt, R. *Wireless Indoor Localization*. UCI Machine Learning Repository, 2017.
- Boyd, S. and Vandenberghe., L. *Convex Optimization..* Cambridge University Press, 2004.
- Chapelle, O., Schölkopf, B., and Zien, A., editors. *Semi-Supervised Learning*. MIT Press, 2006.
- Chen, X., Chen, W., Chen, T., Yuan, Y., Gong, C., and Chen, K. Self-pu: Self boosted and calibrated positive-unlabeled training. In *ICML*, 2020.
- du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *NeurIPS*, pp. 703—711, 2014.
- du Plessis, M., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386—1394, 2015.
- du Plessis, M. C., Niu, G., and Sugiyama, M. (2016). Class-prior estimation for learning from positive and unlabeled data. In *ACML*, 221–236.
- du Plessis, M. C. and Sugiyama, M. (2014). Class prior estimation from positive and unlabeled data. In *IEICE Trans. Inform. Sys.* **97**(5), 1358–1362.
- Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.
- Fei, G. and Liu, B. Social media text classification under negative covariate shift. In Proc. of *EMNLP*, pp. 2347—2356, 2015.

- Hsieh, Y.G., Niu, G., and Sugiyama, M. Classification from positive, unlabeled and biased negative data, In *ICML, PMLR*, pp. 2820–2829, 2019.
- Ishida, T., Niu, G., and Sugiyama, M. Binary classification from positive-confidence data. In *NeurIPS 2018*, pp. 5921–5932, 2018.
- Kato, M., Teshima, T., and Honda, J. Learning from positive and unlabeled data with a selection bias. In *ICLR*, 2019
- Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, pp. 1675–1685, 2017.
- LeCun, Y. and Cortes, C. *The mnist database of handwritten digits*. 1998.
- Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.
- Ledoux, M. and Talagrand, M. *Probability in Banach spaces*. Classics in Mathematics. SpringerVerlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.
- Li, X.-L., Liu, B., and Ng, S.-K. Negative training data can be harmful to text classification. In *Proc. of EMNLP*, pp. 218–228, 2010.
- McDiarmid, C., On the method of bounded differences. In Siemons, J. editor, *Surveys in Combinatorics*, Cambridge University Press, pp. 148–188, 1989.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.
- Natarajan, N., S. Dhillon, I., Ravikumar, P., and Tewari, A. Learning with noisy labels. In *NeurIPS*, pp 1196–1204, 2013.

- Sakai, T., du Plessis, M. C. d., Niu, G., and Sugiyama, M. Semi-supervised classification based on classification from positive and unlabeled data. In *ICML*, volume 70, pp. 2998—3006, 2017.
- Schölkopf, B. and Smola, A. *Learning with Kernels*. MIT Press, 2001.
- Shi, H., Pan, S., Yang, J., and Gong, C. Positive and unlabeled learning via loss decomposition and centroid estimation. In *IJCAI*, pp. 2689—2695, 2018.
- Shinoda, K., Kaji, H., and Sugiyama, M. Binary classification from positive data with skewed confidence. In *IJCAI*, pp. 3328—3334, 2020.
- Yao, Y., Liu, T., Han, B., Gong, M., Niu, G., Sugiyama, M., and Tao, D. Rethinking class-prior estimation for positive-unlabeled learning. In *ICLR*, 2021.
- Watanabe, S. and Matsui, H. Classification from Positive and Biased Negative Data with Skewed Labeled Posterior Probability. In *Neural Computation*, To appear.
- Zhang, Z., Zhao, P., Jiang, Y., and Zhou, Z. Learning from incomplete and inaccurate supervision. In *KDD*, pp. 1017—1025, 2019.

A 付録

A.1 PUbN 分類リスクの導出

$s = -1$ となるデータは, すべて N データになるので, $p(x|y = -1, s = -1) = p(x|s = -1)$ より, 分類リスク (2.1) の第 2 項は, 次のように表される.

$$\begin{aligned}(1 - \pi)R_N^-(g) &= (1 - \pi) \int p(x|y = -1)\ell(-g(x))dx \\ &= \int p(x, y = -1)\ell(-g(x))dx \\ &= \int p(x, y = -1, s = +1)\ell(-g(x))dx + \int p(x, y = -1, s = -1)\ell(-g(x))dx \\ &= p(y = -1, s = +1) \int p(x|y = -1, s = +1)\ell(-g(x))dx \\ &\quad + p(y = -1, s = -1) \int p(x|s = -1)\ell(-g(x))dx \\ &= \rho R_{bN}^-(g) + (1 - \pi - \rho)R_{s=-1}^-(g).\end{aligned}$$

A.2 Pconf 分類リスクの導出

$p(y = +1|x) > 0$ と仮定しているので

$$\begin{aligned}(1 - \pi)p(x|y = -1) &= p(x, y = -1) \\ &= p(x) - p(x, y = +1) \\ &= \frac{p(x, y = +1)}{p(y = +1|x)} - p(x, y = +1) \\ &= \frac{p(x, y = +1)(1 - p(y = +1|x))}{p(y = +1|x)} \\ &= \pi p(x|y = +1) \left(\frac{1 - r}{r} \right),\end{aligned}$$

となる. したがって, 分類リスク (2.1) の第2項は, 次のように表される.

$$\begin{aligned}
(1 - \pi)R_N^-(g) &= \int (1 - \pi)p(x|y = -1)\ell(-g(x))dx \\
&= \int \pi p(x|y = +1) \left(\frac{1-r}{r} \right) \ell(-g(x))dx \\
&= \pi R_P^- \left(\frac{1-r}{r} g \right).
\end{aligned}$$

A.3 補題1の導出

仮定により

$$\frac{1-r}{r} \leq \frac{1}{C_r}$$

となる. ある x_i を x'_i に置き換えたときの経験リスクの変化は, $\hat{R}_P^+(g) \leq C_\ell$, $\hat{R}_P^-\left(\frac{1-r}{r}g\right) \leq C_\ell/C_r$ より, $(C_\ell + C_\ell/C_r)/n_P$ 以上にはならない.

一様偏差の一つの方向である $\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g)$ を考える. $\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g)$ は $\hat{R}(g)$ の変化と上界を共有していることに注意すると, マクダイアミッドの不等式 (McDiamid 1989) より,

$$\Pr \left\{ \sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) - E_{\mathcal{X}_P} \left[\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) \right] \geq \epsilon \right\} \leq \exp \left(- \frac{2\epsilon^2 n_P}{(C_\ell + C_\ell/C_r)^2} \right)$$

が成り立つ. また, $\delta = 2 \exp(-2\epsilon^2 n_P / (C_\ell + C_\ell/C_r)^2)$ と置くと $\epsilon = \sqrt{\frac{\log 2/\delta}{2n_P} (C_\ell + C_\ell/C_r)^2}$ となるので,

$$\Pr \left\{ \sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) - E_{\mathcal{X}_P} \left[\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) \right] \leq \sqrt{\frac{\log 2/\delta}{2n_P} (C_\ell + C_\ell/C_r)^2} \right\} \leq 1 - \delta/2.$$

したがって, 少なくとも $1 - \delta/2$ の確率で以下の式が成り立つ.

$$\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) \leq E_{\mathcal{X}_P} \left[\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) \right] + \left(C_\ell + \frac{C_\ell}{C_r} \right) \sqrt{\frac{\log(2/\delta)}{2n_P}}.$$

$\hat{R}(g)$ は不偏で, $\ell(z)$ はすべての $|z| \leq C_g$ に対してリプシッツ連続であり, リプシッツ定数 L_ℓ を持つと仮定しているので, Mohri et al. (2012) とタラグラントの補題 (Ledoux &

Talagrand 2011) より

$$\begin{aligned}
E_{\mathcal{X}_P} \left[\sup_{g \in \mathcal{G}} \hat{R}(g) - R(g) \right] &\leq 2\mathfrak{R}_{n_P} \left(\left(1 + \frac{1-r}{r}\right) \circ \ell \circ \mathcal{G} \right) \\
&\leq 2 \left(1 + \frac{1}{C_r}\right) \mathfrak{R}_{n_P}(\ell \circ \mathcal{G}) \\
&\leq 2 \left(L_\ell + \frac{L_\ell}{C_r}\right) \mathfrak{R}_{n_P}(\mathcal{G})
\end{aligned}$$

となり, この方向が導けた. 他の方向 $\sup_{g \in \mathcal{G}} R(g) - \hat{R}(g)$ も同様に導ける.

A.4 定理 1 の導出

補題 1 より, 以下のように導出される.

$$\begin{aligned}
R(\hat{g}) - R(g^*) &= (\hat{R}(\hat{g}) - \hat{R}(g^*)) + (R(\hat{g}) - \hat{R}(\hat{g})) + (\hat{R}(g^*) - R(g^*)) \\
&\leq 0 + 2 \sup_{g \in \mathcal{G}} |\hat{R}(g) - R(g)| \\
&\leq 4 \left(L_\ell + \frac{L_\ell}{C_r}\right) \mathfrak{R}_{n_P}(\mathcal{G}) + 2 \left(C_\ell + \frac{C_\ell}{C_r}\right) \sqrt{\frac{\log(2/\delta)}{2n_P}},
\end{aligned}$$

ここで, \hat{R} の定義より, $\hat{R}(\hat{g}) \leq \hat{R}(g^*)$ となる.

A.5 PbN 分類リスクの導出

分類リスク (2.1) の第 2 項は, $p(x, s = -1, y = +1) = 0$ より, $p(x, s = -1) = p(x, s = -1, y = -1)$ となるので, 次のように分解できる.

$$\begin{aligned}
(1 - \pi)R_N^-(g) &= (1 - \pi) \int \ell(-g(x))p(x|y = -1)dx \\
&= \int p(y = -1)p(x|y = -1)\ell(-g(x))dx \\
&= \int p(x, y = -1)\ell(-g(x))dx \\
&= \int [p(x, y = -1, s = +1) + p(x, s = -1)] \ell(-g(x))dx \\
&= \int p(y = -1, s = +1)p(x|y = -1, s = +1)\ell(-g(x))dx \\
&\quad + \int p(x, s = -1)\ell(-g(x))dx \\
&= \rho R_{bN}^-(g) + \int p(x, s = -1)\ell(-g(x))dx.
\end{aligned} \tag{A.1}$$

ここで, $p(s = +1|x) \neq 0$ と仮定しているので

$$\begin{aligned} p(x, s = -1) + p(x, s = +1) &= p(x) \\ &= \frac{p(x, s = +1)}{p(s = +1|x)} \end{aligned}$$

となり,

$$\begin{aligned} p(x, s = -1) &= \frac{p(x, s = +1) [1 - p(s = +1|x)]}{p(s = +1|x)} \\ &= \frac{1 - \sigma}{\sigma} p(x, s = +1). \end{aligned}$$

となる. したがって, (A.1) の第 2 項は次のように表される.

$$\begin{aligned} \int p(x, s = -1) \ell(-g(x)) dx &= \int \frac{1 - \sigma}{\sigma} p(x, s = +1) \ell(-g(x)) dx \\ &= \int \frac{1 - \sigma}{\sigma} p(s = +1) p(x|s = +1) \ell(-g(x)) dx \\ &= (\pi + \rho) R_{s=+1}^- \left(\frac{1 - \sigma}{\sigma} g \right). \end{aligned}$$

A.6 補題 2 の導出

劣加法性により

$$\begin{aligned} &\sup_{g \in \mathcal{G}} |\hat{R}(g) - R(g)| \\ &\leq \pi \sup_{g \in \mathcal{G}} |\hat{R}_P^+(g) - R_P^+(g)| + \rho \sup_{g \in \mathcal{G}} |\hat{R}_{bN}^-(g) - R_{bN}^- P(g)| + (\pi + \rho) \sup_{g \in \mathcal{G}} |\hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g)| \end{aligned} \quad (\text{A.2})$$

が成り立つ. また, 仮定により

$$\frac{1 - \sigma}{\sigma} \leq \frac{1}{C_\sigma}$$

となる. 次に, $\sup_{g \in \mathcal{G}} |\hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g)|$ の一方向 $\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g)$ について考える. ある x_i を x'_i に置き換えたときの経験リスク $\hat{R}_{s=+1}^-$ の変化は, $(C_\ell/C_\sigma)/(n_P + n_{bN})$ 以上にはならない.

したがって, $\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g)$ は $\hat{R}_{s=+1}^-$ の変化と上界を共有していることに注意すると, マクダイアミッドの不等式 (McDiamid 1989) より,

$$\begin{aligned} \Pr \left\{ \sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) - E_{\mathcal{X}_{s=+1}} \left[\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) \right] \geq \epsilon \right\} \\ \leq \exp \left(- \frac{2\epsilon^2(n_P + n_{bN})}{(C_\ell/C_\sigma)^2} \right) \end{aligned}$$

が成り立つ. また, $\delta = 6 \exp(-2\epsilon^2(n_P + n_{bN})/(C_\ell/C_\sigma)^2)$ と置くと $\epsilon = \sqrt{\frac{\log 6/\delta}{2(n_P + n_{bN})}}(C_\ell/C_\sigma)^2$ となるので,

$$\begin{aligned} \Pr \left\{ \sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) - E_{\mathcal{X}_{s=+1}} \left[\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) \right] \leq \sqrt{\frac{\log 6/\delta}{2(n_P + n_{bN})}}(C_\ell/C_\sigma)^2 \right\} \\ \leq 1 - \delta/6 \end{aligned}$$

となる. したがって, 少なくとも $1 - \delta/6$ の確率で以下の式が成り立つ.

$$\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) \leq E_{\mathcal{X}_{s=+1}} \left[\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) \right] + \frac{C_\ell}{C_\sigma} \sqrt{\frac{\log(6/\delta)}{2(n_P + n_{bN})}}.$$

$\hat{R}_{s=+1}^-(g)$ は不偏で, $\ell(z)$ はすべての $|z| \leq C_g$ に対してリプシッツ連続であり, リプシッツ定数 L_ℓ を持つと仮定しているので, Mohri et al. (2012) とタラグランの補題 (Ledoux & Talagrand 2011) より

$$\begin{aligned} E_{\mathcal{X}_{s=+1}} \left[\sup_{g \in \mathcal{G}} \hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g) \right] &\leq 2\mathfrak{R}_{s=+1} \left(\frac{1-\sigma}{\sigma} \circ \ell \circ \mathcal{G} \right) \\ &\leq \frac{2}{C_\sigma} \mathfrak{R}_{s=+1}(\ell \circ \mathcal{G}) \\ &\leq \frac{2L_\ell}{C_\sigma} \mathfrak{R}_{s=+1}(\mathcal{G}), \end{aligned}$$

となる. $\sup_{g \in \mathcal{G}} R_{s=+1}^-(g) - \hat{R}_{s=+1}^-(g)$ も同様に導かれる.

さらに, $\sup_{g \in \mathcal{G}} |\hat{R}_P^+(g) - R_P^+(g)|$, $\sup_{g \in \mathcal{G}} |\hat{R}_{bN}^-(g) - R_{bN}^-(g)|$ の上界についても同様の手順で導けるので, 少なくとも $1 - \delta/3$ の確率で, 以下の境界が別々に成立する.

$$\sup_{g \in \mathcal{G}} |\hat{R}_P^+(g) - R_P^+(g)| \leq 2L_\ell \mathfrak{R}_{n_P}(\mathcal{G}) + C_\ell \sqrt{\frac{\log 6/\delta}{2n_P}}, \quad (\text{A.3})$$

$$\sup_{g \in \mathcal{G}} |\hat{R}_{bN}^-(g) - R_{bN}^-(g)| \leq 2L_\ell \mathfrak{R}_{n_{bN}}(\mathcal{G}) + C_\ell \sqrt{\frac{\log 6/\delta}{2n_{bN}}}, \quad (\text{A.4})$$

$$\sup_{g \in \mathcal{G}} |\hat{R}_{s=+1}^-(g) - R_{s=+1}^-(g)| \leq \frac{2L_\ell}{C_\sigma} \mathfrak{R}_{s=+1}(\mathcal{G}) + \frac{C_\ell}{C_\sigma} \sqrt{\frac{\log(6/\delta)}{2(n_P + n_{bN})}}. \quad (\text{A.5})$$

したがって, (A.2)-(A.5) より, 補題 2 が導けた.

A.7 定理 2 の導出

補題 2 より, 以下のように導出される.

$$\begin{aligned}
R(\hat{g}) - R(g^*) &= (\hat{R}(\hat{g}) - \hat{R}(g^*)) + (R(\hat{g}) - \hat{R}(\hat{g})) + (\hat{R}(g^*) - R(g^*)) \\
&\leq 0 + 2 \sup_{g \in \mathcal{G}} |\hat{R}(g) - R(g)| \\
&\leq 4\pi L_\ell \mathfrak{R}_{n_P}(\mathcal{G}) + 4\rho L_\ell \mathfrak{R}_{n_{bN}}(\mathcal{G}) + 4(\pi + \rho)(L_\ell/C_\sigma) \mathfrak{R}_{s=+1}(\mathcal{G}) \\
&\quad + 2\pi C_\ell \sqrt{\frac{\log 6/\delta}{2n_P}} + 2\rho C_\ell \sqrt{\frac{\log 6/\delta}{2n_{bN}}} + 2(\pi + \rho)(C_\ell/C_\sigma) \sqrt{\frac{\log 6/\delta}{2(n_P + n_{bN})}},
\end{aligned}$$

ここで, \hat{R} の定義より, $\hat{R}(\hat{g}) \leq \hat{R}(g^*)$ となる.