統計的学習理論読み（Chapter 3）

松井孝太

名古屋大学大学院医学系研究科 生物統計学分野
matsui.k@med.nagoya-u.ac.jp

Table of contents

導入

本スライドは [4] の第 3 章のまとめである.

- 判別問題においてサロゲート損失を用いることの正当化
- 理論的に良いサロゲート損失とは何か？

がメイントピック

Table of contents

判別適合的損失

判別適合的損失

マージン損失

マージン損失

- $\mathcal{X}$: input space, $\mathcal{Y} = \{+1, -1\}$ binary label
- $\mathcal{G} \subset \{g : \mathcal{X} \to \mathbb{R}\}$ a set of classification functions
- $\mathcal{H} = \{\mathrm{sign} \circ g \mid g \in \mathcal{G}\}$: hypothesis set

Definition 1 (margin, margin loss)

判別関数 $g \in \mathcal{G}$ と *data* $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ に対して $yg(\boldsymbol{x})$ を *margin* という. また, 非負値関数 $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ に対して g の *margin loss* を $\phi(yg(\boldsymbol{x}))$ で定義する.

Example 1 (0-1 margin loss)

$m \in \mathbb{R}$ に対して $\phi_{err} : m \mapsto \mathbf{1}[m \leq 0]$ で定義される *margin loss* を考える.

0-1 loss $\ell_{err}(\mathrm{sign}(g(\boldsymbol{x})), y) = \begin{cases} +1 & \textit{if} \quad \mathrm{sign}(g(\boldsymbol{x})) \neq y \\ 0 & \textit{otherwise} \end{cases}$ に対して,

$\ell_{err}(\mathrm{sign}(g(\boldsymbol{x})), y) \begin{cases} = \phi_{err}(yg(\boldsymbol{x})) & \textit{if} \quad g(\boldsymbol{x}) \neq 0 \\ \leq \phi_{err}(0) = 1 & \textit{if} \quad g(\boldsymbol{x}) = 0 \end{cases}$ が成立.

Definition 2 (経験・予測 ϕ-損失)

判別関数 g に対して,

$$\hat{R}_\phi(g) := \frac{1}{n}\sum_{i=1}^{n}\phi(y_i g(\boldsymbol{x}_i)) \qquad \textit{(empirical } \phi\textit{-loss)}$$

$$R_\phi(g) := \mathbb{E}[\phi(Yg(X))] \qquad \textit{(predictive } \phi\textit{-loss)}$$

特に,

$$\hat{R}_{err}(g) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[\mathrm{sign}(g(\boldsymbol{x}_i)) \neq y] \qquad \textit{(empirical classification error)}$$

$$R_{err}(g) := \mathbb{E}[\mathbf{1}[\mathrm{sign}(g(X)) \neq Y]] \qquad \textit{(predictive classification error)}$$

と定義

経験・予測 ϕ-損失 II

Remark 1

$\forall m \in \mathbb{R}$, $\phi_{err}(m) \leq \phi(m)$ のとき,

$$\hat{R}_{err}(g) \leq \hat{R}_{\phi}(g)$$
$$R_{err}(g) \leq R_{\phi}(g)$$

が成立 (ϕ-損失は判別誤差の上界を与える).

Notation

- $R_{\phi}^{*} := \inf_{g:\text{measurable}} R_{\phi}(g)$
- $R_{err}^{*} := \inf_{g:\text{measurable}} R_{err}(g)$

これから考えること

上界が小さければ $\hat{R}_{err}, R_{err}$ も小さい (ので上界で surrogate する)
$\longrightarrow R_{\phi}(g) - R_{\phi}^{*}$ と $R_{err}(g) - R_{err}^{*}$ の関係を評価する

Surrogate loss

Definition 3 (surrogate loss)

- *hinge loss (SVM)*

$$\phi(m) := \max\{0, 1-m\}$$

- *exponential loss (boosting)*

$$\phi(m) := e^{-m}$$

- *logistic loss (logistic regression)*

$$\phi(m) := \log(1+e^{-m})$$

これらは全て単調非増加な凸関数

- $\hat{R}_\phi(g)$ の最小化によりデータ上でマージン $yg(\boldsymbol{x})$ が大きくなり, 多くの学習データで $\mathrm{sign}(g(\boldsymbol{x})) = y$ の成立が期待される.
- 凸性から最適化計算が効率的に実行できる.

判別適合的損失

判別適合的損失

予測 ϕ-損失と予測判別誤差との関係

$R_\phi(g)$ と $R_{err}(g)$ との関係を調べる.

まず定義から,

$$R_\phi(g) = \mathbb{E}[\phi(Yg(X))] = \mathbb{E}_X\left[\mathbb{E}_Y[\phi(Yg(X)) \mid X]\right]$$

と書ける (条件付き期待値の条件に関する期待値).

関数 $C_\eta(\alpha)$ を

$$C_\eta(\alpha) := \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$$

とおくと, 内部の期待値は以下のように書ける:

$$\begin{aligned}
&\mathbb{E}_Y[\phi(Yg(X)) \mid X] \\
=&\Pr(Y=+1 \mid X)\phi(g(X)) + \underbrace{(1-\Pr(Y=+1 \mid X))}_{=\Pr(Y=-1 \mid X)}\phi(-g(X)) \\
=&C_\eta(g(X))
\end{aligned}$$

ここで, $\eta = \Pr(Y=+1 \mid X)$ とおいた.

予測 ϕ-損失と予測判別誤差との関係 II

C_η と Bayes rule との関係

R_{err} を最小にする Bayes rule h_0 は,

$$h_0(x) = \underset{y \in \{+1,-1\}}{\arg\max}\ \Pr(Y = y \mid X = x)$$

で与えられる. $\eta = \Pr(Y = +1 \mid X)$ の言葉で書くと,

$$\eta > \frac{1}{2} \Longrightarrow \hat{y} = +1$$
$$\eta < \frac{1}{2} \Longrightarrow \hat{y} = -1$$

Proposition 1

$\mathrm{sign} \circ g^* \ \big(g^* = \arg\min C_\eta(g(X))\big)$ が *Bayes rule*
$\Longleftrightarrow \forall \eta \in [0,1] \backslash \{\frac{1}{2}\}$ に対して,

$$\inf_{g : g(X)(\eta - \frac{1}{2}) \leq 0} C_\eta(g(X)) > \inf_g C_\eta(g(X))$$

予測 ϕ-損失と予測判別誤差との関係 III

(Prop の説明)

判別器 g について, 以下が成立していてほしい:

$$\mathrm{sign}(g(x)) = \mathrm{sign}(\alpha) = \underbrace{\mathrm{sign}(\eta - \frac{1}{2})}_{Bayes\ rule}$$

すなわち,

$$\alpha\left(\eta - \frac{1}{2}\right) \geq 0. \tag{1}$$

いま, 上記を満たさない領域で C_η を最小化し, その最適解を $\hat{\alpha}$ とおく:

$$\hat{\alpha} = \mathop{\arg\min}_{\alpha:\alpha(\eta-\frac{1}{2})\leq 0} C_\eta(\alpha)$$

もし, $\hat{\alpha} \leq \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha)$ であれば, $\hat{\alpha}$ に対応する判別関数 $\hat{g}$ から構成される仮説 $\mathrm{sign} \circ \hat{g}$ は Bayes rule とはならない ((1) が満たされないから).

Classification calibrated loss

Definition 4 (classification calibrated loss)

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} C_\eta(\alpha), \quad H^-(\eta) = \inf_{\alpha(\eta - \frac{1}{2}) \leq 0} C_\eta(\alpha)$$

とおく. $\forall \eta \in [0,1] \backslash \{\frac{1}{2}\}$ に対して,

$$H^-(\eta) > H(\eta)$$

が成立するとき, マージン損失 ϕ を *classification-calibrated loss (CC-loss)* と呼ぶ.

Remark 2

CC-loss ϕ を採用したとき, 予測 ϕ-損失 $R_\phi(g)$ を最小にする判別関数 g は *Bayes rule* を与える.

$\theta \in [-1, 1]$ に対して, ψ_0 を以下で定義

$$\psi_0(\theta) := H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$$

Proposition 2

マージン損失 ϕ が *CC-loss* $\iff \forall\theta \neq 0,\ \psi_0(\theta) > 0$

($\because$) ($\Rightarrow$) ϕ が CC-loss とすると, $\forall\eta \in [0,1]\backslash\{\frac{1}{2}\}$ に対して $H^-(\eta) > H(\eta)$ が成立. ここで, $\theta \in [-1,1]\backslash\{0\}$ に対して

$$\frac{1+\theta}{2} \in [0,1]\backslash\frac{1}{2}$$

を η とおけば,

$$\psi_0(\theta) = H^-(\eta) - H(\eta) > 0$$

が言える. ($\Leftarrow$) は以上を逆にたどれば良い. $\square$

Classification calibrated loss III

ψ_0 に対して, 以下の条件を満たす凸関数 ψ を考える (ψ_0 の凸包と呼ぶ)

1. $\psi_0 \geq \psi$
2. $\forall \tilde{\psi}$: convex, $\quad \psi_0 \geq \tilde{\psi} \Rightarrow \psi \geq \tilde{\psi}$

Proposition 3

$$\psi(\theta) = \sup_{\tilde{\psi}} \left\{ \tilde{\psi}(\theta) \mid \forall \tilde{\psi} : cvx \text{ s.t. } \psi_0 \geq \tilde{\psi} \right\}$$

($\because$) 1, 2, 及び Prop B6 の 3 より □

Prop B6-3

$\forall I$: index set, $\{f_i\}_{i \in I}$: convex family に対して, $f(x) = \sup_{i \in I} f_i(x)$ は convex function

Classification calibrated loss IV

Lemma 1 (3.3)

1. $\psi_0(\theta)$ の *convex-hull* $\psi(\theta)$ は $(-1,1)$ 上連続関数かつ $\psi(0)=0$
2. ψ_0, ψ は共に $[-1,1]$ 上の偶関数

(proof) (1) convex-hull ψ は $[-1,1]$ 上の凸関数だから, Thm B7 より $(-1,1)$ 上で連続.

Thm B7 (凸関数の連続性)

凸関数 f に対して $\mathrm{int}(\mathrm{dom}(f)) \neq \emptyset$ のとき, f は $\mathrm{int}(\mathrm{dom}(f))$ 上で連続

$\theta = 0$ のとき,

$$H\left(\frac{1}{2}\right) = \inf_{\alpha} C_{1/2}(\alpha) = H^{-}\left(\frac{1}{2}\right)$$

となる ($\eta - 1/2 = 0$ なので H^{-} における α の制約はなくなる). よって $\psi_0(\frac{1}{2}) = 0$.

(proof つづき)

- ϕ が CC-loss のとき, $H^-(\eta) - H(\eta) > 0, \forall \eta \neq \frac{1}{2}$ であるから, 上の事実と合わせると $\psi_0 \geq 0$ (非負関数) であることが分かる.
- $\tilde{\psi} = 0$ (定値関数) とすると, ψ_0 の convex-hull ψ に対して定義より $\psi \geq 0$ が成立

以上より,

$$0 = \psi_0(0) \geq \psi(0) \geq 0$$

だから, $\psi = 0$ が成立 (真中の不等式は ψ の定義の (1) より).

<u>偶関数であること</u> 定義より $C_\eta(\alpha) = C_{1-\eta}(-\alpha)$ なので,

$$H(\eta) = H(1-\eta)$$

が成立 (α の前の符号は inf で吸収される)

(proof つづき) また, $\alpha(2\eta - 1) \geq 0 \iff -\alpha(2(1-\eta) - 1) \geq 0$ なので,

$$H^-(\eta) = \inf_{\alpha(2\eta-1)\leq 0} C_\eta(\alpha) = \inf_{-\alpha(2(1-\eta)-1)\leq 0} C_{1-\eta}(-\alpha) = H^-(1-\eta)$$

よって,

$$\begin{aligned}\psi_0(\theta) &= H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right) = H^-\left(1 - \frac{1+\theta}{2}\right) - H\left(1 - \frac{1+\theta}{2}\right) \\ &= H^-\left(\frac{1-\theta}{2}\right) - H\left(\frac{1-\theta}{2}\right) = \psi_0(-\theta)\end{aligned}$$

だから, ψ_0 は偶関数. いま, $\forall\tilde{\psi}$ s.t. $\psi_0 \geq \tilde{\psi}$ に対して

$$\psi(\theta) = \max\{\tilde{\psi}(\theta), \tilde{\psi}(-\theta)\}$$

とおくと, ψ は偶関数で, $\psi_0 \geq \psi \geq \tilde{\psi}$ となるので, ψ は ψ_0 の convex-hull. □

Theorem 2 (3.4 予測 ϕ-loss と予測判別誤差の関係)

$\forall \phi$: margin loss, $\forall f$: classifier, $\forall D$: distribution,

$$\psi(R_{err}(f) - R^*_{err}) \leq R_\phi(f) - R^*_\phi$$

i.e. expected risk の上界の *expected ϕ-risk* が与える.

(proof) $\eta(X) = \Pr(Y = +1|X)$ とおく. このとき, Bayes rule は

$$\eta(X) - \frac{1}{2} > 0 \Longrightarrow y = +1$$
$$\eta(X) - \frac{1}{2} < 0 \Longrightarrow y = -1$$

よって,

$$\begin{aligned} R_{err}(f) - R^*_{err} =& \mathbb{E}_X[\mathbb{E}_Y[\mathbf{1}_{\mathrm{sign}(f(X)) \neq Y}|X]] \\ &- \mathbb{E}_X[\mathbb{E}_Y[\mathbf{1}_{\mathrm{sign}(\eta(X)-\frac{1}{2}) \neq Y}|X]] \\ =& \mathbb{E}_X[\mathbb{E}_Y[\mathbf{1}_{\mathrm{sign}(f(X)) \neq Y} - \mathbf{1}_{\mathrm{sign}(\eta(X)-\frac{1}{2}) \neq Y}|X]] \end{aligned}$$

さらに,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbf{1}_{\mathrm{sign}(f(X))\neq Y} - \mathbf{1}_{\mathrm{sign}(\eta(X)-\frac{1}{2})\neq Y}|X] \\
&= \left\{\mathbf{1}_{\mathrm{sign}(f(X))\neq +1} - \mathbf{1}_{\mathrm{sign}(\eta(X)-\frac{1}{2})\neq +1}\right\}\eta(X) \\
&\quad + \left\{\mathbf{1}_{\mathrm{sign}(f(X))\neq -1} - \mathbf{1}_{\mathrm{sign}(\eta(X)-\frac{1}{2})\neq -1}\right\}(1-\eta(X))
\end{aligned}$$

ここで, $\mathrm{sign}(f(X)) = +1$ かつ $\mathrm{sign}(\eta(X) - \frac{1}{2}) = -1$ のとき,

$$\mathrm{r.h.s} = (0-1)\eta(X) + (1-0)(1-\eta(X)) = 1 - 2\eta(X)$$

また, $\mathrm{sign}(f(X)) = -1$ かつ $\mathrm{sign}(\eta(X) - \frac{1}{2}) = +1$ のとき,

$$\mathrm{r.h.s.} = (1-0)\eta(X) + (0-1)(1-\eta(X)) = 2\eta(X) - 1$$

以上を合わせると,

$$\mathrm{r.h.s.} = \mathbf{1}_{\mathrm{sign}(f(X))\neq \mathrm{sign}(\eta(X)-\frac{1}{2})} \times |2\eta(X) - 1|$$

Classification calibrated loss !X

よって,

$$
\begin{aligned}
&\psi(R_{err}(f) - R^*_{err}) \\
(\text{Jensen } \to) \leq &\mathbb{E}[\psi(\mathbf{1}_{\mathrm{sign}(f(X))\neq\mathrm{sign}(\eta(X)-\frac{1}{2})} \times |2\eta(X)-1|)] \\
= &\mathbb{E}[\mathbf{1}_{\mathrm{sign}(f(X))\neq\mathrm{sign}(\eta(X)-\frac{1}{2})} \times \psi(|2\eta(X)-1|)] \\
\leq &\mathbb{E}[\mathbf{1}_{\mathrm{sign}(f(X))\neq\mathrm{sign}(\eta(X)-\frac{1}{2})} \times \psi_0(|2\eta(X)-1|)]
\end{aligned}
$$

- 真中の $=$ は, $\mathbf{1} = 1$ or 0 より,

$$
\begin{cases}
\psi(\mathbf{1} \times x) = \mathbf{1} \times \psi(x) & \text{if } \mathbf{1} = 1 \\
\psi(\mathbf{1} \times x) = 0 = \mathbf{1} \times \psi(x) & \text{if } \mathbf{1} = 0
\end{cases}
$$

  から従う.
- 最後の $\leq$ は, $\psi_0 \geq \psi$ から従う.

Classification calibrated loss X

また,

$$
\begin{aligned}
&\psi_0(|2\eta(X)-1|)\\
=&H^-\left(\frac{1+|2\eta(X)-1|}{2}\right)-H\left(\frac{1+|2\eta(X)-1|}{2}\right)\\
=&\begin{cases} H^-(\eta(X))-H(\eta(X)) & \text{if } 2\eta(X)-1>0\\ H^-(1-\eta(X))-H(1-\eta(X)) & \text{otherwise}\end{cases}
\end{aligned}
$$

かつ $H^-(\eta)=H^-(1-\eta),\ \ H(\eta)=H(1-\eta)$ より,

$$
\begin{aligned}
\mathbb{E}[\{\cdot\}\times\psi_0(|2\eta(X)-1|)] &= \mathbb{E}[\{\cdot\}\times(H^-(\eta(X))-H(\eta(X)))]\\
&\le \mathbb{E}[\{\cdot\}\times(C_{\eta(X)}(f(X))-H(\eta(X)))]\\
&\quad\left((\because)H^-(\eta(X))=\inf_{f(X)}C_{\eta(X)}(f(X))\right)\\
&\le \mathbb{E}[C_{\eta(X)}(f(X))-H(\eta(X))]\\
&= R_\phi(f)-R_\phi^*
\end{aligned}
$$

以上より, $\psi(R_{err}(f)-R_{err}^*)\le R_\phi(f)-R_\phi^* \quad \square$

判別適合的損失

凸マージン損失

Convex Margin Loss

convex function ϕ に対して ϕ-margin loss が C.C. かどうかを判定する

Theorem 3 (3.5 C.C. Theorem for convex margin loss)

ϕ が *convex, differentiable,* $\phi'(0) < 0$ のとき, 以下が成立

1. ϕ*-margin loss* は *C.C.*
2. $\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right)$
3. $\forall\{f_i\}$ *: measurable functions,* $\forall D$ *: distribution on* $\mathcal{X} \times \{\pm 1\}$

$$R_\phi(f_i) \to R_\phi^* \Longrightarrow R_{err}(f_i) \to R_{err}^*$$

Remark

予測 ϕ 損失が小さい $\Rightarrow$ 予測判別誤差 $\approx$ Bayes error

よって $\hat{g} = \arg\min_g \hat{R}_\phi(g)$ に対して, $R_\phi(\hat{g})$ が小さい $\Rightarrow R_{err}(\hat{g}) \approx R_{err}^*$

3 の証明は 3.4 節の定理 3.6 で一般の margin loss について行う

(proof) 1 の証明

$\eta > \frac{1}{2}$ のとき, $C_\eta(\alpha) = \eta\phi(\alpha) + (1-\eta)\phi(\alpha)$ が $\alpha \leq 0$ で最小値をとらないことを示す ($\eta > \frac{1}{2}$ のときも同様のことが $\alpha \geq 0$ に対して示せる)

これが言えると

$$H(\eta) = \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha), \quad H^-(\eta) = \inf_{\alpha(2\eta-1)\leq 0} C_\eta(\alpha)$$

に対して, $\eta > \frac{1}{2}$ のとき, $\alpha(2\eta-1) \geq 0 \iff \alpha \leq 0$ だから, $C_\eta(\alpha)$ が $\alpha \leq 0$ で最小値を取らなければ

$$H^-(\eta) > H(\eta)$$

が成立. 一方, $\eta < \frac{1}{2}$ のとき, $\alpha(2\eta-1) \geq 0 \iff \alpha \geq 0$ だから, $C_\eta(\alpha)$ が $\alpha \geq 0$ で最小値を取らなければ, 同様に $H^-(\eta) > H(\eta)$ が成立.

両者を合わせると, Def 3.2 より ϕ が C.C. であると言える.

Convex Margin Loss III

(proof) 1 の証明つづき $C_\eta(\alpha)$ を $\alpha = 0$ で微分すると,

$$C'_\eta(0) = \frac{\mathrm{d}}{\mathrm{d}\alpha} C_\eta(\alpha)\Big|_{\alpha=0} = \eta\phi'(0) - (1-\eta)\phi'(0) = (2\eta - 1)\phi'(0)$$

$\phi'(0) < 0$ より, $\eta > \frac{1}{2}$ に対して, $C'_\eta(0) < 0$ が成立. このとき以下が成立

$$\exists \alpha_0 > 0 \ \text{ s.t. } \ \frac{C_\eta(\alpha_0) - C_\eta(0)}{\alpha_0} < \frac{C'_\eta(0)}{2} < 0 \quad (3.5)$$

($\because$) def より, $\forall \varepsilon > 0$, $\exists \delta > 0$ s.t. $\forall \alpha$ with $|\alpha| < \delta$, $\left|\frac{C_\eta(\alpha) - C_\eta(0)}{\alpha} - C'_\eta(0)\right| < \varepsilon$ であり, $C'_\eta(0) < 0$ より, $\frac{C'_\eta(0)}{2} > C'_\eta(0)$ が成立. 一方実数の連続性から次が成立:

$$\forall \varepsilon > 0, \exists \alpha_0 \ \text{ s.t. } \ \frac{C_\eta(\alpha_0) - C_\eta(0)}{\alpha_0} < C'_\eta(0) + \varepsilon$$

特に, $C'_\eta(0) + \varepsilon < \frac{C'_\eta(0)}{2}$ となるように ε をとれば良い.

Convex Margin Loss IV

(proof) 1 の証明つづき 一方, $C_\eta(\alpha)$ は convex (ϕ が convex なので) だから,

$$\forall\alpha \in \mathbb{R}, \quad C_\eta(\alpha) \geq C_\eta(0) + C'_\eta(0)(\alpha - 0)$$

が成立 (convex function の特徴付け). よって, $\alpha \leq \frac{\alpha_0}{2}$ なる任意の $\alpha \in \mathbb{R}$ で,

$$\begin{aligned} C_\eta(\alpha) &\geq C_\eta(0) + \alpha C'_\eta(0) \\ &\geq C_\eta(0) + \frac{\alpha_0}{2} C'_\eta(0) \\ \text{by (3.5)} \rightarrow &> C_\eta(\alpha_0) \end{aligned}$$

が成立. 以上より, $C_\eta(\alpha)$ は $\alpha \leq 0$ では最小値をとらない.

Remark

赤字の部分は, 恐らく "$\forall\alpha \leq 0$ で" で良い

Convex Margin Loss V

(proof) 2 の証明

$$\begin{aligned}\phi(0) = \eta\phi(0) + (1-\eta)\phi(0) &= C_\eta(0) \\ &\geq \inf_{\alpha(2\eta-1)\leq 0} C_\eta(\alpha) = H^-(\eta)\end{aligned}$$

が成立. また, ϕ の convexity から, $\forall\alpha \in \mathbb{R},\ \ \phi(\alpha) \geq \phi(0) + \alpha\phi'(0)$. よって,

$$\begin{aligned}\phi(0) &\geq \inf_{\alpha(2\eta-1)\leq 0} \eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \\ &\geq \inf_{\alpha(2\eta-1)\leq 0} \eta(\phi(0) + \alpha\phi'(0)) + (1-\eta)(\phi(0) - \alpha\phi'(0)) \\ &= \phi(0) + \inf_{\alpha(2\eta-1)\leq 0} \alpha(2\eta-1)\phi'(0) \\ &= \phi(0)\end{aligned}$$

- 最後の等号は, $\phi'(0) < 0$ より, $\alpha(2\eta-1)\phi'(0) \geq 0$ なので, 第 2 項が 0 となることから従う.

Convex Margin Loss VI

(proof) <u>2 の証明つづき</u> 従って $\phi(0) = H^{-}(\eta)$ が言えるので,

$$\psi_0(\theta) = \phi(0) - H\left(\frac{1 = \theta}{2}\right)$$

を得る. いま, $\alpha^* = \arg\inf C_{\frac{1+\theta}{2}}(\alpha)$ とすると,

$$\begin{aligned}
H\left(\frac{1+\theta}{2}\right) &= \inf_{\alpha \in \mathbb{R}} C_{\frac{1+\theta}{2}}(\alpha) \\
&= \frac{1+\theta}{2}\phi(\alpha^*) + \frac{1-\theta}{2}\phi(-\alpha^*) \\
&= \frac{\phi(\alpha^*) - \phi(-\alpha^*)}{2}\theta + \frac{\phi(\alpha^*) - \phi(-\alpha^*)}{2}
\end{aligned}$$

より, $H\left(\frac{1+\theta}{2}\right)$ は θ に関して線形 (特に, concave).

よって, ψ_0 は convex - concave = convex + convex = convex だから,

$\psi_0(\theta) = \psi(\theta)$ □

Convex Margin Loss VII

Remark

定理 3.5 の逆も成立つ. i.e.

convex margin loss ϕ が C.C. $\Longrightarrow$ $\phi(\alpha)$ は $\alpha = 0$ で微分可能で $\phi'(0) < 0$

Example : Exponential Loss I

$$\phi(m) = e^{-m}$$

とすると,

- ϕ は $m = 0$ で微分可能
- $\phi'(0) = -1 < 0$

だから, 定理 3.5 の仮定を満たす. よって ϕ は C.C. loss

また,

$$\begin{aligned}
&C_\eta(\alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha) = \eta e^{-\alpha} + (1-\eta)e^{\alpha}, \\
&\frac{\mathrm{d}}{\mathrm{d}\alpha} C_\eta(\alpha) = -\eta e^{-\alpha} + (1-\eta)e^{\alpha} = 0 \\
&\iff \log\eta + \log e^{-\alpha} = \log(1-\eta) + \log e^{\alpha}
\end{aligned}$$

より, $C_\eta(\alpha)$ は $\alpha = \frac{1}{2}\log\frac{\eta}{1-\eta}$ で最小値をとる.

よって,

$$
\begin{aligned}
H(\eta) &= C_\eta\left(\frac{1}{2}\log\frac{\eta}{1-\eta}\right) \\
&= \eta\exp\left\{-\frac{1}{2}\log\frac{\eta}{1-\eta}\right\} + (1-\eta)\exp\left\{\frac{1}{2}\log\frac{\eta}{1-\eta}\right\} \\
&= \eta\left(\frac{\eta}{1-\eta}\right)^{-1/2} + (1-\eta)\left(\frac{\eta}{1-\eta}\right)^{1/2} \\
&= \eta\times\sqrt{\frac{1-\eta}{\eta}} + (1-\eta)\times\sqrt{\frac{\eta}{1-\eta}} = 2\sqrt{\eta(1-\eta)}
\end{aligned}
$$

となり,

$$
\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right) = 1 - \sqrt{1-\theta^2}
$$

を得る ($[0,1]$ 上 strictly monotone increase)

Example : Logistic Loss I

$$\phi(m) = \log(1 + e^{-m})$$

とすると,

- ϕ は $m = 0$ で微分可能
- $\phi'(0) = -\frac{1}{2} < 0$

だから, 定理 3.5 の仮定を満たす. よって ϕ は C.C. loss

また,

$$\begin{aligned}
&C_\eta(\alpha) = \eta \log(1 + e^{-\alpha}) + (1 - \eta)(1 + e^{\alpha}), \\
&\frac{\mathrm{d}}{\mathrm{d}\alpha} C_\eta(\alpha) = \frac{-\eta e^{-\alpha}}{1 + e^{-\alpha}} + \frac{(1 - \eta)e^{\alpha}}{1 + e^{\alpha}} = 0 \\
&\iff (1 - \eta)e^{\alpha} = \frac{\eta(e^{-\alpha} + 1)}{1 + e^{-\alpha}} = \eta
\end{aligned}$$

Example : Logistic Loss II

よって,

$$\begin{aligned}
H(\eta) &= C_\eta\left(\log\frac{\eta}{1-\eta}\right)\\
&= \eta\log\left(1+\exp\left\{-\log\frac{\eta}{1-\eta}\right\}\right)+(1-\eta)\log\left(1+\exp\left\{\log\frac{\eta}{1-\eta}\right\}\right)\\
&= \eta\left(1+\frac{1-\eta}{\eta}\right)+(1-\eta)\left(1+\frac{\eta}{1-\eta}\right)\\
&= \eta\times\log\frac{1}{\eta}+(1-\eta)\times\log\frac{1}{1-\eta} = -\eta\log\eta-(1-\eta)\log(1-\eta)
\end{aligned}$$

となり, これは binary random variable に対する entropy に相当する. また,

$$\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right) = \log 2 + \frac{1+\theta}{2}\log\frac{1+\theta}{2} + \frac{1-\theta}{2}\log\frac{1-\theta}{2}$$

を得る ($[0,1]$ 上 strictly monotone increase)

$$\phi(m) = \max\{1 - m, 0\}$$

とすると,

- ϕ は $m = 0$ で微分可能
- $\phi'(0) = -1 < 0$

だから, 定理 3.5 の仮定を満たす. よって ϕ は C.C. loss. また,

$$C_\eta(\alpha) = \eta \max\{1 - \alpha, 0\} + (1 - \eta) \max\{1 + \alpha, 0\}$$
$$= \begin{cases} \eta(1 - \alpha) & \text{if } \alpha \le -1 \\ (1 - 2\eta)\alpha + 1 & \text{if } -1 < \alpha \le 1 \\ (1 - \eta)(1 + \alpha) & \text{if } 1 < \alpha \end{cases}$$

より,

$$\min C_\eta(\alpha) = \begin{cases} 2\eta \quad \text{at} \quad \alpha = -1 & \text{if } 0 \le \eta < \frac{1}{2} \\ 2(1 - \eta) \quad \text{at} \quad \alpha = 1 & \text{if } \frac{1}{2} < \eta \le 1 \\ 1 & \text{if } \eta = \frac{1}{2} \end{cases}$$

Example : Hinge Loss II

よって,

$$
\begin{aligned}
\psi(\theta) &= \phi(0) - H\left(\frac{1+\theta}{2}\right) \\
&= 1 + \begin{cases} 1+\theta & \text{if } -1 < \theta < 0 \\ 1 & \text{if } \theta = 0 \\ 1-\theta & \text{if } 0 < \theta \leq 1 \end{cases} \\
&= \begin{cases} -\theta & \text{if } -1 < \theta < 0 \\ 0 & \text{if } \theta = 0 \\ 1 & \text{if } 0 < \theta < 1 \end{cases} \\
&= |\theta|
\end{aligned}
$$

を得る ($[0, 1]$ 上 strictly monotone increase)

Example : Squared Hinge Loss I

$$\phi(m) = (\max\{1-m, 0\})^2$$

とすると,

- ϕ は $m=0$ で微分可能
- $\phi'(0) = -2 < 0$

だから, 定理 3.5 の仮定を満たす. よって ϕ は C.C. loss. また,

$$\begin{aligned} C_\eta(\alpha) &= \eta(\max\{1-\alpha, 0\})^2 + (1-\eta)(\max\{1+\alpha, 0\})^2, \\ \frac{\mathrm{d}}{\mathrm{d}\alpha} C_\eta(\alpha) &= -2\eta \max\{1-\alpha, 0\} + 2(1-\eta)\max\{1+\alpha, 0\} = 0 \\ &\iff \alpha = 2\eta - 1 \end{aligned}$$

$(\because)$

- $\alpha \le -1$ とすると, $-2\eta(1-\alpha) = 0 \Leftrightarrow \alpha = 1$ となり矛盾
- $\alpha \ge 1$ とすると, $2(1-\eta)(1+\alpha) = 0 \Leftrightarrow \alpha = -1$ となり矛盾
- $-1 < \alpha < 1$ とすると, $-2\eta(1-\alpha) + 2(1-\eta)(1+\alpha) = 0 \Leftrightarrow \alpha = 2\eta - 1$

Example : Squared Hinge Loss II

よって,

$$
\begin{aligned}
H(\eta) &= C_\eta(2\eta - 1) \\
&= \eta(\max\{1 - 2\eta + 1, 0\})^2 + (1 - \eta)(\max\{1 + 2\eta - 1, 0\})^2 \\
&= \eta(4 - 8\eta + \eta^2) + 4(1 - \eta)\eta^2 \\
&= 4\eta(1 - \eta)
\end{aligned}
$$

となるので,

$$
H\left(\frac{1+\theta}{2}\right) = 4 \times \frac{1+\theta}{2} \times \frac{1-\theta}{2} = (1+\theta)(1-\theta) = 1 - \theta^2
$$

より,

$$
\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right) = 1 - 1 + \theta^2 = \theta^2
$$

を得る ($[0, 1]$ 上 strictly monotone increase)

Example : Non Classification Calibrated Loss I

$$\phi(m) = \max\{0, -m\}$$

とすると, ϕ は $m = 0$ で微分不可能.

$$\begin{aligned} C_\eta(\alpha) &= \eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \\ &= \eta \max\{-\alpha, 0\} + (1-\eta)\max\{\alpha, 0\} \\ &= \begin{cases} -\eta\alpha & \text{if } \alpha \leq 0 \\ (1-\eta)\alpha & \text{if } \alpha \geq 0 \end{cases} \end{aligned}$$

なので, $\forall\eta \in [0,1]$, $C_\eta(\alpha) \geq 0$ だから, $\alpha = 0$ で最小値 $C_\eta(0) = 0$ をとる.

よって, $H^-(\eta) = H(\eta) = 0$ となり, ϕ は C.C. loss ではない. また,

$$\psi_0(\theta) = \psi(\theta) = 0$$

となる (constant function).

Example : Non Classification Calibrated Loss II

このとき, $\psi(R_{err}(f) - R^*_{err}) = 0 \le R_\phi(f) - R^*_\phi$ であり, 定理 3.5 の 3 が成り立たないので,

$$R_\phi(f) \to R^*_\phi \implies R_{err}(f) \to R^*_{err}$$

は保証されない.

実際, $R^*_\phi = \inf_g R_\phi(g) = \inf_g \mathbb{E}[\phi(Yg(X))] = 0$ であり, $f(x) = c$ なる constant function に対して,

$$R_\phi(f) = \Pr(Y = +1)\phi(c) + \Pr(Y = -1)\phi(-c) = \Pr(Y = -\mathrm{sign}(c))|c|$$

より, $R_\phi(f) \to R^*_\phi$ as $c \to 0$ が成立. 一方,

$$R_{err}(f) = \Pr(Y = -\mathrm{sign}(c))$$

より, $R_{err}(f) \to \Pr(Y = -1) \neq R^*_{err}$ as $c \searrow 0$ となる.

判別適合的損失

判別適合性定理: 一般のマージン損失

convex とは限らない margin loss に関する性質

Theorem 4 (3.6 C.C. Theorem)

次の *1, 2* は同値

1. ϕ*-margin loss* が *classification calibrated*
2. $\forall\{f_i\}$ *: measurable functions,* $\forall P$ *: distribution on* $\mathcal{X} \times \{\pm 1\}$ に対して

$$R_\phi(f_i) \to R_\phi^* \quad as \quad i \to \infty \quad \Longrightarrow \quad R_{err}(f_i) \to R_{err}^* \quad as \quad i \to \infty$$

Lemma 5 (3.7)

ϕ*-margin loss* から定義される $H(\eta), H^-(\eta), \psi(\theta)$ に対して以下が成立*.*

1. $H(\eta)$ と $H^-(\eta)$ は $[\frac{1}{2}, 1]$ 上 *concave* かつ $(\frac{1}{2}, 1]$ 上 *continuous*
2. ϕ が *C.C. loss* のとき*,* $\forall\theta \in (0, 1]$*,* $\psi(\theta) > 0$

Classification Calibration Theorem II

(proof of Theorem 3.6) $\underline{1 \Rightarrow 2}$

ϕ を C.C. loss とする. このとき, $\psi(\theta)$ は,

- convex (by definition)
- $[-1,1]$ 上偶関数かつ $\psi(0)=0$ かつ $(0,1]$ 上連続 (by Lemma 3.3)
- $\forall\theta \in (0,1]$, $\psi(\theta) > 0$ (by Lemma 3.7)

を満たす. このような ψ は $[0,1]$ 上狭義単調増加

$(\because)$ $0 \le \theta_1 < \theta_2 \le 1$ として, $0 \le \alpha 1$ に対して $\theta_1 = \alpha\theta_2$ とおくと,

$$\psi(\theta_1) \le (1-\alpha)\psi(0) + \alpha\psi(\theta_2) = \alpha\psi(\theta_2) < \psi(\theta_2)$$

となる. ここで, 最初の不等号は凸性, 真中の等号は $\psi(0)=0$ であること, 最後の不等号は $\alpha < 1$ であることをそれぞれ用いた.

(proof of Theorem 3.6) $\underline{1 \Rightarrow 2}$ つづき

よって, 正数列 $\{\theta_i\} \subset [0,1]$ に対して,

$$\psi(\theta_i) \to 0 \implies \theta_i \to 0$$

が成立.

定理 3.4 より,

$$R_\phi(f_i) \to R_\phi^* \implies \psi(R_{err}(f_i) - R_{err}^*) \to 0$$

が成立つから, 上と合わせると,

$$R_\phi(f_i) \to R_\phi^* \implies R_{err}(f_i) \to R_{err}^*$$

を得る.

Classification Calibration Theorem III

(proof of Theorem 3.6) 1 でない $\Rightarrow$ 2 でない

ϕ が C.C. でないとする. このとき, $\exists\eta \neq \frac{1}{2}$, $\exists\{\alpha_i\}$ s.t.

$$\alpha_i(2\eta - 1) \leq 0, \quad C_\eta(\alpha_i) \to H(\eta)$$

が成立つ.

Remark

$\forall\eta \neq \frac{1}{2}$ で $H^-(\eta) > H(\eta)$ となるのが C.C. loss の定義であるが, 上の条件はある $\eta \neq \frac{1}{2}$ で $H^-(\eta) = H(\eta)$ となることを意味する.

ここで, $x_0 \in \mathcal{X}$ に対して,

$$\Pr(X = x_0) = 1, \quad \Pr(Y = +1|X = x_0) = \eta$$

なる確率分布を考える. また, 関数列 $\{f_i\}$ は constant function $f_i(x) = \alpha_i$, $\forall x \in \mathcal{X}$ からなるとする.

(proof of Theorem 3.6) 1 でない ⇒ 2 でないつづき

このとき, $\eta \neq \frac{1}{2}$, $\alpha_i(2\eta - 1) \leq 0$ より, 以下が成立:

$$\lim_{i\to\infty} R_{err}(f_i) = \lim_{i\to\infty} \mathbb{E}[\mathbf{1}_{\mathrm{sign}(f_i(X))\neq Y}]$$
$$> 1 - \mathbb{E}_X\left[\max_{y\in\{\pm 1\}} \Pr(Y=y|X)\right] = R^*_{err}.$$

(∵) $\mathbb{E}_X\left[\max_{y\in\{\pm 1\}} \Pr(Y=y|X)\right] = \max\{\eta, 1-\eta\}$ と書ける. 一方,

$$\mathbb{E}[\mathbf{1}_{\mathrm{sign}(f_i(X))\neq Y}] = \mathbb{E}[\mathbf{1}_{\mathrm{sign}(\alpha_i)\neq Y}]$$

だから, $\eta > \frac{1}{2}$ とすると, 右辺 =$1-\eta$, 左辺 =$\mathbb{E}[\mathbf{1}_{Y=+1}] = \eta$ より, 左辺 $>$ 右辺となる. $\eta < \frac{1}{2}$ の場合も同様.

他方, $C_\eta(\alpha_i) \to H(\eta)$ だから, $R_\phi(g) = \mathbb{E}_X[C_\eta(g(X))]$ より $\lim_{i\to\infty} R_\phi(f_i) = R^*_\phi$ が成立. これは, 2 の不成立を意味する. □

Example : Ramp Loss I

$$\phi(m) = \min\{1, \max\{1-m, 0\}\}$$

は non-convex な margin loss なので, 定理 3.5 は使えない.

$$C_\eta(\alpha) = \begin{cases} \eta & \text{if } \alpha \le -1 \\ (1-\eta)\alpha + 1 & \text{if } -1 < \alpha \le 0 \\ -\eta\alpha + 1 & \text{if } 0 < \alpha < 1 \\ 1-\eta & \text{if } 1 \le \alpha \end{cases}$$

$$H(\eta) = \begin{cases} \eta & \text{if } 0 \le \eta \le \frac{1}{2} \\ 1-\eta & \text{if } \frac{1}{2} < \eta < 1 \end{cases}$$

$$H^-(\eta) = \begin{cases} 1-\eta & \text{if } 0 \le \eta \le \frac{1}{2} \\ \eta & \text{if } \frac{1}{2} < \eta < 1 \end{cases}$$

なので, $\theta \in [0,1]$ に対して,

$$\psi_0(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right) = \frac{1+\theta}{2} - \left(1 - \frac{1+\theta}{2}\right) = \theta$$

Example : Ramp Loss II

$\psi_0(\theta) > 0$ $(\theta > 0)$ であるから, 特に $\eta > \frac{1}{2}$ に対して,

$$H^-(\eta) > H(\eta)$$

となるので, ϕ は C.C. loss である.

$[-1, 1]$ 上では, $\psi_0(\theta) = |\theta|$ であり, 特に $\psi(\theta) = |\theta|$ であるから,

$$\psi(R_{err}(f) - R^*_{err}) = R_{err}(f) - R^*_{err} \leq R_\phi(f) - R^*_\phi$$

が成立 (hinge loss の評価と同じ)

Remark

non-convex な margin loss は, 0 で微分不可能でも C.C. となる場合がある.

References

[1] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[3] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[4] 金森敬文. 統計的学習理論 (機械学習プロフェッショナルシリーズ). 講談社, 2015.