統計解析特論 講義メモ $\boxed{\mathbf{4}}$ 判別分析

学習データ：$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.

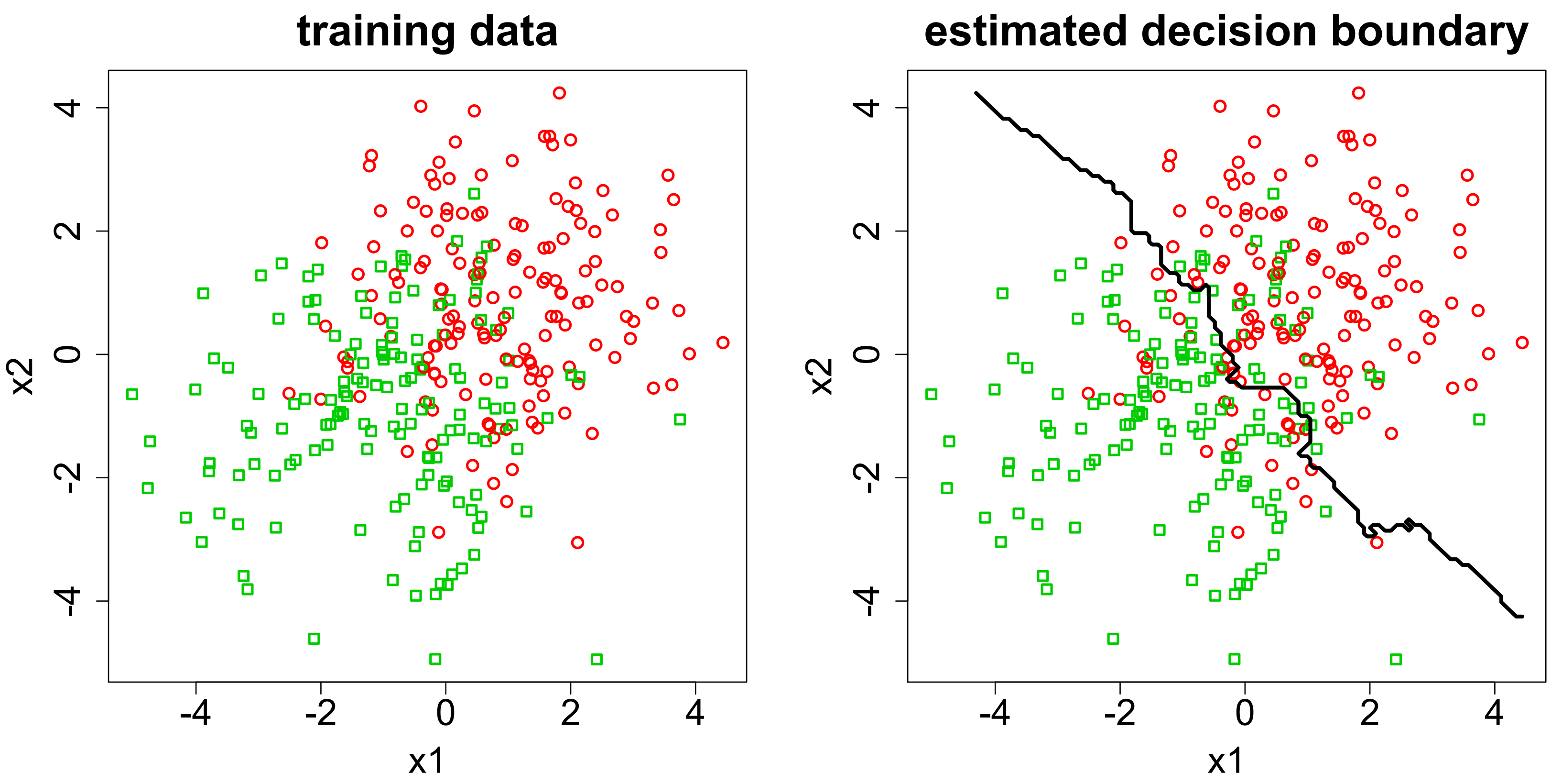
- ２値判別：$y \in \{+1, -1\}$
  - 迷惑メールの判別，デジカメの顔検出．特定の疾患の診断.
- 多値判別： $y \in \mathcal{Y} = \{1, 2, \ldots, G\}$
  - 文字認識，自然言語処理.

目的

学習データと同じ分布にしたがう新たな入力 $\boldsymbol{x}$ に対して，
ラベル y を予測する.

２値判別

- 学習データ $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.
  - 入力：$\boldsymbol{x} \in \mathcal{X}$
  - 出力，ラベル：$y \in \{+1, -1\}$



予測の方法

- 学習データから

$$\underline{\text{仮説}}\quad h : \mathcal{X} \to \{+1, -1\}\quad \text{を推定.}$$

  仮説：判別器 **(classifier)** ともいう.

- $\boldsymbol{x}$ のラベルを $h(\boldsymbol{x})$ で予測.

入力 $\boldsymbol{x}$ のラベル y を予測する手順

1. データ $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ に対して
できるだけ $h(\boldsymbol{x}_i) = y_i$ となるような仮説を学習（推定）.

2. 新たな入力 $\boldsymbol{x}$ のラベルを $h(\boldsymbol{x}) \in \{+1, -1\}$ で予測.

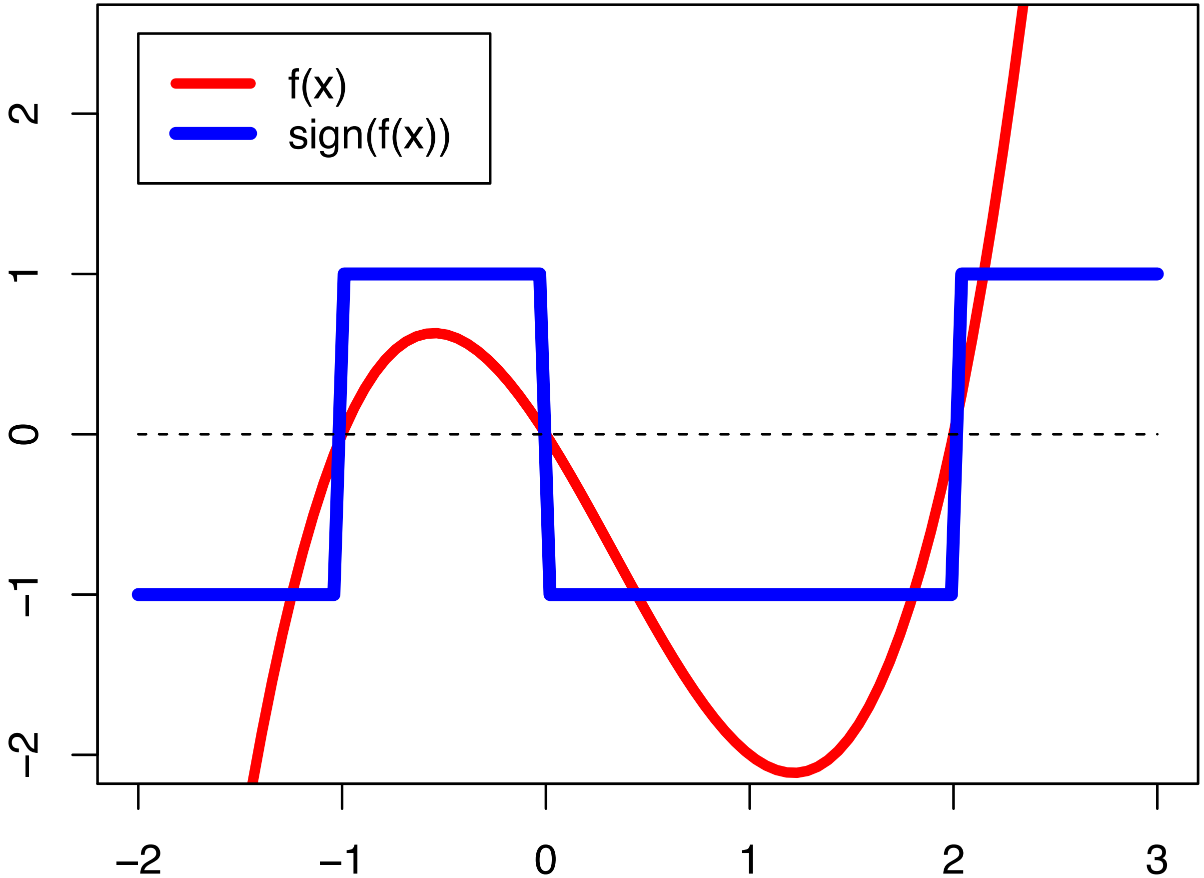
判別器のモデリング

- 判別関数 $f : \mathcal{X} \longrightarrow \mathbb{R}$
- 判別関数 $f(\boldsymbol{x})$ から仮説 $h(\boldsymbol{x})$ を構成

$$h(\boldsymbol{x}) = \mathrm{sign}(f(\boldsymbol{x}))$$

- 符号関数：$\mathrm{sign}(z) = +1\,(z > 0),\ -1\,(z < 0),\ 0\,(z = 0)$
  $h(\boldsymbol{x}) = 0$ のとき：「判別不能」,「± 1 のどちらかに適当に割りふる」など.

判別関数と仮説のプロット

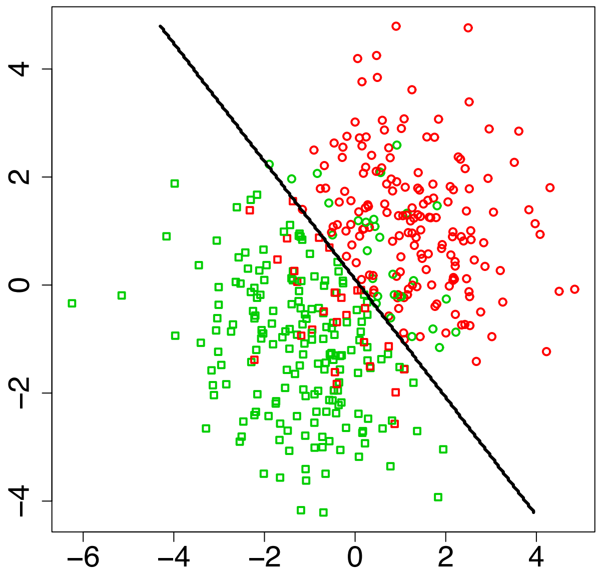


線形判別器

基底関数 $\phi : \mathcal{X} \to \mathbb{R}^d$ を定める． $\phi(x) = (\phi_1(x), \dots, \phi_d(x))$.

$$\text{判別関数の集合 } \mathcal{F} = \{\, f(x) = \phi(x)^T w + b \mid b \in \mathbb{R},\ w \in \mathbb{R}^d \,\}$$
$$\text{線形判別器の集合 } \mathcal{H} = \{\, \mathrm{sign}(f(x)) \mid f(x) \in \mathcal{F} \,\}$$

データから，適切なパラメータ
w, b を推定する



0-1 損失

データ (x, y) を判別器 $h \in \mathcal{H}$ で学習

- $h(x) = y \;\Rightarrow$ 正しい：**loss =** 0
- $h(x) \neq y \;\Rightarrow$ 間違い：**loss =** 1

定義関数を使って表現：

$$\textbf{0-1 損失：} \quad \mathbf{1}[h(x) \neq y], \qquad \mathbf{1}[A] = \begin{cases} 1, & A \text{が真}, \\ 0, & A \text{が偽}. \end{cases}$$

補足：非対称損失を用いることもある.
データ (x, y) に対して

- $h(x) = y \ \Rightarrow$ 正しい：損失 0
- $h(x) = 1 \ \& \ y = 0 \ \Rightarrow$ ：損失 1
- $h(x) = 0 \ \& \ y = 1 \ \Rightarrow$ ：損失 10

例： 健康なら $y = 0$, 病気なら $y = 1$.

講義内容は，非対称損失に一般化可能.

学習誤差・予測誤差

仮説 $h : \mathcal{X} \to \{+1, -1\}$ の予測誤差と学習誤差.

- 学習誤差：学習データ $(x_1, y_1), \ldots, (x_n, y_n)$ に対して

$$\widehat{\mathrm{e}}(h) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\,h(x_i) \neq y_i\,]$$

- 予測誤差：分布 $(x, y) \sim P$ に対して

$$\mathrm{e}(h) \;=\; \Pr(h(x) \neq y) \;=\; E[\mathbf{1}[h(x) \neq y]]$$

[確率密度 $p(x, y)$ から定まる**]**

仮説の学習

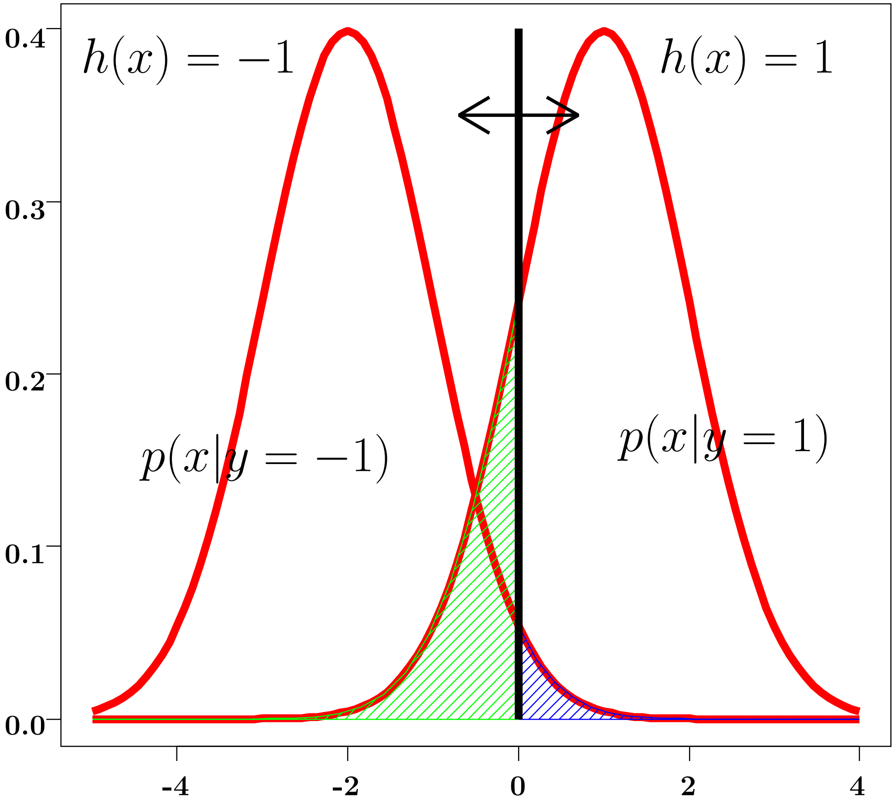
- $h(x)$ の予測精度が高い $\iff e(h)$ が小さい
  $\mathrm{e}(h)$ を小さくする仮説 h を選びたい.
  - 確率分布 P は未知なので $\mathrm{e}(h)$ の値は分からない
  - 案：代わりに $\widehat{\mathrm{e}}(h)$ を小さくする h を選ぶ

根拠：各 h に対して $\widehat{\mathrm{e}}(h) \xrightarrow{p} \mathrm{e}(h),\ n \to \infty,$ （大数の法則）

$$\Longrightarrow\ [\,e(h)\text{ を最小にする } h\,]\ \simeq\ [\,\widehat{e}(h)\text{ を最小にする } h\,]$$

例：予測誤差の計算

$x \in \mathbb{R},\ \ y \in \{1, -1\}, \quad p(x, y) = p(x|y)p(y)$ とする



$$
\begin{aligned}
e(h) &= \Pr(h(X) = -1, Y = +1) \\
&\qquad + \Pr(h(X) = +1, Y = -1) \\
&= \Pr(h(X) = -1 \,|\, Y = 1) \Pr(Y = 1) \\
&\qquad + \Pr(h(X) = 1 \,|\, Y = -1) \Pr(Y = -1) \\
&= \ [\text{■の面積}] \times \Pr(Y = 1) \\
&\qquad + [\text{■の面積}] \times \Pr(Y = -1)
\end{aligned}
$$

例 **1.** 線形判別器 $h(x) = \mathrm{sign}(w^T\phi(x_i) + b)$ に対して

$$\widehat{e}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\,\mathrm{sign}(w^T\phi(x_i) + b) \neq y_i\,] \;\to\; \min_{w,b}$$

- $\widehat{\mathrm{e}}(h)$ の最小化は数値的に困難

- 計算しやすい他の方法が提案されている **(**代替損失の理論**)**

講義の予定：

1. 分布 P のもとで最適な仮説：ベイズルール，ベイズ誤差.

2. 学習誤差最小による学習法の予測誤差

3. 補足：代替損失による学習

ベイズ規則，ベイズ誤差

- ベイズルール **(Bayes rule)**：予測誤差を最小にする仮説

$$\mathrm{e}(h_0) = \inf_{h:\text{任意の仮説}} \mathrm{e}(h) \quad \longrightarrow \quad h_0 \text{ はベイズルール}$$

- $\mathrm{e}(h_0)$ をベイズ誤差（**Bayes error**）という：予測精度の下限
- 学習の目標：学習データからベイズルール h_0 を推定.

Theorem 1.
2値判別 $(y \in \{\pm 1\})$ でデータの分布を $p(x,y) = \Pr(Y = y|x)p(x)$ とする.

ベイズルール $h_0(x) = \begin{cases} +1, & \Pr(Y=+1|x) \ \geq \ \Pr(Y=-1|x), \\ -1, & \Pr(Y=-1|x) \ > \ \Pr(Y=+1|x). \end{cases}$

ベイズ誤差 $e(h_0) = \displaystyle\int \min_y \{p(x,y)\} dx$

出現しやすいラベルを予測ラベルとするのが最適.

note: $\Pr(Y=+1|x) \ \geq \ \Pr(Y=-1|x)$

$\iff \ p(x,+1) \ \geq \ p(x,-1)$

補足：$\mathrm{Pr}(Y = y|x)p(x)$, $x \in \mathbb{R}^d$, $y \in \{+1, -1\}$ について.

$$p(x, y) = p(x|Y = y)\mathrm{Pr}(Y = y) = \mathrm{Pr}(Y = y|x)p(x)$$

Y に関する確率を表すときは Pr と書く.

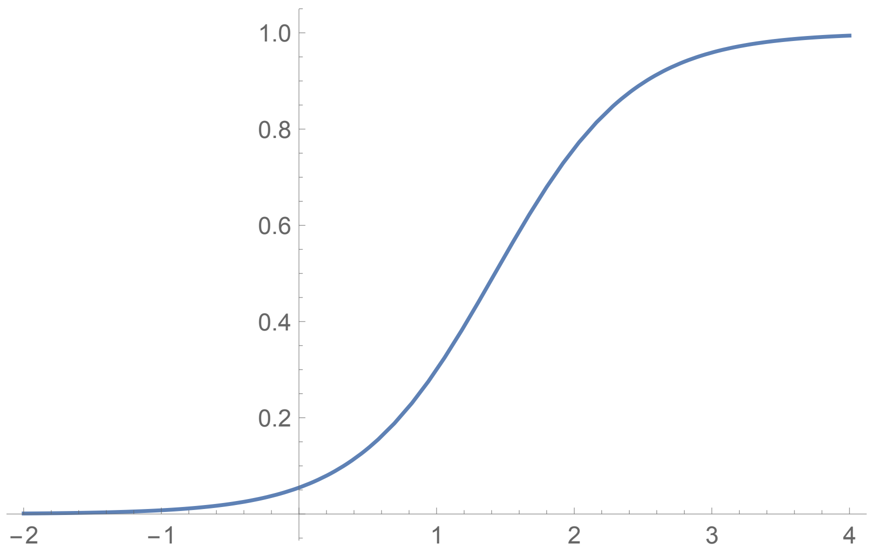
例：$x \in \mathbb{R}$, $y \in \{+1, -1\}$ とする.

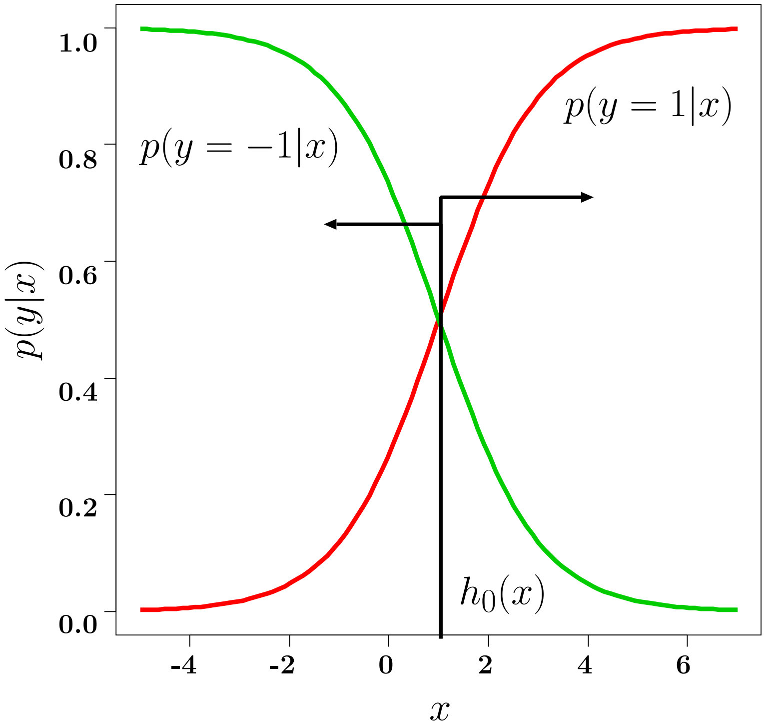
$$\Pr(Y=+1) = 0.3, \quad p(x|Y=+1) : N(2, 1^2) \text{の密度}$$

$$\Pr(Y=-1) = 0.7, \quad p(x|Y=-1) : N(0, 1^2) \text{の密度}$$

$$\begin{aligned}
&\Pr(Y=+1|x) \\
&= \frac{p(x|Y=+1)\Pr(Y=+1)}{\sum_{y=\pm 1} p(x|Y=y)\Pr(Y=y)} \\
&= \frac{0.3\frac{1}{\sqrt{2\pi}}e^{-(x-2)^2/2}}{0.3\frac{1}{\sqrt{2\pi}}e^{-(x-2)^2/2} + 0.7\frac{1}{\sqrt{2\pi}}e^{-x^2/2}} = \frac{1}{1+\frac{7}{3}\exp\{-2x+2\}}
\end{aligned}$$

$\Pr(Y = +1|x)$ のグラフ (x の関数としてプロット)





Theorem 1 の証明. ベイズルールの導出.

$$\begin{aligned}
e(h) &= \int \sum_{y=\pm 1} \mathbf{1}[h(x) \neq y] p(x, y) dx \\
&\geq \int \min_{y'} p(x, y') dx \\
&= \int p(x, -h_0(x)) dx \\
&= \int \sum_{y=\pm 1} \mathbf{1}[h_0(x) \neq y] p(x, y) dx \\
&= e(h_0)
\end{aligned}$$

多値判別のベイズルール

- $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} = \{1, \ldots, G\}$
- 確率分布：$p(x, y) = \Pr(Y = y|x)p(x)$

$$\text{ベイズルール：} \quad h_0(x) = \arg\max_{y \in \mathcal{Y}} \ \Pr(Y = y|x)$$

$$\text{ベイズ誤差：} \quad e(h_0) = \int (1 - \max_y p(x, y))dx$$

note: 多値のときは**(2**値の拡張として**)** 以下が成立：

$$\sum_{y=\pm 1} \mathbf{1}[h(x) \neq y]p(x, y) = 1 - p(x, h(x)) \geq 1 - \max_y p(x, y).$$

推定された仮説の精度

目標：推定された仮説の予測誤差を評価.

- データ $(x_1, y_1), \ldots, (x_n, y_n) \sim_{i.i.d.} P$,　ベイズルール $h_0(x)$.

  2 値判別を考える．多値でも同様.

- 推定方法：
  - 仮説集合：$\mathcal{H} = \{h_1, \ldots, h_M\}, \quad |\mathcal{H}| = M < \infty$.
  - 学習誤差の最小化 **(empirical risk minimization, ERM)**

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \ \widehat{e}(h) = [\,\mathcal{H} \text{ のなかで } \widehat{e}(h) \text{ を最小にする仮説}\,]$$

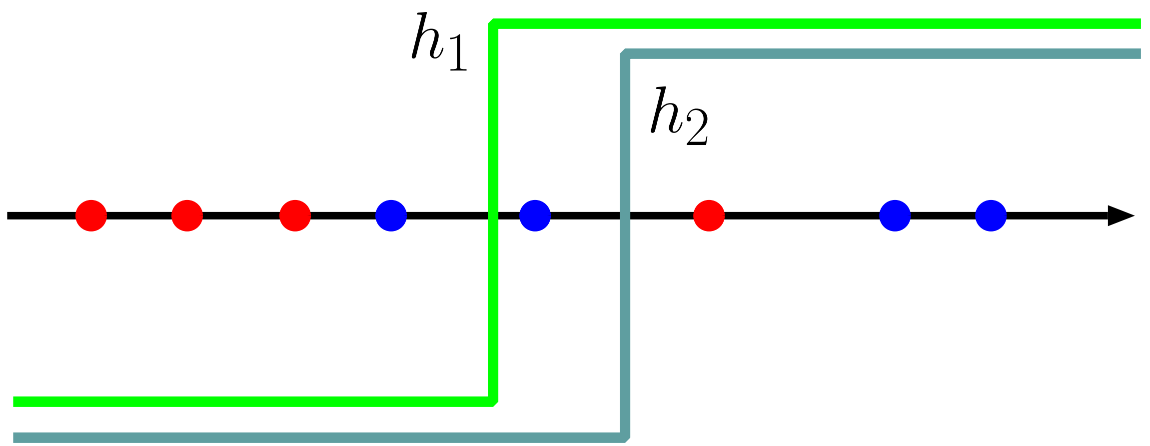
補足：$\widehat{e}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[h(x_i) \neq y_i]$.

- 目標：推定量 $\widehat{h}$ の予測誤差 $e(\widehat{h})$ を評価.

note: $e(\widehat{h})$ はデータに依存するので確率変数.

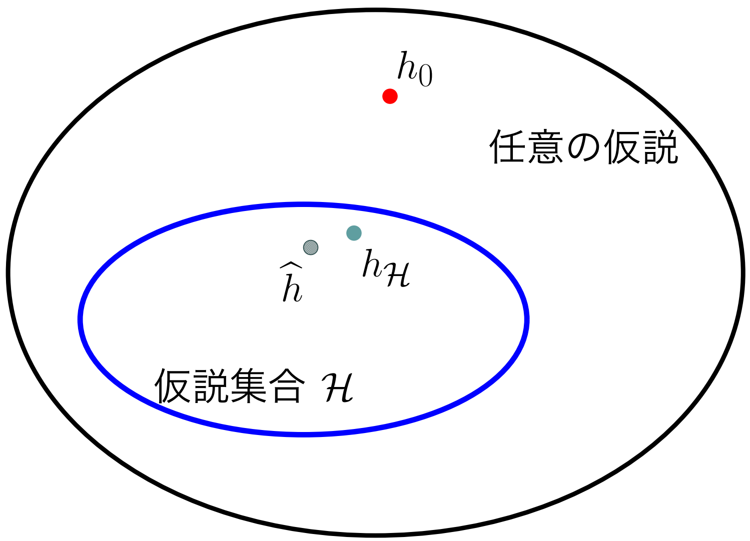
例 **2.**

$$\mathcal{H} = \{h_1,\, h_2\}, \quad \widehat{e}(h_1) = \frac{2}{8}, \quad \widehat{e}(h_2) = \frac{3}{8} \quad \Longrightarrow \quad \widehat{h} = h_1$$



仮説集合 $\mathcal{H} = \{h_1, \ldots, h_M\}$ で **Bayes**ルール h_0 を推定

- 一般に $h_0 \notin \mathcal{H}$



- $h_{\mathcal{H}} := \arg\min_{h \in \mathcal{H}} e(h)$
  $\widehat{h} := \arg\min_{h \in \mathcal{H}} \widehat{e}(h)$

$$e(h_0) \le e(h_{\mathcal{H}}) \le e(\widehat{h}), \quad \widehat{e}(\widehat{h}) \le \widehat{e}(h_{\mathcal{H}})$$

- 一般に $h_0 \neq h_{\mathcal{H}}$

- $e(\widehat{h}) - e(h_0)$ が小さいほどよい
- $e(\widehat{h}) - e(h_0) = \underbrace{\left[e(\widehat{h}) - e(h_{\mathcal{H}})\right]}_{\text{推定誤差}} + \underbrace{\left[e(h_{\mathcal{H}}) - e(h_0)\right]}_{\text{近似誤差}}$

推定誤差 (≥ 0)： 観測データに依存する確率変数.

近似誤差 (≥ 0)： 仮説集合 $\mathcal{H}$ と ベイズルール h_0 から定まる定数

推定誤差について調べる

推定量の誤差評価

$$\text{推定誤差の評価式：}\quad \Pr\left(e(\widehat{h}) - e(h_{\mathcal{H}}) < \varepsilon\right) \;\geq\; 1-\delta$$

が成立するような ε と δ の関係を調べる.
$\varepsilon > 0,\ \delta \in (0,1)$ は小さな値.

$\Pr(\cdots)$： **i.i.d.** データ $(X_1, Y_1), \ldots, (X_n, Y_n)$ に関する確率

- δ が与えられたとき, ε が小さいほどよい.

以下の確率を評価する.

$$\Pr\left(e(\widehat{h}) - e(h_{\mathcal{H}}) \geq \varepsilon\right) \;\leq\; \delta$$

Lemma 1. 以下の不等式が成り立つ：

$$\Pr\left(e(\widehat{h}) - e(h_{\mathcal{H}}) \geq \varepsilon\right)$$
$$\leq \ \Pr\left(e(\widehat{h}) - \widehat{e}(\widehat{h}) \geq \varepsilon/2\right) + \Pr\left(\widehat{e}(h_{\mathcal{H}}) - e(h_{\mathcal{H}}) \geq \varepsilon/2\right)$$

Proof.

$$\begin{aligned} e(\widehat{h}) - e(h_{\mathcal{H}}) &= (e(\widehat{h}) - \widehat{e}(\widehat{h})) + \underline{(\widehat{e}(\widehat{h}) - \widehat{e}(h_{\mathcal{H}}))} + (\widehat{e}(h_{\mathcal{H}}) - e(h_{\mathcal{H}})) \\ &\leq (e(\widehat{h}) - \widehat{e}(\widehat{h})) + (\widehat{e}(h_{\mathcal{H}}) - e(h_{\mathcal{H}})) \end{aligned}$$

したがって $e(\widehat{h}) - e(h_{\mathcal{H}}) \geq \varepsilon$ が成立するとき，以下のどちらかが成立する：

$$(i) \ \ e(\widehat{h}) - \widehat{e}(\widehat{h}) \geq \varepsilon/2, \qquad (ii) \ \ \widehat{e}(h_{\mathcal{H}}) - e(h_{\mathcal{H}}) \geq \varepsilon/2$$

■

ヘフディング不等式

Lemma 2 (Hoeffding's inequality)**.**

$$Z_1, \ldots, Z_n \sim_{i.i.d.} P, \quad 0 \le Z_i \le 1, \quad \mu = E[Z_i]$$

$$\Longrightarrow \ \Pr\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \ge \varepsilon \right) \le \ e^{-2n\varepsilon^2}.$$

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \le -\varepsilon \right) \ \le \ e^{-2n\varepsilon^2}$$

note: 大数の法則から $\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mu$ は分かる.

- ヘフディング不等式から収束スピードが分かる.
- チェビシェフ不等式より良い上界.

- $\widehat{e}(h) - e(h) = \dfrac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\ h(X_i) \neq Y_i\ ] - \boldsymbol{E}[\ \mathbf{1}[h(X) \neq Y]\ ]$

Hoeffding 不等式 で $Z_i = \mathbf{1}[\ h(X_i) \neq Y_i\ ]$ とすると

$$\Pr\left(\widehat{e}(h_{\mathcal{H}}) - e(h_{\mathcal{H}})) \geq \varepsilon/2\right) \leq e^{-2n(\varepsilon/2)^2},$$

$$\begin{aligned}\Pr\left(e(\widehat{h}) - \widehat{e}(\widehat{h}) \geq \varepsilon/2\right) &\leq \Pr\left(\bigcup_{h\in\mathcal{H}} \{e(h) - \widehat{e}(h) \geq \varepsilon/2\}\right) \\ &\leq \sum_{h\in\mathcal{H}} \Pr\left(e(h) - \widehat{e}(h) \geq \varepsilon/2\right) \leq |\mathcal{H}|e^{-n\varepsilon^2/2}.\end{aligned}$$

note: $\mathbf{1}[\ \widehat{h}(X_i) \neq Y_i\ ],\ i = 1, \dots, n$ は独立でない.
$|\mathcal{H}|$：$\mathcal{H}$ の要素数.

$$
\begin{aligned}
\therefore \quad & \Pr\left(e(\widehat{h}) - e(h_{\mathcal{H}}) \geq \varepsilon\right) \\
& \leq \Pr\left(\widehat{e}(\widehat{h}) - e(\widehat{h}) \geq \varepsilon/2\right) + \Pr\left(\widehat{e}(h_{\mathcal{H}}) - e(h_{\mathcal{H}}) \geq \varepsilon/2\right) \\
& \leq (|\mathcal{H}| + 1)e^{-n\varepsilon^2/2} \\
& = \exp\{-n\varepsilon^2/2 + \log(|\mathcal{H}| + 1)\}
\end{aligned}
$$

したがって，$\varepsilon > 0$ に対して

$$
e(\widehat{h}) \;\geq\; \varepsilon + e(h_{\mathcal{H}})
$$

となる確率は $\exp\{-n\varepsilon^2/2 + \log(|\mathcal{H}| + 1)\}$ 以下

$\delta = \exp\{-n\varepsilon^2/2 + \log(|\mathcal{H}| + 1)\}$ として書き直すと・・・

確率 $1 - \delta$ 以上の確率で

$$e(\widehat{h}) < e(h_0) + \underbrace{\left[e(h_{\mathcal{H}}) - e(h_0)\right]}_{\text{近似誤差}} + \underbrace{\sqrt{\frac{2}{n}\log\frac{|\mathcal{H}|+1}{\delta}}}_{\text{推定誤差}} \qquad (1)$$

- 近似誤差： $b_{\mathcal{H}} := e(h_{\mathcal{H}}) - e(h_0)$.
- 推定誤差： $v_{\mathcal{H}}(n, \delta) := \sqrt{\frac{2}{n}\log\frac{|\mathcal{H}|+1}{\delta}}$.

  推定量 $\widehat{h}$ の推定誤差の上界　($|\mathcal{H}|$ は $\mathcal{H}$ の要素数)

(1) を書き直すと：データ数 n のとき $1 - \delta$ 以上の確率で

$$e(\widehat{h}) < e(h_0) + b_{\mathcal{H}} + v_{\mathcal{H}}(n, \delta).$$

モデル選択

複数の仮説集合 $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_K$ について次の包含関係を仮定：

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_K$$

このとき以下が成立

- 近似誤差は減少：$\min_{h \in \mathcal{H}_1} e(h) \geq \min_{h \in \mathcal{H}_2} e(h) \geq \cdots \geq \min_{h \in \mathcal{H}_K} e(h)$ より

$$b_{\mathcal{H}_1} \;\geq\; b_{\mathcal{H}_2} \;\geq\; \cdots \;\geq\; b_{\mathcal{H}_K}$$

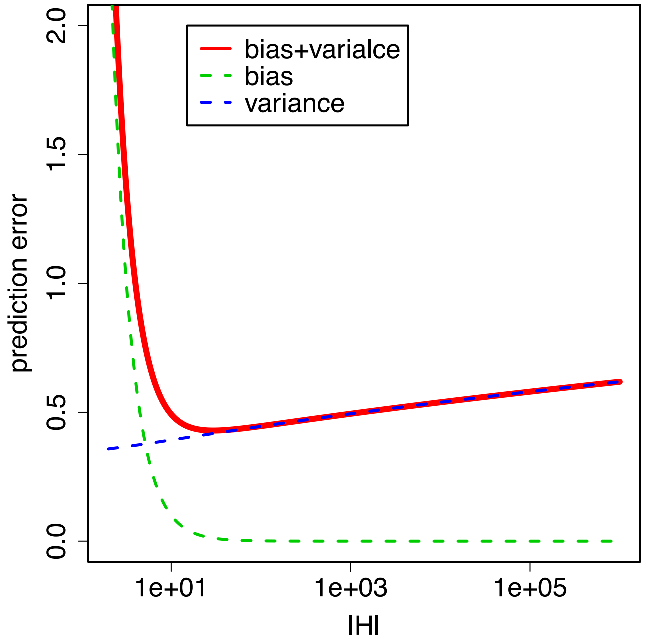
- 推定誤差は増加：データ数 n，確率 δ が一定のとき
  $|\mathcal{H}_1| \leq |\mathcal{H}_2| \leq \cdots \leq |\mathcal{H}_K|$ より

$$v_{\mathcal{H}_1}(n, \delta) \;\leq\; v_{\mathcal{H}_2}(n, \delta) \;\leq\; \cdots \;\leq\; v_{\mathcal{H}_K}(n, \delta)$$

♣ $e(\widehat{h})$ の上界 $e(h_0) + b_{\mathcal{H}} + v_{\mathcal{H}}(n,\delta)$ が小さいほうがよい.

♣ 適切な仮説集合： $\min_{k=1,\dots,K} b_{\mathcal{H}_k} + v_{\mathcal{H}_k}(n,\delta)$ を達成する $\mathcal{H}_k$

- 仮説集合：大
  $\Longrightarrow$ 近似誤差：小，推定誤差：大
- 両者の和を小さくする，ほどよい大きさの仮説集合を用いる.



note: 上の考察は $e(\widehat{h})$ の確率的上界に基づいている.
期待値 $E[e(\widehat{h})]$ で考えても同様の結論が得られる

仮説集合の複雑度

- 仮説集合 $\mathcal{H}$ を用いたときの推定誤差：

$$v_{\mathcal{H}}(n,\delta) = \sqrt{\frac{2}{n}\log\frac{|\mathcal{H}|+1}{\delta}} \leq \sqrt{\frac{2\log(|\mathcal{H}|+1)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

$\longrightarrow$ $\mathcal{H}$ の要素数が大きいほど，推定誤差が大きい.

- $\mathcal{H}$ の要素数が ∞ のときは？
  - 仮説集合の本質的な複雑度を測る：
    VC次元，ラデマッハ複雑度，被覆数.

例 **3.** 基底関数 $\phi_1(x), \ldots, \phi_k(x)$,
$\mathcal{H} = \{\mathrm{sign}(f(x)) \mid f(x) = \sum_{j=1}^{k} c_j \phi_j(x),\, c_j \in \mathbb{R}\}$.

$$v_{\mathcal{H}}(n,\delta) = 4\sqrt{\frac{2k}{n}\log\frac{en}{k}} + \sqrt{\frac{2\log(2/\delta)}{n}}$$

$$b_{\mathcal{H}} = e(h_{\mathcal{H}}) - e(h_0)$$

とすると，ERM は確率 $1-\delta$ 以上で次式を満たす：

$$e(\widehat{h}) < e(h_0) + b_{\mathcal{H}} + v_{\mathcal{H}}(n,\delta).$$

線形判別 $h(x) = \mathrm{sign}(w^T x + b)$, $w \in \mathbb{R}^d, b \in \mathbb{R}$ なら $k = d+1$.

VC theory から得られる．cf.「統計的学習理論」2章

代替損失 (surrogate loss)

仮説 $h(x) = \mathrm{sign}(f(x))$ の **0-1** 損失：

$$\mathbf{1}[h(x) \neq y] = \mathbf{1}[yf(x) \leq 0] \qquad \text{(}f(x) = 0\text{ はひとまず無視).}$$

$yf(x)$ を ((x, y) に対する $f(x)$ の) マージンという.

$$\widehat{e}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\, y_i f(\boldsymbol{x}_i) \leq 0\,],$$

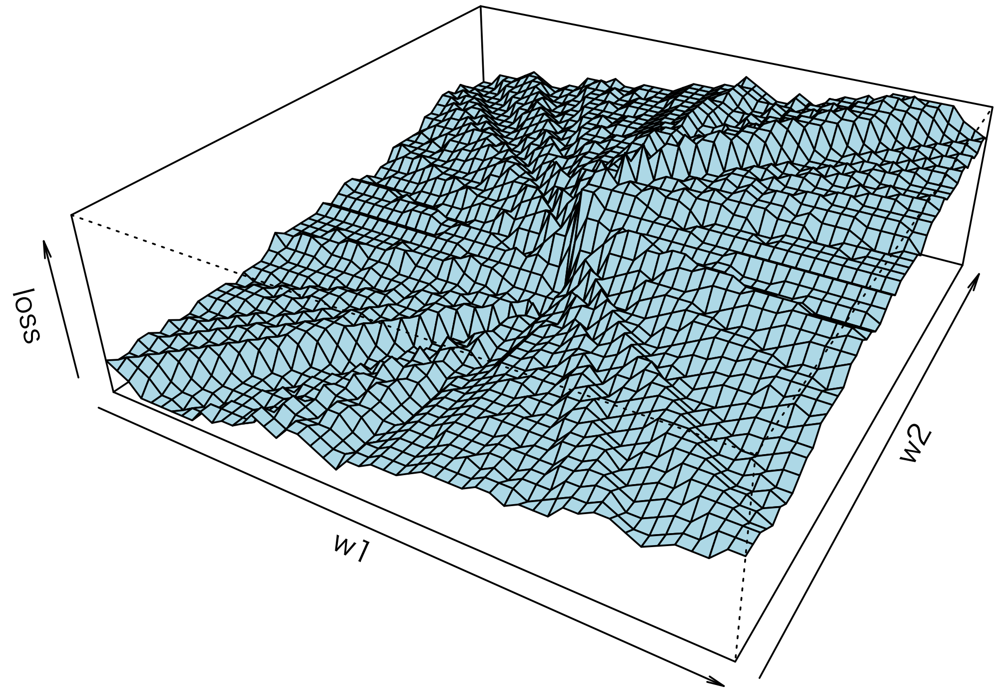
$$e(h) = \Pr(yf(x) \leq 0) = \mathbb{E}[\mathbf{1}[yf(x) \leq 0]]$$

データ上でマージン $yf(x)$ が大きいほどよい.

例：線形仮説 $f(x) = w^T x,\ x, w \in \mathbb{R}^2$

$$\min_w \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y_i(w^T x_i) \le 0] \quad \longrightarrow \text{ 計算困難}$$

0-1 loss



- $\mathbf{1}[yf(x) \leq 0]$ の代わりに，計算しやすい関数 $\ell(yf(x))$ を用いる.

$$\text{例：} \quad \min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i(w^T x_i + b))$$

- $\ell(u)$ が単調減少 $\Longrightarrow$ データ上で大きなマージン $yf(x)$.
- ベイズ・ルールを推定できるか？

例：**(0-1** 損失と共にグラフを描く**)**

サポートベクターマシン： ヒンジ損失

$$\ell(u) = [1-u]_+ \quad ([z]_+ = \max\{z, 0\})$$

ブースティング： 指数損失

$$\ell(u) = e^{-u}$$

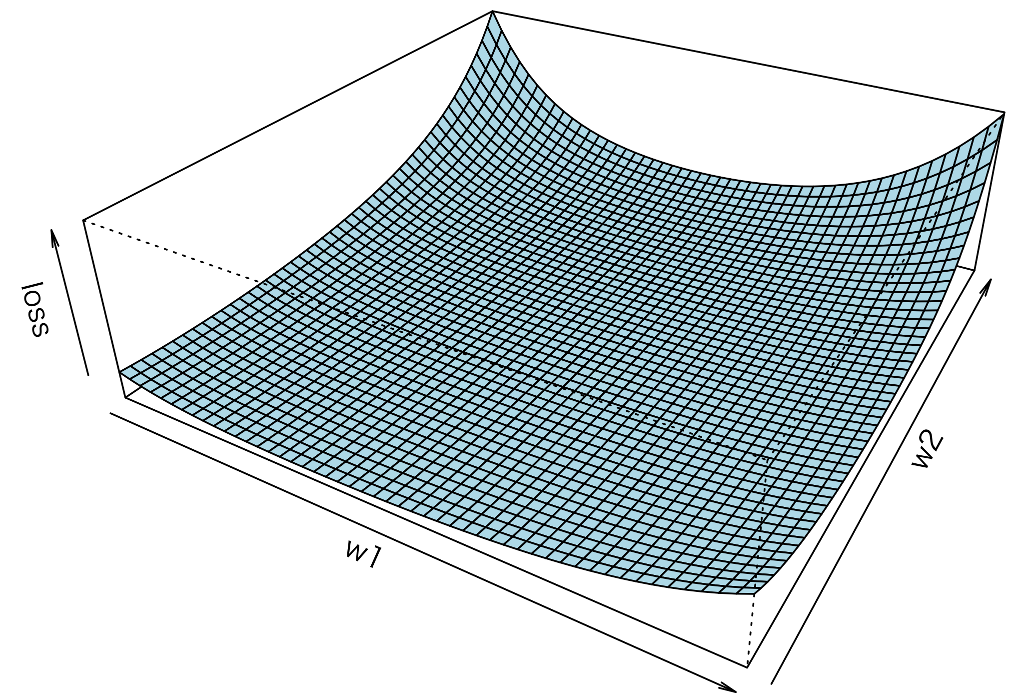
ロジスティック回帰： ロジスティック損失

$$\ell(u) = \log_2(1+e^{-u})$$

例：線形仮説 $f(x) = w^T x,\ x, w \in \mathbb{R}^2$

$$\min_w \frac{1}{n} \sum_{i=1}^{n} e^{-y_i (w^T x_i)} \qquad \longrightarrow \text{ 計算しやすい}$$

exp loss



代替損失の下で最適な判別関数

例．指数損失：

$$\mathbb{E}[\ell(yf(x))] = \int \sum_y e^{-yf(x)} \mathrm{Pr}(Y = y|x) p(x) dx \longrightarrow f \text{ について} \min$$

各 x で $\sum_y e^{-yf(x)} \mathrm{Pr}(Y = y|x)$ を最小にする $f(x)$ が最適：

以下を解く．

$$\frac{\partial}{\partial f(x)} \sum_y e^{-yf(x)} \mathrm{Pr}(Y = y|x) = -\sum_y e^{-yf(x)} y \mathrm{Pr}(Y = y|x) = 0$$

$$
\begin{aligned}
&- e^{-f(x)}\Pr(Y=+1|x) + e^{f(x)}\Pr(Y=-1|x) = 0 \\
\Longrightarrow\ & f(x) = \frac{1}{2}\log\frac{\Pr(Y=+1|x)}{\Pr(Y=-1|x)}
\end{aligned}
$$

$$
\mathrm{sign}(f(x)) = \begin{cases} +1, & \Pr(Y=+1|x) \geq \Pr(Y=-1|x), \\ -1, & \Pr(Y=+1|x) < \Pr(Y=-1|x) \end{cases}
$$

- 期待指数損失を最小にする判別関数：ベイズ・ルールに対応
- ヒンジ損失，ロジスティック損失でも同様の結果.

代替lossの理論

$\ell(u)$ の条件：

(a) $\mathbf{1}[u \le 0] \le \ell(u)$ を満たす.

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(\boldsymbol{x}_i))\text{ が小さい} \implies \widehat{e}(\mathrm{sign}(f))\text{ も小さい}$$

(b) $\ell(u)$ は凸関数**(**定義を示す**)**：凸関数は最適化しやすい．停留点なら最適解.

(c) $u = 0$ で$\ell(u)$は微分可能で $\ell'(0) < 0$.

- 条件 **(a),(b),(c)** を満たす損失を $\ell(u)$ と判別関数の列 $\{f_n(x)\}_{n\in\mathbb{N}}$ に対して以下が成り立つ.

$$\mathbb{E}[\ell(yf_n(x))] \longrightarrow \inf_f \mathbb{E}[\ell(yf(x))], \quad (n\to\infty)$$
$$\Longrightarrow \quad e(\mathrm{sign}(f_n)) \longrightarrow e(h_0), \qquad (n\to\infty)$$

  意味：f_n が $n\to\infty$ で $\ell(u)$ 損失の最小値に収束するとき, $\mathrm{sign}(f_n)$ の予測誤差はベイズ誤差に収束する.

  $\longrightarrow$ $\ell(u)$-損失最小化を正当化

- 例に示した損失 **(**ヒンジ, 指数, ロジスティック**)** は条件を満たす.

- $\ell(u)$ は単調減少でなくてもよい. 例. $\ell(u)=(1-u)^2$.

- **Reference: Bartlett, P. L, et al., Convexity, Classification, and Risk Bounds, Journal of the American Statistical Association, 2006**

- 多値判別では，$\ell'(u) < 0$ のような簡単な条件は得られていない. **ongoing research.**