Convex Formulation for Learning from Positive and Unlabeled Data (ICML2015)

Marthinus Christoffel Du Plessis（東京大学）
Gang Niu (Baidu Inc.)
Masashi Sugiyama（東京大学）

THE UNIVERSITY OF TOKYO

Baidu 百度

概要

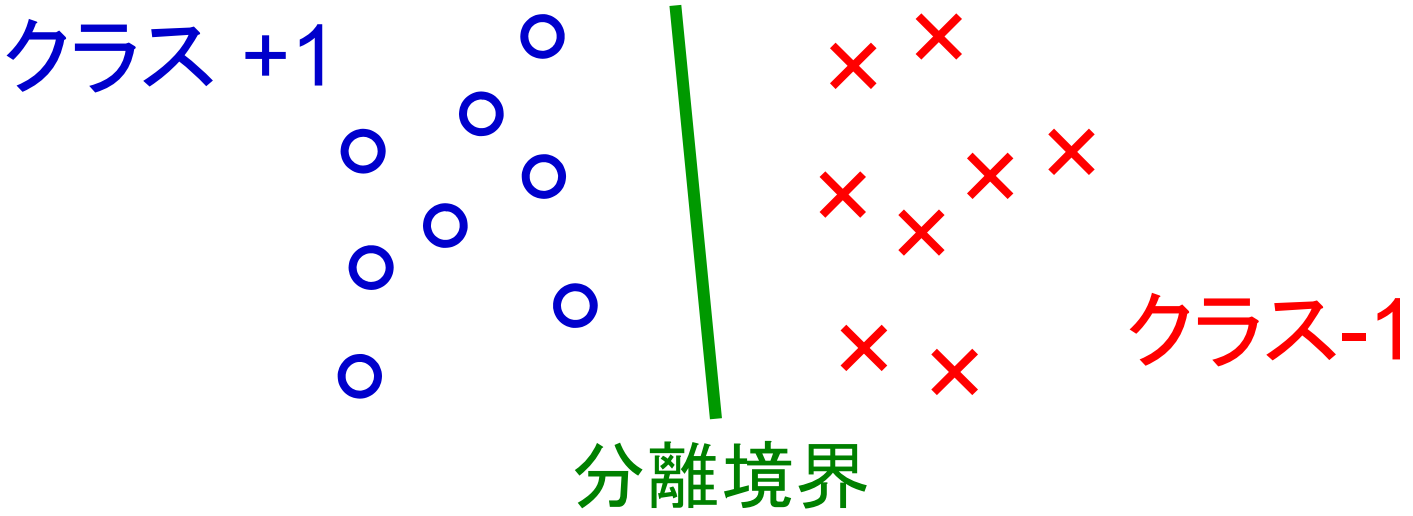
- 分類問題（パターン認識）は，入手できるデータの種類により様々な手法：
  - 教師付き学習：学習精度は良いが，ラベル付けのコストが高い
  - 教師なし学習：ラベル付けのコストは不要だが，学習の信頼性が低い
  - 半教師付き学習：ラベル付けのコストは抑制できるが，学習精度は必ずしも高くない
- より現実的な手法を提案：
  - 正例とラベルなしデータからの分類

発表の流れ

1. 分類問題の分類
2. 正例とラベルなしデータからの分類1
3. 正例とラベルなしデータからの分類2
4. 正例とラベルなしデータからの分類3

2クラスの教師付き分類

- ラベル付きデータ: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
  - 入力 $\boldsymbol{x}$ は d 次元の実ベクトル $\boldsymbol{x} \in \mathbb{R}^d$
  - 出力 y は2値のクラスラベル $y \in \{+1, -1\}$

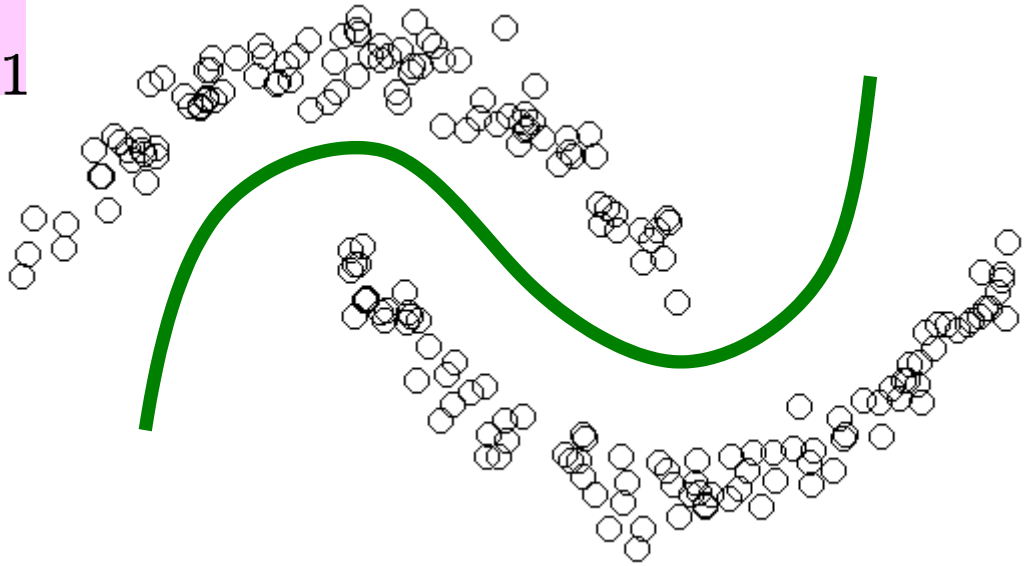


- 大量のラベル付きデータを用いれば，精度良く分類境界が学習できる

教師なし分類

- ラベル付きデータの収集にはコストがかかるため，容易に入手できるラベルなしデータを用いる

$\{\boldsymbol{x}'_i\}_{i=1}^{n'}$

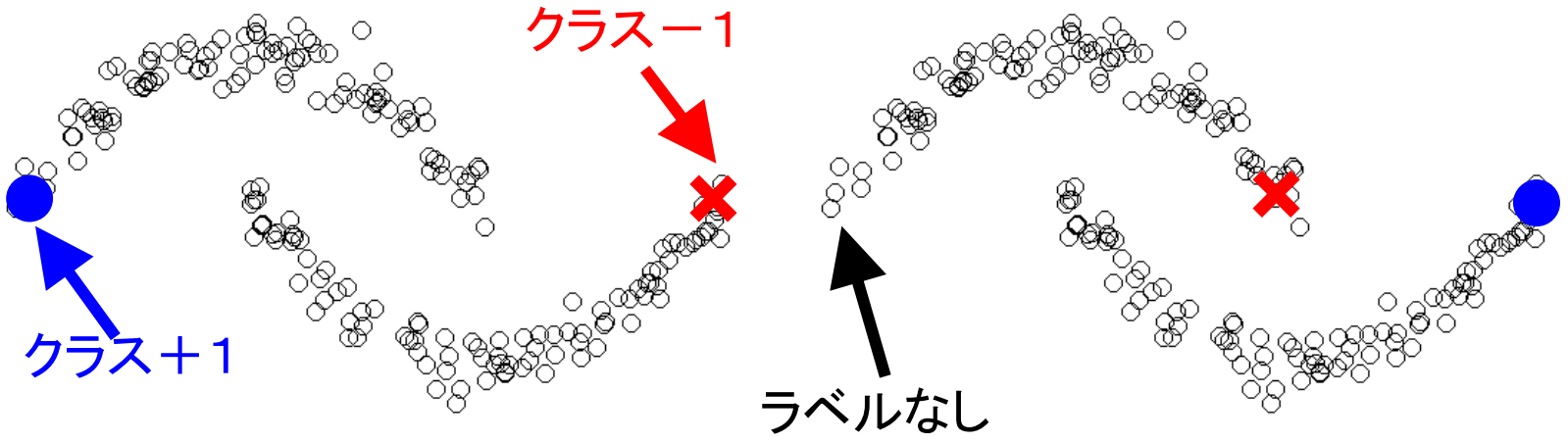


- 教師なし分類はクラスタリングともよぶ
- データがクラス毎にクラスタに分かれていないと，正しく分類できない

半教師付き分類

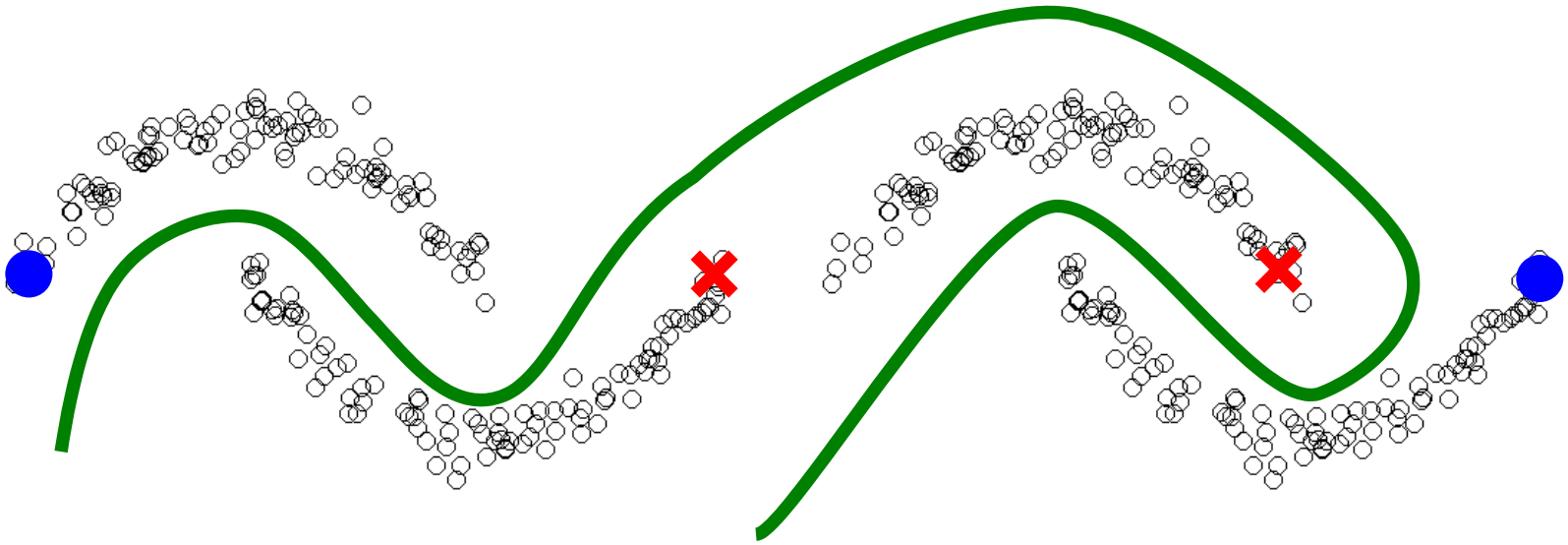
- 少量のラベル付きデータと大量のラベルなしデータを利用する

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \qquad \{\boldsymbol{x}'_i\}_{i=1}^{n'}$$



半教師付き分類(続き)

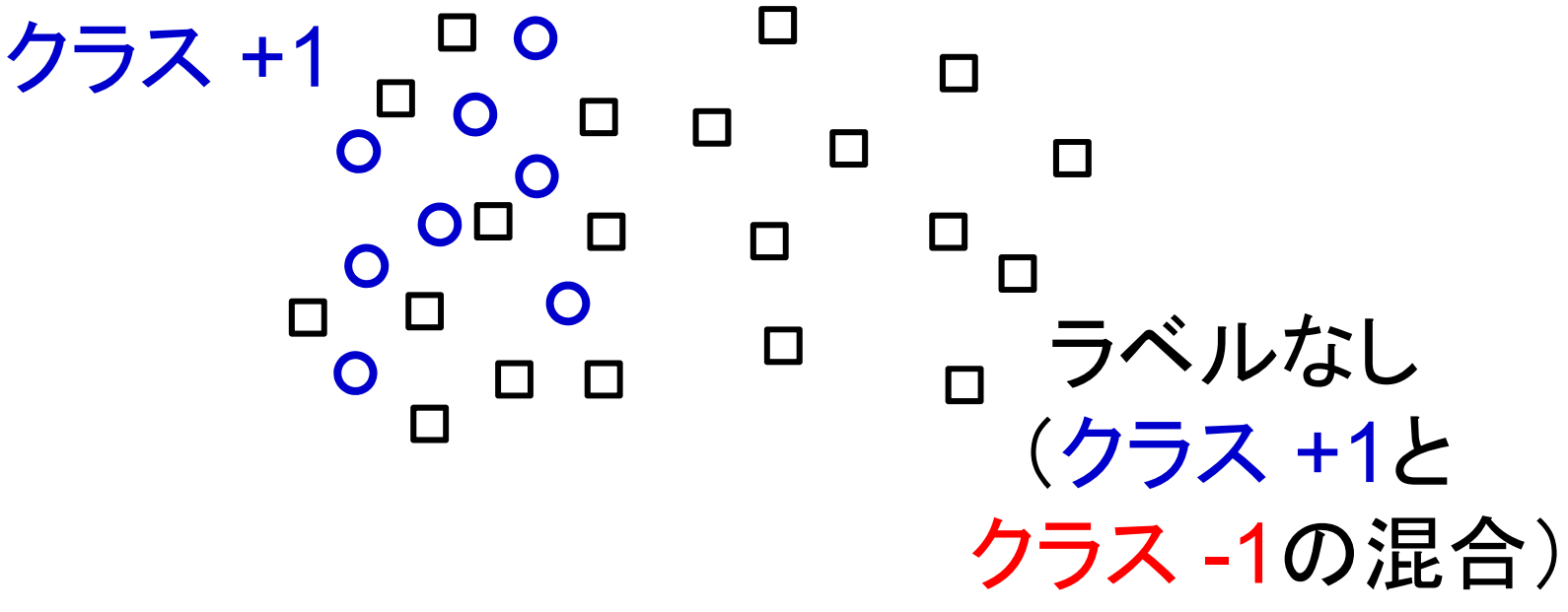
- ラベルなしデータがなすクラスタ構造に従って分類



- 同じクラスタに属するデータが同じラベルを持つとき，うまく分類できる
- そのような仮定が常に成り立つとは限らない

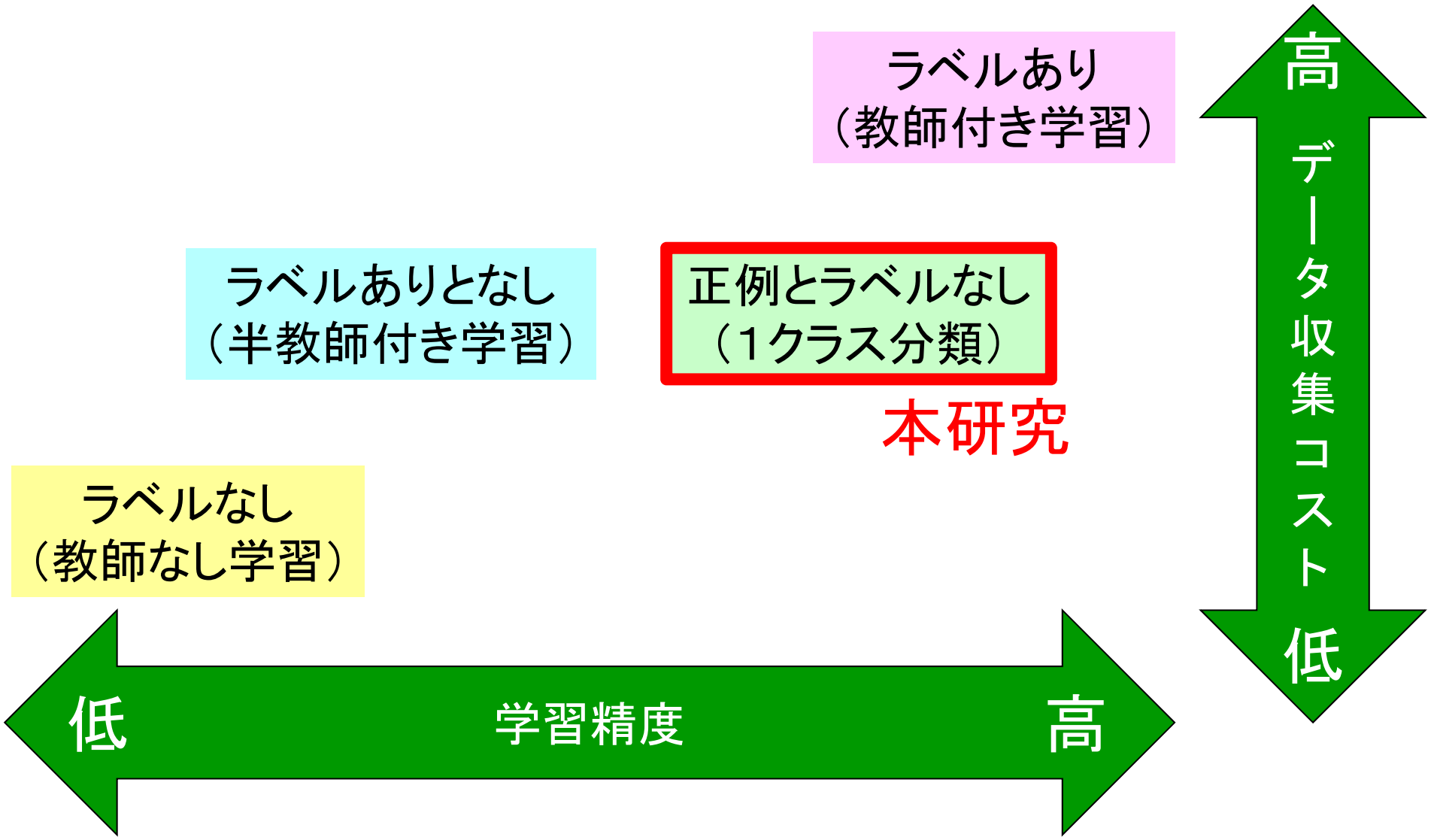
本研究：
正例とラベルなしデータからの分類

- ラベル付きデータの収集にはコストがかかるが，一方のクラス（＋1とする）のラベル付きデータは容易に入手できることがある



分類の分類

- データ収集のコストに合わせて手法を選ぶ



発表の流れ

1. 分類問題の分類
2. 正例とラベルなしデータからの分類1
3. 正例とラベルなしデータからの分類2
4. 正例とラベルなしデータからの分類3

分類器の誤差

Elkan & Noto (KDD2008)

- 分類器 $f(\boldsymbol{x})$ の誤差 $R(f)$ は，
  偽陰性と擬陽性の加重和：

$$R(f) = \pi \int \ell_{0/1}\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

偽陰性（正を負と誤る）

$$+(1-\pi) \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

擬陽性（負を正と誤る）

  - 0/1損失：
$$\ell_{0/1}(t) = \begin{cases} 1 & (t \leq 0) \\ 0 & (t > 0) \end{cases}$$
  - クラス事前確率：$\pi = p(y=+1)$

$$1-\pi = p(y=-1)$$

ラベルなしデータに対する誤差

$$\int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

- 「正例とラベルなしデータからの分類」では負例がないため，ラベルなしデータを負例だと思った場合の誤差を考える:

$$\int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$$

- これを使えば，分類器 $f(\boldsymbol{x})$ の誤差 $R(f)$ は

$$R(f) = 2\pi \int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi$$

と表せる

証明

$$p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$$

$$\int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \pi \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

$$+(1-\pi) \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

$$= \pi\Big(1 - \int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}\Big)$$

$$+(1-\pi) \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

証明（続き）

$$\int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \pi\Big(1-\int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}\Big)$$

$$+(1-\pi)\int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

$$R(f) = \pi\int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

$$+(1-\pi)\int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

$$= \pi\int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$-\pi\Big(1-\int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}\Big)$$

$$= 2\pi\int \ell_{0/1}\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi$$

重み付き分類

$$R(f) = 2\pi \int \ell_{0/1}\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi$$

- これは，正例とラベルなしデータに

$$c_{+1} = \frac{2\pi}{\eta} \quad c_U = \frac{1}{1-\eta}$$

η：正例とラベルなしデータの比率

という重み付けた分類に相当

- 正例とラベルなしデータの重み付き分類によって，正しく分類器が学習できる！

発表の流れ

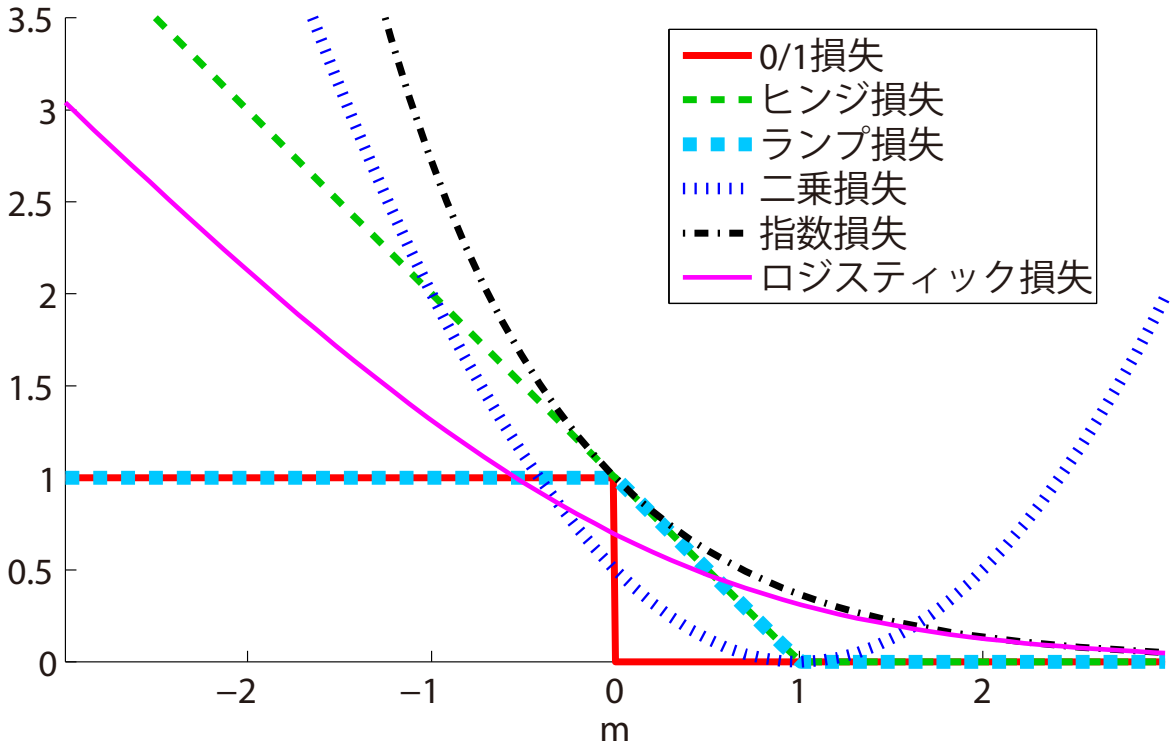
1. 分類問題の分類
2. 正例とラベルなしデータからの分類1
3. 正例とラベルなしデータからの分類2
4. 正例とラベルなしデータからの分類3

代理損失

$$R(f) = 2\pi \int \ell_{0/1}\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi$$

- 0/1損失の最小化はNP困難
- 実際には，最適化しやすい代理損失 $\ell(m)$ を用いる

$$m = yf(\boldsymbol{x})$$



代理損失に対する誤差

$$R(f) = 2\pi \int \ell_{0/1}\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell_{0/1}\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi$$

- 代理損失 ℓ に対する誤差を偽陰性と擬陽性に戻すと，余分な項が出てくる：

$$R_\ell(f) = 2\pi \int \ell\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi$$

$$= \underbrace{\pi \int \ell\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}}_{\text{偽陰性}} + \underbrace{(1-\pi) \int \ell\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}}_{\text{擬陽性}}$$

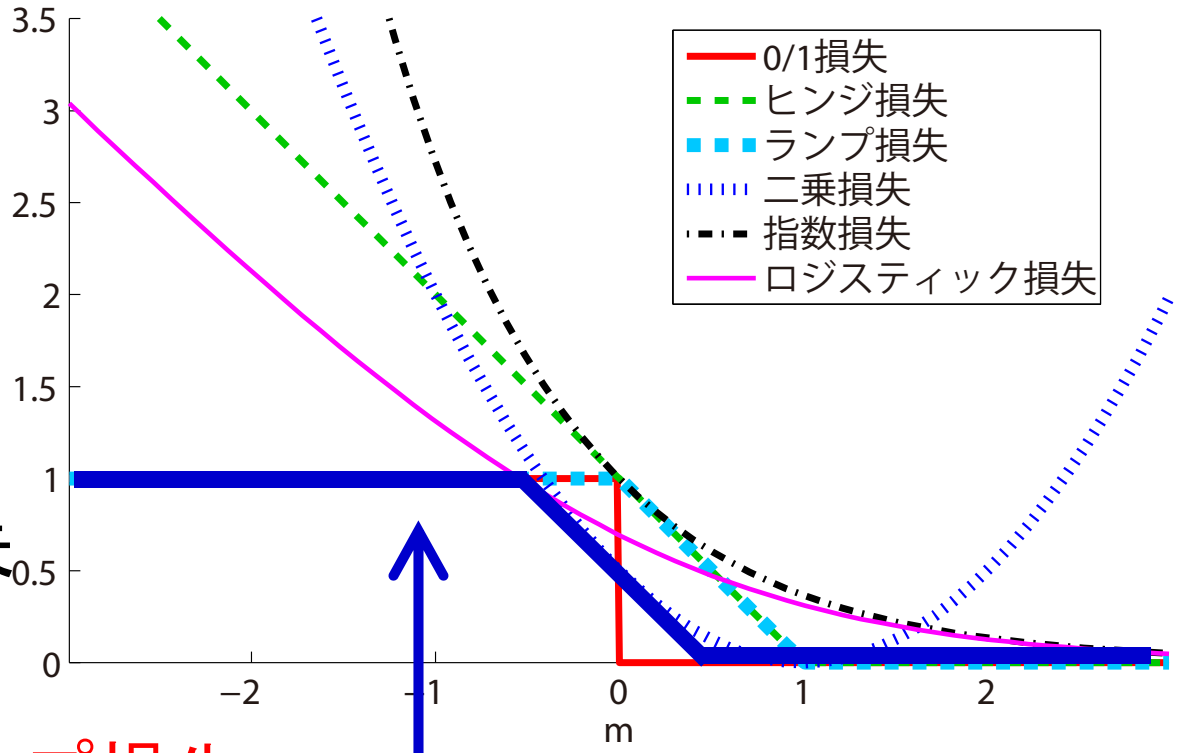
$$\underbrace{+\pi \int \Big[\ell\Big(f(\boldsymbol{x})\Big) + \ell\Big(-f(\boldsymbol{x})\Big)\Big] p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} - \pi}_{\text{余分な項}}$$

- $\ell(m) + \ell(-m) = \mathrm{Const.}$ のとき余分な項が無視できる

代理損失の選択

$$\ell(m) + \ell(-m) = \text{Const.}$$

- NG:
  - ヒンジ損失
  - 二乗損失
  - 指数損失
  - ロジスティック損失
- OK:
  - （少しずらした）ランプ損失

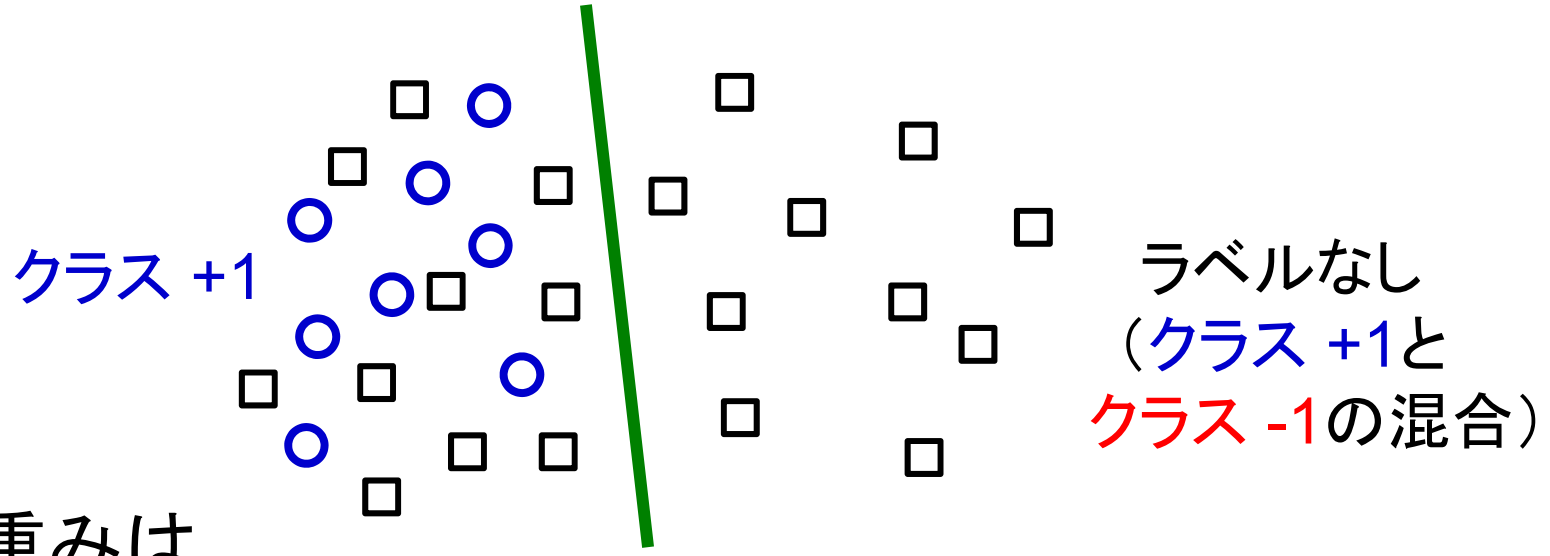


$$\ell(m) = \frac{1}{2}\max\Big(0, \min(2, 1-m)\Big)$$

非凸解法

du Plessis, Niu & Sugiyama (NIPS2014)

- ランプ損失を用いた重み付きロバストサポートベクトルマシンで正例とラベルなしデータを分離（重なりが大きいので普通のSVMはダメ）



- 重みは

$$c_{+1} = \frac{2\pi}{\eta} \qquad c_U = \frac{1}{1-\eta}$$

η : 正例とラベルなしデータの比率

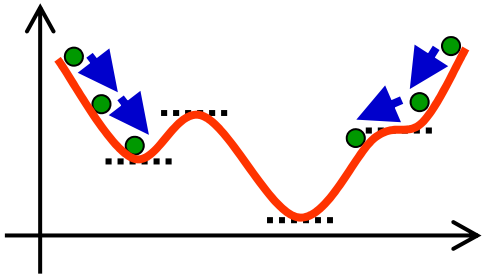
発表の流れ

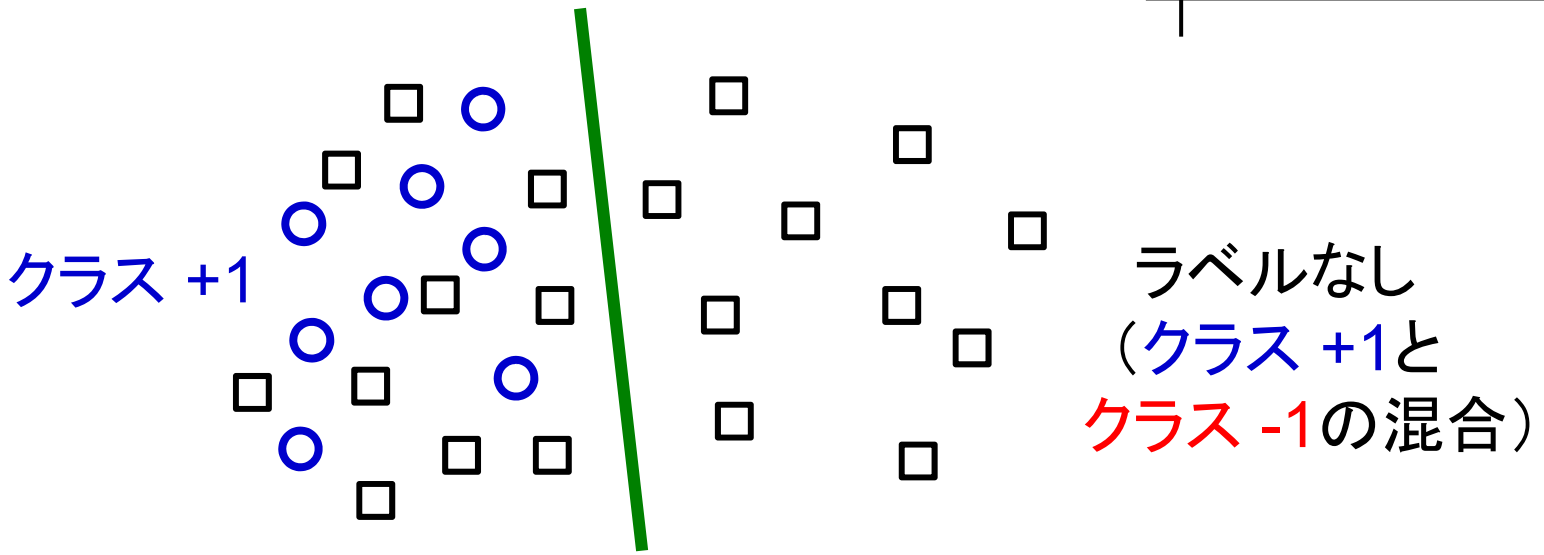
1. 分類問題の分類
2. 正例とラベルなしデータからの分類1
3. 正例とラベルなしデータからの分類2
4. 正例とラベルなしデータからの分類3

凸解法

du Plessis, Niu & Sugiyama (ICML2015)

- ランプ損失は非凸関数なので，ロバストサポートベクトルマシンで大域的最適解を求めるのは困難
- 正例とラベルなしデータに異なる損失を使う





分類器の誤差

- 分類器 $f(\boldsymbol{x})$ の損失 ℓ に対する誤差 $R_\ell(f)$ は，偽陰性と擬陽性の加重和：

$$R_\ell(f) = \pi \int \ell\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

偽陰性（正を負と誤る）

$$+(1-\pi)\int \ell\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

擬陽性（負を正と誤る）

  - クラス事前確率：

$$\pi = p(y=+1)$$

$$1-\pi = p(y=-1)$$

ラベルなしデータに対する誤差

$$\int \ell\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

- 「正例とラベルなしデータからの分類」では負例がないため，ラベルなしデータを負例だと思った場合の誤差を考える:

$$\int \ell\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$$

- これを使えば，分類器 $f(\boldsymbol{x})$ の誤差 $R_\ell(f)$ は

$$R_\ell(f) = \pi \int \Big[\ell\Big(f(\boldsymbol{x})\Big) - \ell\Big(-f(\boldsymbol{x})\Big)\Big]p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

$$+ \int \ell\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

証明

- $p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$ より

$$R_\ell(f) = \pi \int \ell\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

$$+(1-\pi)\int \ell\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

$$= \pi \int \ell\Big(f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

$$+\int \ell\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \pi\int \ell\Big(-f(\boldsymbol{x})\Big)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

合成損失関数

$$R_\ell(f) = \pi \int \Big[\ell\Big(f(\boldsymbol{x})\Big) - \ell\Big(-f(\boldsymbol{x})\Big)\Big] p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}$$

$$+ \int \ell\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

- $\widetilde{\ell}(m) = \ell(m) - \ell(-m)$ という合成損失関数を考えれば

$$R_\ell(f) = \pi \int \widetilde{\ell}\Big(f(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + \int \ell\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

- 定理: $\widetilde{\ell}(m)$ は凸関数 ⇔ $\widetilde{\ell}(m)$ は線形関数

$$\widetilde{\ell}(m) = am + b$$

$$R_\ell(f) = a\pi \int f(\boldsymbol{x}) p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x} + b\pi + \int \ell\Big(-f(\boldsymbol{x})\Big) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

代理損失の選択

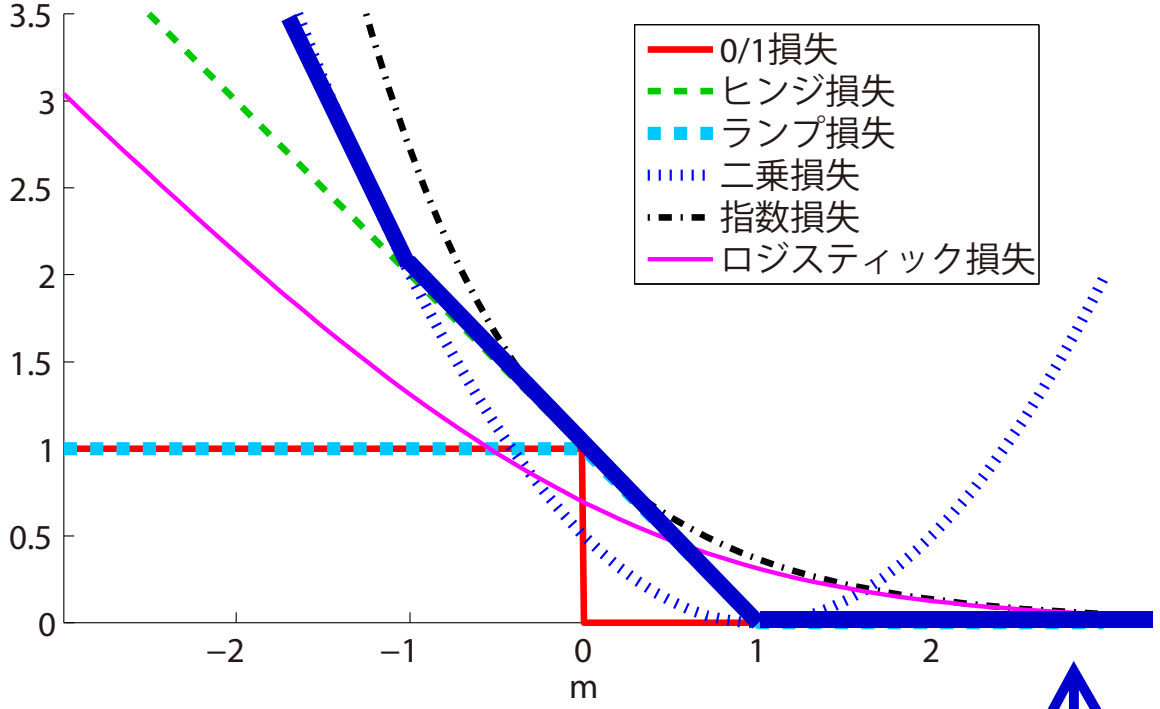
$$\widetilde{\ell}(m) = \ell(m) - \ell(-m) = am + b$$

- NG:
  - ヒンジ損失
  - 指数損失
- OK:
  - ロジスティック損失

    $\ell(m) = \log(1 + e^{-m})$
  - 二乗損失（線形モデルに対して解析解が求まる）

    $\ell(m) = (m-1)^2$
  - 二段ヒンジ損失（線形モデルに対して二次計画）

    $\ell(m) = \max\left(\max(0, 1-m), -2m\right)$



理論解析

- 正例: $\{(\boldsymbol{x}_i, y_i = +1)\}_{i=1}^{n_+}$, ラベルなし: $\{\boldsymbol{x}'_i\}_{i=1}^{n'}$
- 線形モデル: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})$
- 学習規準: $\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \widehat{J}_\ell(f_{\boldsymbol{\theta}})$

$$\widehat{J}_\ell(f_{\boldsymbol{\theta}}) = \frac{\pi}{n_+} \sum_{i=1}^{n_+} \widetilde{\ell}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big) + \frac{1}{n'} \sum_{i=1}^{n'} \ell\Big(-f_{\boldsymbol{\theta}}(\boldsymbol{x}'_i)\Big) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2$$

- 最適解: $\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} J_\ell(f_{\boldsymbol{\theta}})$

$$J_\ell(f_{\boldsymbol{\theta}}) = \pi \int \widetilde{\ell}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big) p(\boldsymbol{x}|y=+1) \mathrm{d}\boldsymbol{x} + \int \ell\Big(-f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2$$

- 定理: ロジスティック損失，二乗損失，二段ヒンジ損失に対して

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = \mathcal{O}_p(n_+^{-1/2} + n'^{-1/2})$$

$$|\widehat{J}_\ell(f_{\widehat{\boldsymbol{\theta}}}) - J_\ell(f_{\boldsymbol{\theta}^*})| = \mathcal{O}_p(n_+^{-1/2} + n'^{-1/2})$$

まとめ

- 教師付き学習：学習精度は良いが，ラベル付けのコストが高い
- 教師なし学習：ラベル付けのコストは不要だが，学習の信頼性が低い
- 半教師付き学習：ラベル付けのコストは抑制できるが，学習精度は必ずしも高くない
- 正例とラベルなしデータからの分類：
  - 正例が簡単に集められる場合に有効
  - 損失を工夫すればうまく学習できる