統計解析特論 講義メモ $\boxed{\mathbf{5}}$ 判別アルゴリズム

- サポートベクトルマシン **(SVM)**
- カーネル-**SVM**
- 多値判別

サポートベクトルマシン(SVM)

- ヒンジ損失を用いた推定法

$$\min_{\boldsymbol{w},b} \frac{1}{n}\sum_{i=1}^{n}[1-y_i(\boldsymbol{w}^T x_i + b)]_+ \quad \longrightarrow \quad \widehat{\boldsymbol{w}}, \widehat{b} \tag{1}$$

仮説： $\widehat{h}(x) = \mathrm{sign}(\,\widehat{\boldsymbol{w}}^T x + \widehat{b}\,)$

(1) は **LP(linear programming prob.)** として表せる.
note: $[a]_+ := \max\{0, a\} = \min\{\xi \,|\, 0 \le \xi,\ a \le \xi\}$.

$$\min_{w,b,\xi} \ \frac{1}{n}\sum_{i=1}^{n}\xi_i, \quad \text{s.t. } 0 \le \xi_i,\ 1 - y_i(\boldsymbol{w}^T x_i + b) \le \xi_i$$

内点法，単体法 **(**シンプレックス法**)** など.

- 正則化項付き**SVM**

$$\min_{\boldsymbol{w},b} \quad \frac{1}{n}\sum_{i=1}^{n}[1 - y_i(\boldsymbol{w}^T x_i + b)]_+ + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \quad \longrightarrow \quad \widehat{\boldsymbol{w}}, \widehat{b}$$

凸**2**次最適化**(**目的関数：凸２次，制約：線形**)**

$$\min_{w,b,\xi} \; \frac{1}{n}\sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}\|\boldsymbol{w}\|^2, \quad \text{s.t. } 0 \le \xi_i,\ 1 - y_i(\boldsymbol{w}^T x_i + b) \le \xi_i$$

効率的に解ける．内点法，有効制約法など．

解釈：判別境界までの距離を最大化

線形仮説 $h(x) = \mathrm{sign}(w^T x + b)$.
正例 **(+1)** と負例 **(-1)** を分ける.

- x_i から 判別境界 $w^T x + b = 0$ までの距離：$\dfrac{|w^T x_i + b|}{\|w\|}$
- ラベル y_i も考慮：符号付き距離 $d_i = \dfrac{y_i(w^T x_i + b)}{\|w\|}$
  - $d_i > 0$：$h(x)$ で (x_i, y_i) を正しく判別．判別境界までの距離 d_i.
  - $d_i < 0$：$h(x)$ で (x_i, y_i) を誤判別．判別境界までの距離 $|d_i|$.

$$\frac{y_i(w^T x_i + b)}{\|w\|} \to \text{大きく} \qquad \overset{y_i(w^T x_i+b)\geq 0 \text{なら}}{\Longleftrightarrow} \qquad \begin{cases} \|w\| & \to \text{小さく} \\ y_i(w^T x_i + b) & \to \text{大きく} \end{cases}$$

$y_i(w^T x_i + b) \to$ 大きく： ヒンジ損失で測る.

$$\frac{1}{n}\sum_{i=1}^{n}[1 - y_i(w^T x_i + b)]_+ + \frac{\lambda}{2}\|w\|^2 \quad \longrightarrow \text{最小化}$$

正則化項 $\|\boldsymbol{w}\|^2$：判別境界までの距離の最大化に対応.

カーネル **SVM**

- 判別関数：$w^T x + b \to w^T \phi(x) + b$,
- カーネル回帰分析と同様に，カーネル関数 $k(x, x')$ を用いる.

$$k(x, x') = \phi(x)^T \phi(x')$$

- $k(x, x')$ に対応する **RKHS** を $\mathcal{H}$ とする.

判別関数 $f(x) + b,\ f \in \mathcal{H}$

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}} \sum_{i=1}^{n} \bigl[1 - y_i\bigl(f(x_i) + b\bigr)\bigr]_+ + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

表現定理より，$f(x) = \sum_j \alpha_j k(x, x_j)$ と表せる.

$$\min_{\alpha, b} \sum_{i=1}^{n} \left[1 - y_i\left(\sum_{j=1}^{n} k(x_i, x_j)\alpha_j + b\right)\right]_+ + \frac{\lambda}{2}\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

$$\longrightarrow \quad \widehat{\alpha},\ \widehat{b}. \quad \text{判別関数} : f(x) = \sum_{i=1}^{n} k(x, x_i)\widehat{\alpha}_i + \widehat{b}$$

$[a]_+ = \min\{\xi \,|\, 0 \le \xi,\ a \le \xi\}$ を用いると，凸**2**次計画問題になる.

CVによるカーネルSVMのモデルパラメータの選択

- モデルパラメータ：{正則化，カーネル}　パラメータ
- ガウシアンカーネル：$\exp\{-\gamma\|x-x'\|^2\}$.

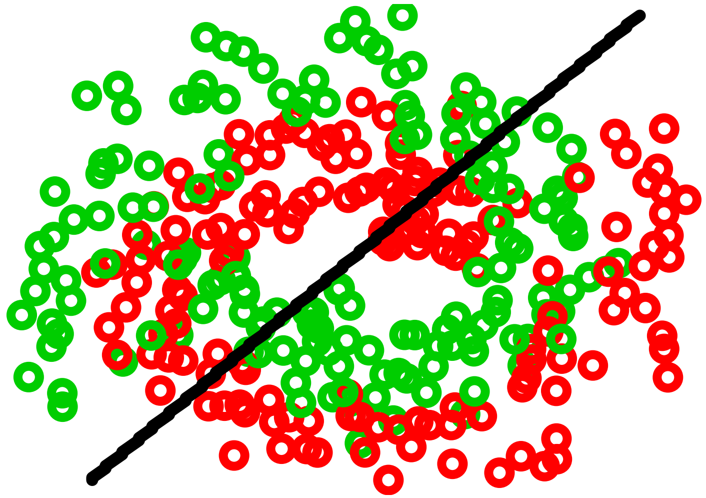
ヒューリスティクス： $\gamma = \mathrm{median}\left\{ \frac{1}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2} \,\middle|\, i,j = 1,\ldots,n,\ i \neq j \right\}$

$\longrightarrow$ 数値計算の桁落ちを防ぐ.

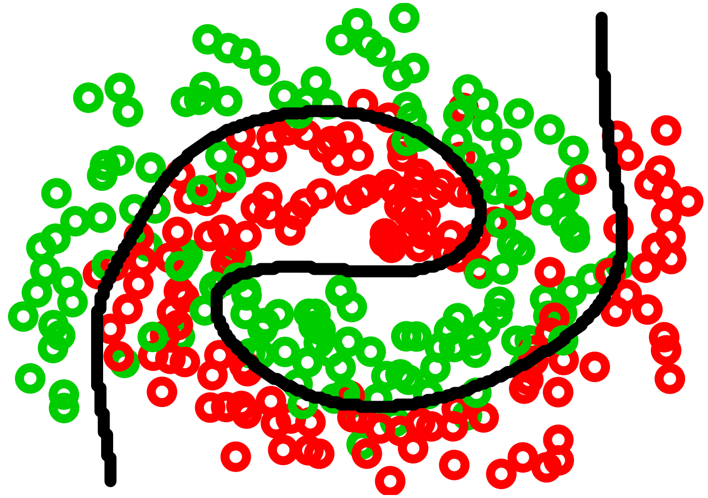
- 交差検証法で予測誤差を推定
  $\Longrightarrow$ 適切なモデルパラメータを決める.

  評価の規準：判別なので **0-1** 損失

例：適切なカーネル関数 $\rightarrow$ 複雑な判別境界の学習



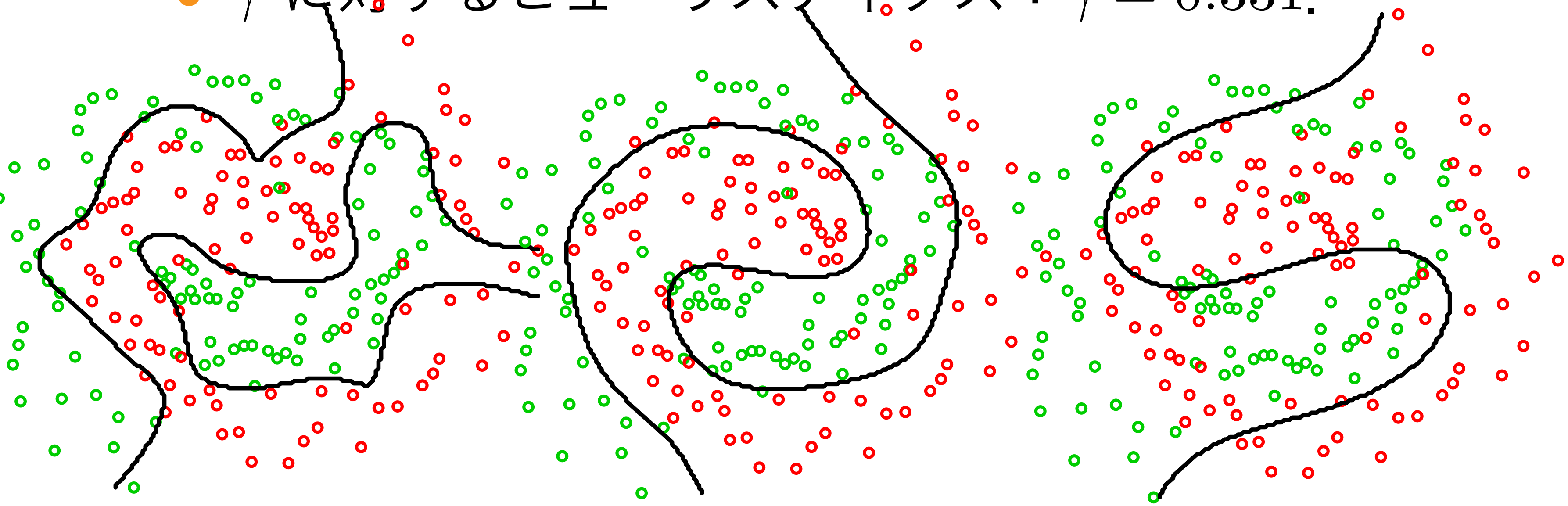
線形カーネル



ガウシアンカーネル

数値例：λ を交差検証法で決定 $(\lambda : 10^{-5} \sim 10^{2})$

- γ に対するヒューリスティクス：$\gamma = 0.331$.



$\lambda = 0.5 \cdot 10^{-4}$　　最適値 $\lambda = 0.247$　　$\lambda = 100$

予測誤差：0.250　　予測誤差：0.193　　予測誤差：0.278

多値判別

- ラベルの種類が３つ以上： $y \in \mathcal{Y} = \{1, 2, \ldots, G\}$
  - 数字の読み取り，アルファベットの読み取り，など
- 方法は大きく分けて２通り

アプローチ１： 複数の２値判別に分割，それぞれ学習，統合.
アプローチ２： 多値判別の判別関数を直接学習

アプローチ１（２値判別法の組合せ）を解説する.
アプローチ２と比較して・・・

- 利点：実装が簡単．すでにある **2** 値判別法 **(SVM** など**)** を組み合わせるだけ.
- 欠点
    - 理論的な精度保証が（あまり）ない
    - 各クラスのデータ数に偏りがあるとき，予測精度が低い傾向.

- **one-vs-one** 法
- **one-vs-all (one-vs-rest)** 法
- 一般化：**Error correcting output coding (ECOC)**

one-vs-one 法

データ：$(x_1, y_1), \ldots, (x_n, y_n)$.
学習法：

1. 2つのラベル $y, y' \in \mathcal{Y}$ を選ぶ.

2. ラベル y, y' のデータのみを使う．2値判別の仮説を学習.

$$h_{yy'}(x) = \begin{cases} +1, & x \text{ のラベルを } y \text{ と予測} \\ -1, & x \text{ のラベルを } y' \text{ と予測} \end{cases}$$

3. ラベルのすべての組合せ y, y' について $h_{yy'}(x)$ を求める.
   - $G(G-1)/2$ の仮説.
   - $h_{yy'}(x) = -h_{y'y}(x)$ とする.

予測法：多数決で決める．新たな入力 x

1. ラベル y のスコア $\mathrm{score}(y)$ を計算：

$$\mathrm{score}(y) = |\{y' \in \mathcal{Y} \mid h_{yy'}(x) = +1\}|$$

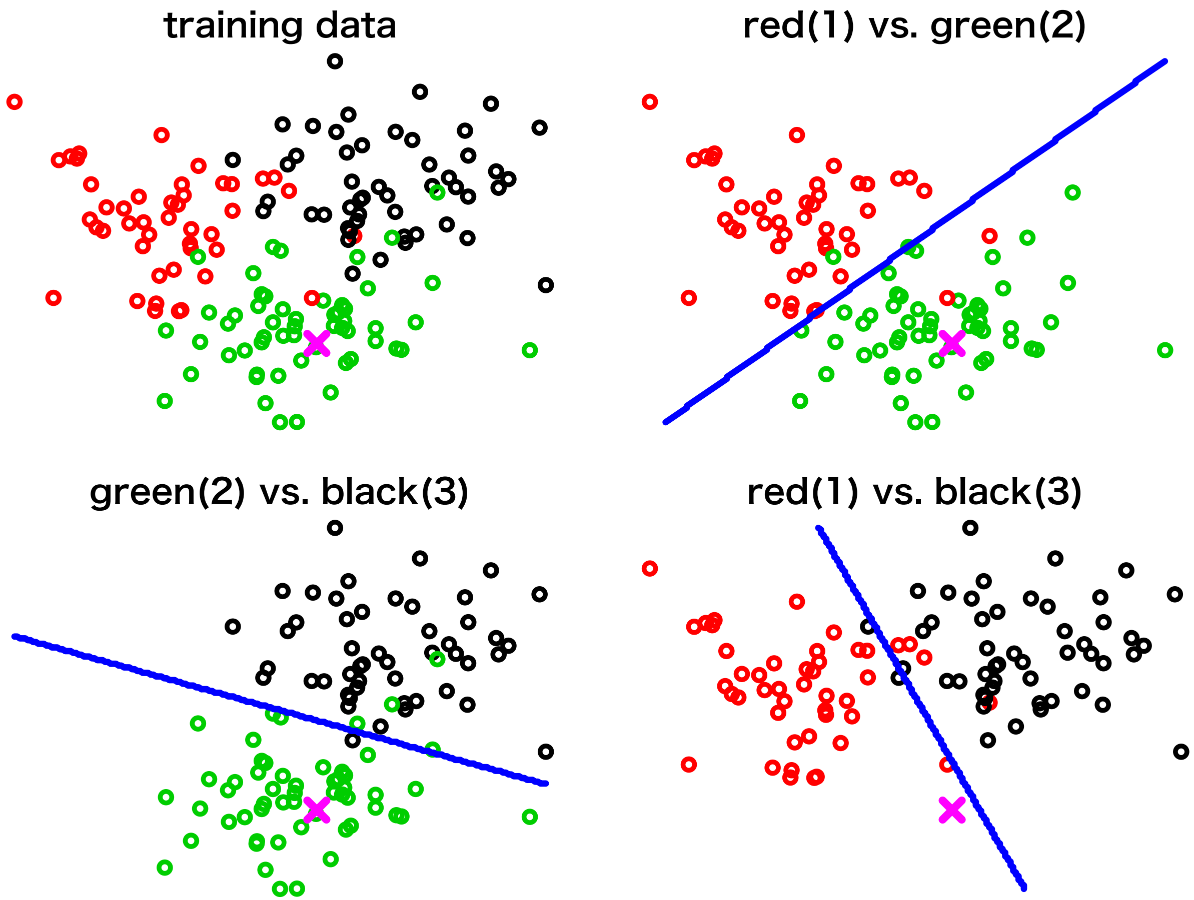
意味：ラベル y に関する仮説 $h_{y1}, \ldots, h_{yG}$ の中で，
x のラベルを y と予測する仮説の数．

2. $\mathrm{score}(y)$ の値が最大になる $\widehat{y}$ を x の予測ラベルとする．

$$\widehat{y} = \mathop{\arg\max}_{y \in \mathcal{Y}} \ \mathrm{score}(y)$$

例：one-vs-one 法

- 学習：多値判別を複数の２値判別に分割



($h_{a,b}(x)$: ラベルが a なら $+1$, b なら -1 を返す)

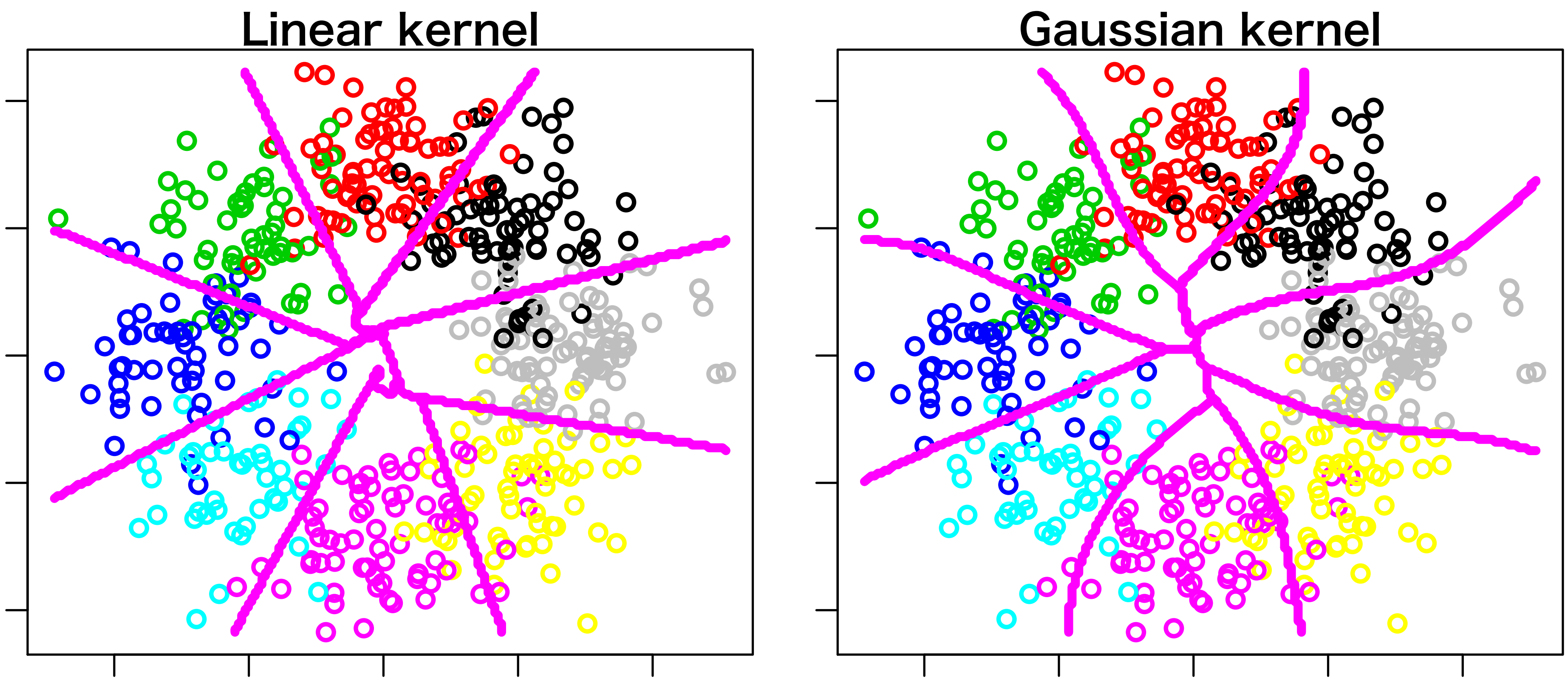
- 予測：多数決による予測.
  図の X のラベルを予測.

  $\mathrm{score}(y) :=$「X のラベルを y と判別する判別器の数」

y				$\mathrm{score}(y)$
red(1)	–	$h_{1,2} = -1$	$h_{1,3} = +1$	1
green(2)	$h_{2,1} = +1$	–	$h_{2,3} = +1$	2
black(3)	$h_{3,1} = -1$	$h_{3,2} = -1$	–	0

したがって X の予測ラベルは **green.**

例：正規分布モデル



one-vs-all (one-vs-rest) 法

学習法：

1. ラベル y とそれ以外を判別．判別関数 $f_y(x)$ を求める．

$$\mathrm{sign}(f_y(x)) = \begin{cases} +1, & x \text{ のラベルを } y \text{ と予測} \\ -1, & x \text{ のラベルを } y \text{ 以外と予測} \end{cases}$$

2. $f_y(x)$, $y \in \mathcal{Y}$ を求める．

x のラベルを予測：

判別関数による予測

1. すべての判別関数 $f_y(x)$ の出力を計算 **(実数値)**

2. $f_y(x)$ の値で比較. $\widehat{y} = \arg\max_{y \in \mathcal{Y}} f_y(x)$.

仮説による予測

1. すべての仮説の出力 $h_y(x) = \mathrm{sign}(f_y(x)),\ y \in \mathcal{Y}$ を計算.

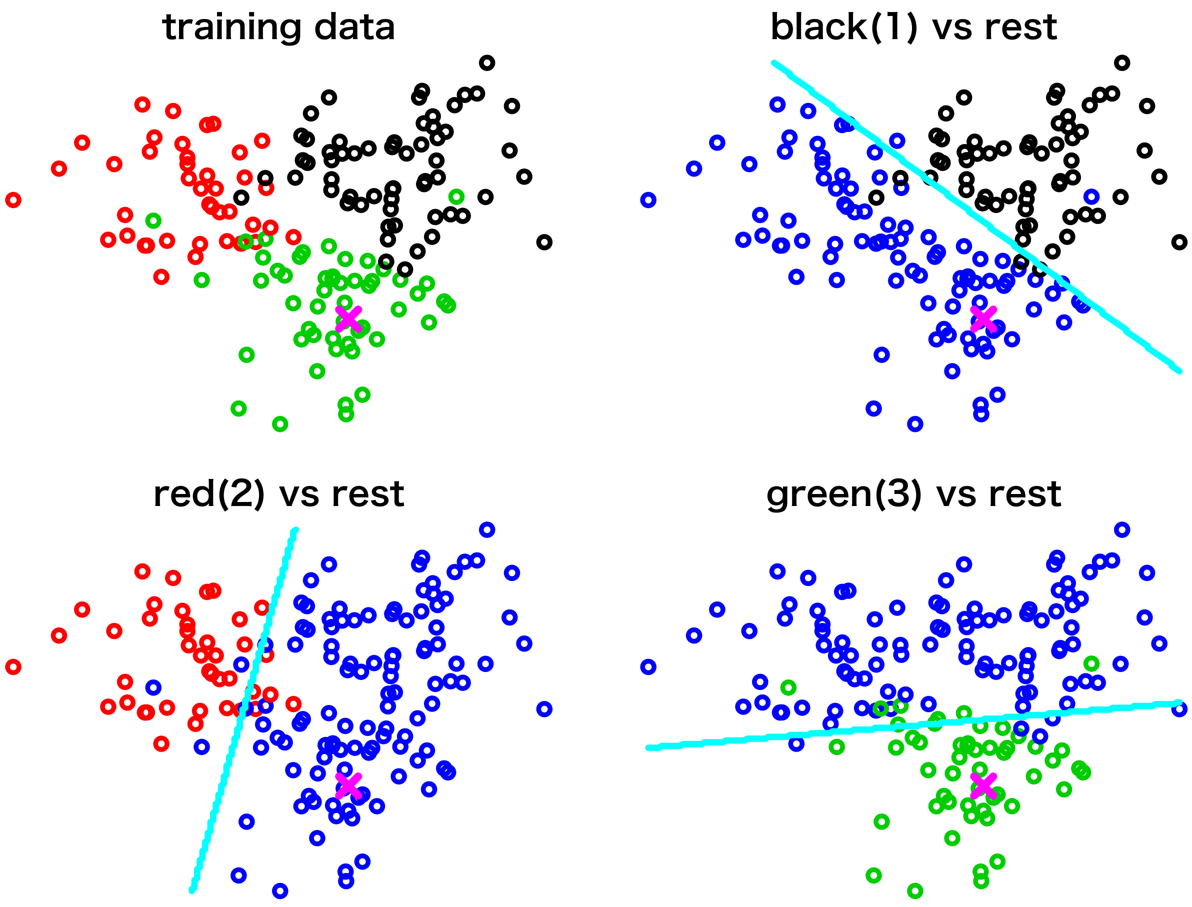
2. $h_y(x) = +1$ となるラベル y を予測ラベルとする.

$$\widehat{y} = \{y \in \mathcal{Y} \,|\, h_y(x) = +1\}.$$

note: $+1$ が複数 **or** ひとつもない $\Rightarrow$ 予測ラベルが決まらない

例：**one-vs-all**法

ラベル：● ● ●，残りのラベルを●でプロット．



X の予測ラベルは **green.**

誤り訂正出力符号化法(ECOC)

- 誤り訂正符号の考え方を応用
- **one-vs-one**, **one-vs-all** の一般化：

References:

Dietterich, Bakiri, Solving Multiclass Learning Problems via Error Correcting Output Codes, Journal of Artificial Intelligence Research, vol. 2 Issue 1, pp. 263-286, 1994.

Allwein, Schapire, Singer, Reducing multiclass to binary: a unifying approach for margin classifiers, The Journal of Machine Learning Research, vol. 1, pp. 113-141, 2001.

学習：

- 各ラベル y に符号語 $c_y = (c_{y1}, \ldots, c_{yT}) \in \{+1, -1, 0\}^T$ を対応させる.

例：

	t $(T=5)$				
label	**1**	**2**	**3**	**4**	**5**
1 (c_1)	-1	-1	-1	-1	0
2 (c_2)	$+1$	$+1$	-1	0	-1
3 (c_3)	0	-1	$+1$	$+1$	-1
4 (c_4)	$+1$	0	-1	$+1$	$+1$

- $t = 1, \ldots, T$ に対して
  - (x_i, y_i) を 2値データ $(x_i, c_{y_i t})$ に変換 $(i = 1, \ldots, n)$.
  - $c_{y_i t} \neq 0$ のデータのみ用いて，2値仮説 $h_t(x)$ を学習.

予測：

- x に対して $h_t(x) \in \{+1, -1\}$, $t = 1, \ldots, T$ を計算.
- $(h_1(x), \ldots, h_T(x))$ に最も近い $c_y = (c_{y1}, \ldots, c_{yT})$ を探す.
  対応するラベル $\widehat{y}$ を予測ラベルとする.

  近さの測り方：ハミング距離．$a, b \in \{+1, -1, 0\}$ に対して

$$d(a,b) = \frac{1-ab}{2} = \begin{cases} 1, & a \neq b, ab \neq 0 \\ 1/2, & ab = 0 \\ 0, & a = b \neq 0 \end{cases}$$

$$\widehat{y} = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{t=1}^{T} d(c_{yt}, h_t(x))$$

(この定義だと距離の公理は満たさないが，便宜上距離とよぶ**)**

符号語 c_y (例：ラベル数4)

- **one-vs-one**：

	t					
label	**1**	**2**	**3**	**4**	**5**	**6**
1 (c_1)	$+1$	$+1$	$+1$	0	0	0
2 (c_2)	-1	0	0	$+1$	$+1$	0
3 (c_3)	0	-1	0	-1	0	$+1$
4 (c_4)	0	0	-1	0	-1	-1

- 各列に対応して，元データを2値に変換して学習．
  例：$t = 1$ ではラベル $1, 2$ のデータのみ使用．
- ハミング距離で予測：$\mathrm{score}(y)$ と同じ：
  c_{yt} が 0 でない要素どうしを比較 (どの c_y も 0 の数は同じ)．

- **one-vs-all**：仮説による予測

	t			
label	**1**	**2**	**3**	**4**
1 (c_1)	$+1$	-1	-1	-1
2 (c_2)	-1	$+1$	-1	-1
3 (c_3)	-1	-1	$+1$	-1
4 (c_4)	-1	-1	-1	$+1$

符号間の最小距離：$\rho = \min_{y,y' \in \mathcal{Y},\, y \neq y'} \sum_{t=1}^{T} d(c_{yt}, c_{y't})$.

- t 列目の符号化 $\to$ 仮説 $h_t(x) \in \{+1, -1\}$
- **ECOC** で得られる仮説：$H(x) \in \mathcal{Y}$

$\widetilde{e}(h_t) = \frac{1}{n} \sum_{i=1}^{n} d(c_{y_i t}, h_t(x_i))$: $\{(x_i, c_{y_i t})\}_{i=1}^{n}$ に対する学習誤差.

($c_{y_i t} = 0$ なら **loss** は $1/2$, そうでなければ **0-1 loss** で測る**)**

$H(x)$ の学習誤差について次式が成立：

$$\widehat{e}(H) \leq \frac{2T}{\rho} \cdot \frac{1}{T} \sum_{t=1}^{T} \widetilde{e}(h_t),$$

$\frac{\rho}{T} \to$ 大きい $\Longrightarrow$ 学習誤差小さい

$\widehat{e}(H)$ の評価式の証明. $H(x_i) \neq y_i$ のとき,

$$\exists y' \neq y_i, \quad \sum_{t=1}^{T} d(c_{y_i t}, h_t(x_i)) \geq \sum_{t=1}^{T} d(c_{y' t}, h_t(x_i)) \tag{2}$$

符号語 c_{y_i}, $c_{y'}$ に対して $\mathcal{T}_0, \mathcal{T}_{-1}$ を定義：

$$\mathcal{T}_0 = \{t \,|\, c_{y_i t} c_{y' t} = 0\}, \ \mathcal{T}_{-1} = \{t \,|\, c_{y_i t} c_{y' t} = -1\}.$$

以下が成立

$$\begin{aligned} t \in \mathcal{T}_0 \ &\Rightarrow \ d(c_{y_i t}, h_t(x_i)) + d(c_{y' t}, h_t(x_i)) \geq \frac{1}{2}, \\ t \in \mathcal{T}_{-1} \ &\Rightarrow \ d(c_{y_i t}, h_t(x_i)) + d(c_{y' t}, h_t(x_i)) = 1. \end{aligned}$$

(2) より

$$
\begin{aligned}
\sum_{t=1}^{T} d(c_{y_i t}, h_t(x_i)) &\geq \frac{1}{2} \sum_{t=1}^{T} \left\{ d(c_{y_i t}, h_t(x_i)) + d(c_{y' t}, h_t(x_i)) \right\} \\
&\geq \frac{1}{2} \sum_{t \in \mathcal{T}_0 \cup \mathcal{T}_{-1}} \left\{ d(c_{y_i t}, h_t(x_i)) + d(c_{y' t}, h_t(x_i)) \right\} \\
&\geq \frac{1}{4} |\mathcal{T}_0| + \frac{1}{2} |\mathcal{T}_{-1}| = \frac{1}{2} \sum_{t=1}^{T} d(c_{y_i t}, c_{y' t}) \geq \frac{\rho}{2}
\end{aligned}
$$

したがって,

$$H(x_i) \neq y_i \implies 1 \leq \frac{2}{\rho} \sum_{t=1}^{T} d(c_{y_i t}, h_t(x_i))$$

$$\implies \widehat{e}(H) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[H(x_i) \neq y_i]$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \frac{2}{\rho} \sum_{t=1}^{T} d(c_{y_i t}, h_t(x_i)) = \frac{2T}{\rho} \cdot \frac{1}{T} \sum_{t=1}^{T} \widetilde{e}(h_t)$$

■

note: 予測誤差 $e(H)$ と符号化 c_{yt} に関係ついて，理論的な結果はあまりない.

例：$\mathcal{Y} = \{1, \ldots, G\}$

- **one-vs-one:**

$$\rho = \frac{G^2 - G + 2}{4},\ T = \frac{G(G-1)}{2} \implies \frac{\rho}{T} = \frac{1}{2} + \frac{1}{G-1} - \frac{1}{G} > \frac{1}{2}$$

ρについて：$c_y, c_{y'}$ の各要素は，$c_{yt}c_{y't} = -1$となる t が 1つ，残りはどちらかが 0.

- **one-vs-all:** $\rho = 2,\ T = G \implies \dfrac{\rho}{T} = \dfrac{2}{G}$

$G \to \infty$ のとき

- **one-vs-one:** $\rho/T \to 1/2$.
- **one-vs-all:** $\rho/T \to 0$.

$\widetilde{e}(h_t)$も併せて考える必要があるが，**one-vs-one** のほうが **training error** は小さい傾向．**(**特に G が大きいとき**)**

— 補足：学習アルゴリズムの統計的性質 —

- **Rademacher** 複雑度
- **SVM** の予測誤差

参考文献：金森 敬文 著「統計的学習理論」，講談社，**2015.**

Rademacher 複雑度

集合 $\mathcal{Z}$ から $\mathbb{R}$ への関数の集合を $\mathcal{G} \subset \mathcal{Z}^{\mathbb{R}}$ とする.
$g \in \mathcal{G}$ に対して $g(z) \in \mathbb{R}$, $z \in \mathcal{Z}$.
Definition 1 (経験 Rademacher(ラデマッハ)複雑度)**.** 関数集合 $\mathcal{G}$ と $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ に対して経験 Rademacher 複雑度 $\widehat{\mathfrak{R}}_S(\mathcal{G})$ を

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)\right]$$

と定義する．ここで $\sigma_1, \ldots, \sigma_n$ は独立に確率分布 $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$ にしたがう確率変数，$\mathbb{E}_{\boldsymbol{\sigma}}$ は $\sigma_1, \ldots, \sigma_n$ に関する期待値を意味する.

例 **1.** $S = \{z_1, z_2\}$ のとき

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{4}\Bigg[&\sup_{g\in\mathcal{G}}\frac{1}{2}(g(z_1)+g(z_2)) + \sup_{g\in\mathcal{G}}\frac{1}{2}(g(z_1)-g(z_2)) \\ &+ \sup_{g\in\mathcal{G}}\frac{1}{2}(-g(z_1)+g(z_2)) + \sup_{g\in\mathcal{G}}\frac{1}{2}(-g(z_1)-g(z_2))\Bigg]\end{aligned}$$

Rademacher複雑度の解釈

- ラベルを $\mathrm{sign}(g(x))$ で予測することを考える
- 仮説集合 $\mathcal{G}$ でランダムに割り当てられるラベルに，どの程度対応できるか？
- $\widehat{\mathfrak{R}}_S(\mathcal{G})$ が大きい程，$\mathcal{G}$ は複雑な関数を含む.

Rademacher複雑度による推定誤差の評価

集合 $\mathcal{Z}$ から $[0,1]$ への関数の集合を $\mathcal{G} \subset \mathcal{Z}^{[0,1]}$ とする.
$\mathcal{Z}$に値をとる確率変数$Z_1, \dots, Z_n \sim_{i.i.d.} P$に対して,
以下の不等式が成り立つ.

$$\Pr\left\{\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right| \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}\right\} \geq 1-\delta$$

ここで $S = \{Z_1, \dots, Z_n\}$. すなわち, $1-\delta$ 以上の確率で

$$\forall g \in \mathcal{G}, \quad \mathbb{E}[g(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} g(Z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

が成り立つ (任意の$\mathcal{G}$で成立).

証明：**reference**を参照のこと.

- **Bartlett, P. L., Mendelson, S., Rademacher and Gaussian Complexities: Risk Bounds and Structural Results, Journal of Machine Learning Research 3 (2002) 463-482.**
- **Mohri, M. et al., Foundations of Machine Learning, The MIT Press, 2012. (Chap. 3)**

Rademacher複雑度の性質

$1 \sim 3$ まで紹介：

1. $\mathcal{G}_1 \subset \mathcal{G}_2 \Longrightarrow \widehat{\mathfrak{R}}_S(\mathcal{G}_1) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}_2)$

2. $c \in \mathbb{R}$ に対して$\widehat{\mathfrak{R}}_S(c\mathcal{G}_1) = |c|\widehat{\mathfrak{R}}_S(\mathcal{G})$.
ここで $c\mathcal{G} = \{cg : g \in \mathcal{G}\}$.

3. $\phi : \mathbb{R} \to \mathbb{R}$はリプシッツ定数$L$のリプシッツ連続関数のとき,
$\widehat{\mathfrak{R}}_S(\phi \circ \mathcal{G}) \leq L\widehat{\mathfrak{R}}_S(\mathcal{G})$.
ここで$\phi \circ \mathcal{G} = \{\phi \circ g : g \in \mathcal{G}\}$.

4. 任意の関数hに対して$\widehat{\mathfrak{R}}_S(\mathcal{G} + h) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \frac{\|h\|_\infty}{\sqrt{m}}$.
ここで$\mathcal{G} + h = \{g + h : g \in \mathcal{G}\}$.

5. $\mathcal{G}$ の凸包を $\mathrm{conv}\mathcal{G}$ とすると, $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_S(\mathrm{conv}\mathcal{G})$.

Proof. **1.** は自明.

2. $c = 0$ なら自明. $c \neq 0$ のとき

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(c\mathcal{G}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i c g(z_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i |c| \mathrm{sing}(c) g(z_i)\right] \\
&= |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \mathrm{sing}(c) g(z_i)\right] \\
&= |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\right]
\end{aligned}
$$

最後の行：σ_i と $\sigma_i\,\mathrm{sign}(c)$ はどちらも確率 $1/2$ で ± 1 の値をとるので分布は同じ. ■

3. は証明が少々面倒なのでパス. 結果は重要.

4.

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}+h) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i(g(z_i)+h(z_i))\right] \\
&\leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^{m} \sigma_i h(z_i)\right|\right] \\
&\leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^{m} \sigma_i h(z_i)\right|^2\right]^{1/2} \\
&= \widehat{\mathfrak{R}}_S(\mathcal{G}) + \frac{1}{m}\sqrt{\sum_{i=1}^{m} h(z_i)^2} \\
&\leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \frac{\|h\|_\infty}{\sqrt{m}}
\end{aligned}
$$

Rademacher複雑度の例

仮説集合 $\mathcal{H}$ が有限のとき

$$\mathcal{G} = \{g(z) = \mathbf{1}[h(x) \neq y] : h \in \mathcal{H}\}, \quad z = (x, y)$$

に対する $\widehat{\mathfrak{R}}_S(\mathcal{G})$ を評価する.

- 以下の不等式を使う
  - **Jensen**不等式：凹関数ϕに対して$\mathbb{E}[\phi(Z)] \leq \phi(\mathbb{E}[Z])$.
  - **Hoeffding's lemma**：
    $\Pr\{a \leq X \leq b\} = 1$ のとき $\mathbb{E}[e^X] \leq e^{(b-a)^2/8}$.
    Hoeffding不等式の証明に用いられる.

$t > 0$ に対して

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}) &= \frac{1}{t}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\max_g \frac{t}{n}\sum_{i=1}^n \sigma_i g(z_i)\Big] \le \frac{1}{t}\log \mathbb{E}_{\boldsymbol{\sigma}}\Big[\exp\Big\{\max_g \frac{t}{n}\sum_{i=1}^n \sigma_i g(z_i)\Big\}\Big] \\
&= \frac{1}{t}\log \mathbb{E}_{\boldsymbol{\sigma}}\Big[\max_g \exp\Big\{\frac{t}{n}\sum_{i=1}^n \sigma_i g(z_i)\Big\}\Big] \le \frac{1}{t}\log \sum_g \mathbb{E}_{\boldsymbol{\sigma}}\Big[\exp\Big\{\frac{t}{n}\sum_{i=1}^n \sigma_i g(z_i)\Big\}\Big] \\
&= \frac{1}{t}\log \sum_g \Pi_{i=1}^n \mathbb{E}_{\sigma_i}[e^{t\sigma_i g(z_i)/n}] \le \frac{1}{t}\log \sum_g \Pi_{i=1}^n e^{t^2/2n^2} \\
&\le \frac{1}{t}\log \sum_g e^{t^2/2n} = \frac{1}{t}\log |\mathcal{G}| e^{t^2/2n} = \frac{\log|\mathcal{G}|}{t} + \frac{t}{2n}
\end{aligned}$$

t について最適化して　$\widehat{\mathfrak{R}}_S(\mathcal{G}) \le 2\sqrt{\dfrac{\log|\mathcal{G}|}{2n}} \le 2\sqrt{\dfrac{\log|\mathcal{H}|}{2n}}$

ほとんど同じ不等式が得られる.

- 前に得られた結果

$$
\begin{aligned}
e(\widehat{h}) &\leq e(h_0) + \textbf{bias} + \sqrt{\frac{2}{n}\log\frac{|\mathcal{H}|+1}{\delta}} \\
&\leq e(h_0) + \textbf{bias} + \sqrt{\frac{2\log 2|\mathcal{H}|}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}.
\end{aligned}
$$

- **Rademacher**複雑度による評価

$$
e(\widehat{h}) \leq e(h_0) + \textbf{bias} + \sqrt{\frac{\log|\mathcal{H}|}{n}} + \sqrt{\frac{\log(4/\delta)}{n}}.
$$

SVMの予測誤差

- ガウスカーネル $k(x,x') = \exp\{-\gamma\|x-x'\|^2\}$ を用いたSVMで判別関数 $\widehat{f} \in \mathcal{H}$ を学習．$k(x,x)=1$ を使う．
- データ $(x_1,y_1),\ldots,(x_n,y_n) \in \mathcal{X} \times \{+1,-1\}$, $\mathcal{X}$.
  i.i.d. from $P(X,Y)$**.**
- ヒンジ損失を $\ell(z) = \max\{1-z, 0\}$ とおく．以下のように $\widehat{f}$ を推定.

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f(x_i)) + \lambda_n \|f\|_{\mathcal{H}}^2 \ \longrightarrow\ \widehat{f} \in \mathcal{H}$$

  - 正則化パラメータ λ_n はデータ数に依存してよい．
  - 簡単のため，バイアス項 b は付けない．

- 証明すること：
  仮定：$\mathcal{X}$ はコンパクト集合，また $\lambda_n \to 0,\ n\lambda_n \to \infty$.
  このとき $e(\mathrm{sign} \circ \widehat{f}) \xrightarrow{p} e(h_0)$ (Bayes error) となる.

- 手順：

1. **Rademacher** 複雑度による推定誤差の評価式から次式を示す.

$$\mathbb{E}[\ell(Y\widehat{f}(x))] \xrightarrow{p} \inf_{f:\text{可測}} \mathbb{E}[\ell(Yf(X))]$$

2. 代替損失の理論から次式を示す.

$$e(\mathrm{sign} \circ \widehat{f}) \xrightarrow{p} \text{Bayes error}$$

以下では，$\mathcal{X}$ はコンパクト集合，$\lambda_n \to 0,\ n\lambda_n \to \infty$ を仮定する.

予備知識

- (X,Y) の任意の分布 P に対して

$$\inf_{f:\text{可測}} \mathbb{E}[\ell(Yf(X))] = \inf_{f\in\mathcal{H}} \mathbb{E}[\ell(Yf(X))].$$

- 補題：データ数が n のとき，$\|\widehat{f}\|_{\mathcal{H}} \leq 1/\sqrt{\lambda_n}$.

$$\lambda_n\|\widehat{f}\|_{\mathcal{H}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\ell(y_i\widehat{f}(x_i)) + \lambda_n\|\widehat{f}\|_{\mathcal{H}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\ell(0) + \lambda_n\|0\|_{\mathcal{H}}^2 = 1$$

より $\|\widehat{f}\|_{\mathcal{H}} \leq 1/\sqrt{\lambda_n}$.

RKHSモデルのRademacher複雑度

ガウスカーネル k に対応するRKHSを $\mathcal{H}$ とする.
また$\mathcal{H}$の部分集合 $\mathcal{F}_n$ を

$$\mathcal{F}_n = \left\{ f \in \mathcal{H} \,:\, \|f\|_{\mathcal{H}} \leq 1/\sqrt{\lambda_n} \right\}$$

とする. 任意の$S = \{x_1, \ldots, x_n\} \subset \mathcal{X}$に対して

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq \frac{1}{\sqrt{n\lambda_n}}$$

Proof.

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\|f\|\leq 1/\sqrt{\lambda_n}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\Big] = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\|f\|\leq 1/\sqrt{\lambda_n}} \Big\langle f, \sum_{i=1}^{n}\sigma_i k(\cdot, x_i)\Big\rangle\Big] \\
&= \frac{1/\sqrt{\lambda_n}}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\Big\|\sum_{i=1}^{n}\sigma_i k(\cdot, x_i)\Big\|_{\mathcal{H}}\Big] \leq \frac{1/\sqrt{\lambda_n}}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\Big\|\sum_{i=1}^{n}\sigma_i k(\cdot, x_i)\Big\|_{\mathcal{H}}^2\Big]^{1/2} \\
&= \frac{1/\sqrt{\lambda_n}}{n}\Big(\sum_{i,j} k(x_i, x_j)\mathbb{E}_{\boldsymbol{\sigma}}\big[\sigma_i\sigma_j\big]\Big)^{1/2} = \frac{1/\sqrt{\lambda_n}}{n}\Big(\sum_{i=1}^{n} k(x_i, x_i)\Big)^{1/2} \leq \frac{1}{\sqrt{n\lambda_n}}
\end{aligned}
$$

■

関数集合 $\mathcal{G}_n$ を

$$\mathcal{G}_n = \{(x,y) \mapsto \ell(yf(x)) \,:\, f \in \mathcal{F}_n\}, \quad z = (x,y) \in \mathcal{X} \times \{+1,-1\}$$

とすると $\widetilde{S} = \{(x_1,y_1),\ldots,(x_n,y_n)\}, S = \{x_1,\ldots,x_n\}$ に対して

$$\widehat{\mathfrak{R}}_{\widetilde{S}}(\mathcal{G}_n) \le \widehat{\mathfrak{R}}_S(\mathcal{F}_n) \le \frac{1}{\sqrt{n\lambda_n}}$$

となる.

$f \in \mathcal{F}_n$ に対して

$$|f(x)| \leq |\langle f, k(\cdot, x)\rangle| \leq \frac{1}{\sqrt{\lambda_n}}$$

より $0 \leq \ell(yf(x)) \leq \dfrac{1}{\sqrt{\lambda_n}} + 1$.

Rademacher 複雑度による推定誤差の評価式より・・・
確率 $1-\delta$ 以上で

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}_n}\left|\mathbb{E}[\ell(Yf(X))]-\frac{1}{n}\sum_{i=1}^{n}\ell(Y_if(X_i))\right| \\
&\leq 2\widehat{\mathfrak{R}}_{\widetilde{S}}(\mathcal{G}_n)+3\max_{\substack{f\in\mathcal{F}_n\\ x,y}}\ell(yf(x))\cdot\sqrt{\frac{\log(1/\delta)}{2n}} \\
&\leq C\sqrt{\frac{\log(2/\delta)}{2n\lambda_n}}
\end{aligned}
$$

となる．$C>0$ はある定数．以下を用いた．

$$
\max_{\substack{f\in\mathcal{F}_n\\ x,y}}\ell(yf(x))\leq\frac{1}{\sqrt{\lambda_n}}+1
$$

詳細：

$$
\begin{aligned}
&\mathbb{E}[\ell(Yf(X))] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i f(X_i)) + 2\mathfrak{R}_{\widetilde{S}}(\mathcal{G}_n) + 3\left(\frac{1}{\sqrt{\lambda_n}} + 1\right)\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i f(X_i)) + 2\mathfrak{R}_{S}(\mathcal{F}_n) + 3\left(\frac{1}{\sqrt{\lambda_n}} + 1\right)\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i f(X_i)) + \frac{2}{\sqrt{n\lambda_n}} + 3\sqrt{\frac{\log(2/\delta)}{2n\lambda_n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}
\end{aligned}
$$

$1-\delta$以上の確率で次が成り立つ.

$$
\begin{aligned}
\mathbb{E}[\ell(Y\widehat{f}(x))] &= \mathbb{E}[\ell(Yf^*(x))] + \mathbb{E}[\ell(Y\widehat{f}(x))] - \mathbb{E}[\ell(Yf^*(x))] \\
&= \mathbb{E}[\ell(Yf^*(x))] + \mathbb{E}[\ell(Y\widehat{f}(x))] - \widehat{\mathbb{E}}[\ell(Y\widehat{f}(x))] - \lambda_n\|\widehat{f}\|^2 \\
&\quad + \widehat{\mathbb{E}}[\ell(Y\widehat{f}(x))] + \lambda_n\|\widehat{f}\|^2 - \mathbb{E}[\ell(Yf^*(x))] \\
&\leq \mathbb{E}[\ell(Yf^*(x))] + \mathbb{E}[\ell(Y\widehat{f}(x))] - \widehat{\mathbb{E}}[\ell(Y\widehat{f}(x))] \\
&\quad + \widehat{\mathbb{E}}[\ell(Yf^*(x))] + \lambda_n\|f^*\|^2 - \mathbb{E}[\ell(Yf^*(x))] \\
&\leq \inf_{f:\text{可測}} \mathbb{E}[\ell(Yf(X))] + \varepsilon + 2C\sqrt{\frac{\log(1/\delta)}{n\lambda_n}} + \lambda_n\|f^*\|^2
\end{aligned}
$$

$\lambda_n \to 0,\ n\lambda_n \to \infty$ とすると,

$$\mathbb{E}[\ell(Y\widehat{f}(x))] \xrightarrow{p} \inf_{f:\text{可測}} \mathbb{E}[\ell(Yf(X))]$$

となることが分かる．代替損失の理論から,

$$e(\mathrm{sign} \circ \widehat{f}) \xrightarrow{p} e(h_0), \quad \textbf{(Bayes error)}$$

となる.