Foundations of Machine Learning

3.1 ラデマッハ複雑度

3.2 成長関数

joeyoji

June 22, 2021

目次

本発表資料の見方

- 式番号や定理番号は FML と一致させています.
- 用語の省略があります.
- 題目に*が付いているのは本編の流れにはない箇所です. 時間の都合上飛ばす可能性があります. また間違っている可能性もあります.
- 証明を省略している箇所は補足資料で証明を行なっています.

3章の目標

これまで

2章：仮説集合が有限のときの汎化誤差の上界を計算

but

機械学習で扱う仮説集合はふつう無限集合

⇓

目標

有限集合での結果を一般化

and

無限集合での学習保証を計算

方法

アイデア：無限の場合を有限の場合の分析に還元

↓

還元の為にいくつかの異なる複雑さ (complexity) の概念を導入

- ラデマッハ複雑度 (Rademacher complexity)
- 成長関数 (growth function)
- VC 次元 (次回)

ラデマッハ複雑度の導入

ラデマッハ複雑度は...

- マクダイアミッドの不等式 (D.16) を使えば証明が比較的簡単
- 高い精度の上界を獲得 (データ依存するものも含む)
- 後々の章でも頻繁に利用

however...

いくつかの仮説集合に対しては計算が NP 困難

$\rightarrow$ これを解消する為に成長関数及び VC 次元の概念を導入

成長関数と VC 次元の導入

成長関数と VC 次元は組み合わせ的な概念

1. ラデマッハ複雑度を成長関数に関連付ける
2. VC 次元に関して成長関数を上から抑える

VC 次元は上から抑えたり推定するのがしばしば容易

記法等準備

記号	定義, 意味
$\mathcal{X}$	入力集合
$\mathcal{Y}$	ラベル集合
$\mathcal{H}$	仮説 $h : \mathcal{X} \to \mathcal{Y}$ の集合
L	損失関数 $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

$\mathcal{G}$ は $\mathcal{H}$ に関する損失関数の族 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ を想定

$$\mathcal{G} = \{g : (x, y) \mapsto L(h(x), y) \mid h \in \mathcal{H}\}$$

但し以下の定義では $\mathcal{G}$ は任意の $\mathcal{Z}$ から $\mathbb{R}$ に対する関数の族という条件で十分

経験ラデマッハ複雑度

定義 3.1 (経験ラデマッハ複雑度)

$\mathcal{G}$ を $\mathcal{Z}$ から $[a,b]$ への関数の族とし $S=(z_1,\ldots,z_m)^\top$ を $\mathcal{Z}$ の標本とする. このとき**標本 S に関する $\mathcal{G}$ の経験ラデマッハ複雑度**は以下で定義される.

$$\widehat{\mathfrak{R}}_S(\mathcal{G})=\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] \tag{3.1}$$

ここで $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_m)^\top$ の各成分は, 互いに独立で一様に $\{-1,1\}$ を値取る確率変数で**ラデマッハ変数**と呼ぶ.

経験ラデマッハ複雑度の意味

$\boldsymbol{g}_S = (g(z_1), \ldots, g(z_m))^\top$ とすると

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_S}{m} \right]$$

内積 $\boldsymbol{\sigma} \cdot \boldsymbol{g}_S$ は標本の損失 $\boldsymbol{g}_S$ と雑音 $\boldsymbol{\sigma}$ の相関

$\rightarrow$　$\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_S}{m}$ は S 上で $\mathcal{G}$ がどれくらい $\boldsymbol{\sigma}$ と相関しているかという量

$\rightarrow$　$\widehat{\mathfrak{R}}_S(\mathcal{G})$ は S 上で $\mathcal{G}$ と雑音が平均的にどれくらい相関しているかという量

単調増大性

関数族が豊か, 或いは複雑な方が複雑度は大きい

$\because \mathcal{G} \subset \mathcal{F}$ とすると

$$\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_S}{m} \leq \sup_{g \in \mathcal{F}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_S}{m}$$

なので

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_S}{m} \right] \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{F}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_S}{m} \right] = \widehat{\mathfrak{R}}_S(\mathcal{F})$$

*経験ラデマッハ複雑度の上界

$\frac{b-a}{2}$ は $\widehat{\mathfrak{R}}_S(\mathcal{G})$ の上界

$\because k_{\boldsymbol{\sigma}}$ を $(\sigma_i)_{i=1}^m$ のうち $\sigma_i = 1$ である添字の個数とする.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_S}{m}\right] &\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{bk_{\boldsymbol{\sigma}} - a(m-k_{\boldsymbol{\sigma}})}{m}\right] \\
&= \frac{1}{2^m}\sum_{k=0}^{m}\binom{m}{k}\left(-a + \frac{a+b}{m}k\right) \\
&= -a + \frac{a+b}{2} = \frac{b-a}{2}
\end{aligned}$$

達成の例: $\mathcal{G} = \{0 \mapsto a, 0 \mapsto b\}, S = \{0\}$ のとき

*経験ラデマッハ複雑度の下界

0 は $\widehat{\mathfrak{R}}_S(\mathcal{G})$ の下界

$$
\begin{aligned}
\because \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_S}{m}\right] &\geq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_S}{m}\right] \\
&= \frac{1}{2^m}\sum_{\sigma_1}\cdots\sum_{\sigma_m}\frac{\sigma_1 g(z_1)+\cdots+\sigma_m g(z_m)}{m} \\
&= 0
\end{aligned}
$$

達成の例: $\mathcal{G}=\{g\}$ のとき (Ex 3.6(a))

*斉次性 (Ex 3.8(a))

$\alpha \in \mathbb{R}$ として $\widehat{\mathfrak{R}}_S(\alpha\mathcal{G}) = |\alpha|\widehat{\mathfrak{R}}_S(\mathcal{G})$. 但し $\alpha\mathcal{G} = \{\alpha g \mid g \in \mathcal{G}\}$

$$
\begin{aligned}
\because \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\frac{\sigma_i \alpha g(z_i)}{m}\right] &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\frac{\sigma_i |\alpha| \mathrm{sign}(\alpha) g(z_i)}{m}\right] \\
&= |\alpha|\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\frac{\sigma_i g(z_i)}{m}\right]
\end{aligned}
$$

σ_i と $\sigma_i \mathrm{sign}(\alpha)$ は同様に分布するので成り立つ.

ラデマッハ複雑度

定義 3.2 (ラデマッハ複雑度)

$\mathcal{D}$ を標本が抽出される分布とする．**ラデマッハ複雑度**とは $\mathcal{D}$ に従って抽出されたサイズ m の標本集合全体における経験ラデマッハ複雑度の期待値である．

$$\mathfrak{R}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\widehat{\mathfrak{R}}_S(\mathcal{G})] \tag{3.2}$$

以下ではラデマッハ複雑度のことを RC と書くことがある．

平均ラデマッハ複雑度ということもある．

*計算の例

例題 (同心球)

入力 $\mathcal{X} = \mathbb{R}$, ラベル $\mathcal{Y} = \{-1, 1\}$, 入力の従う分布 $\mathcal{D} = \mathcal{N}(0, 1)$,
真の関数 $c : x \mapsto \mathrm{sgn}(1 - x^2)$,
仮説 $\mathcal{H} = \{h_\theta : x \mapsto \mathrm{sgn}(\theta^2 - x^2)\}_{\theta \in \mathbb{R}}$
損失関数は $L : (a, b) \mapsto \mathbb{I}\{a \neq b\}$ とする. このラデマッハ複雑度 $\mathfrak{R}_m(\mathcal{G})$ を計算せよ.
但し $\mathbb{I}\{\cdot\}$ は指示関数で $\mathcal{G}$ は仮説 $\mathcal{H}$ に関する損失関数の族.

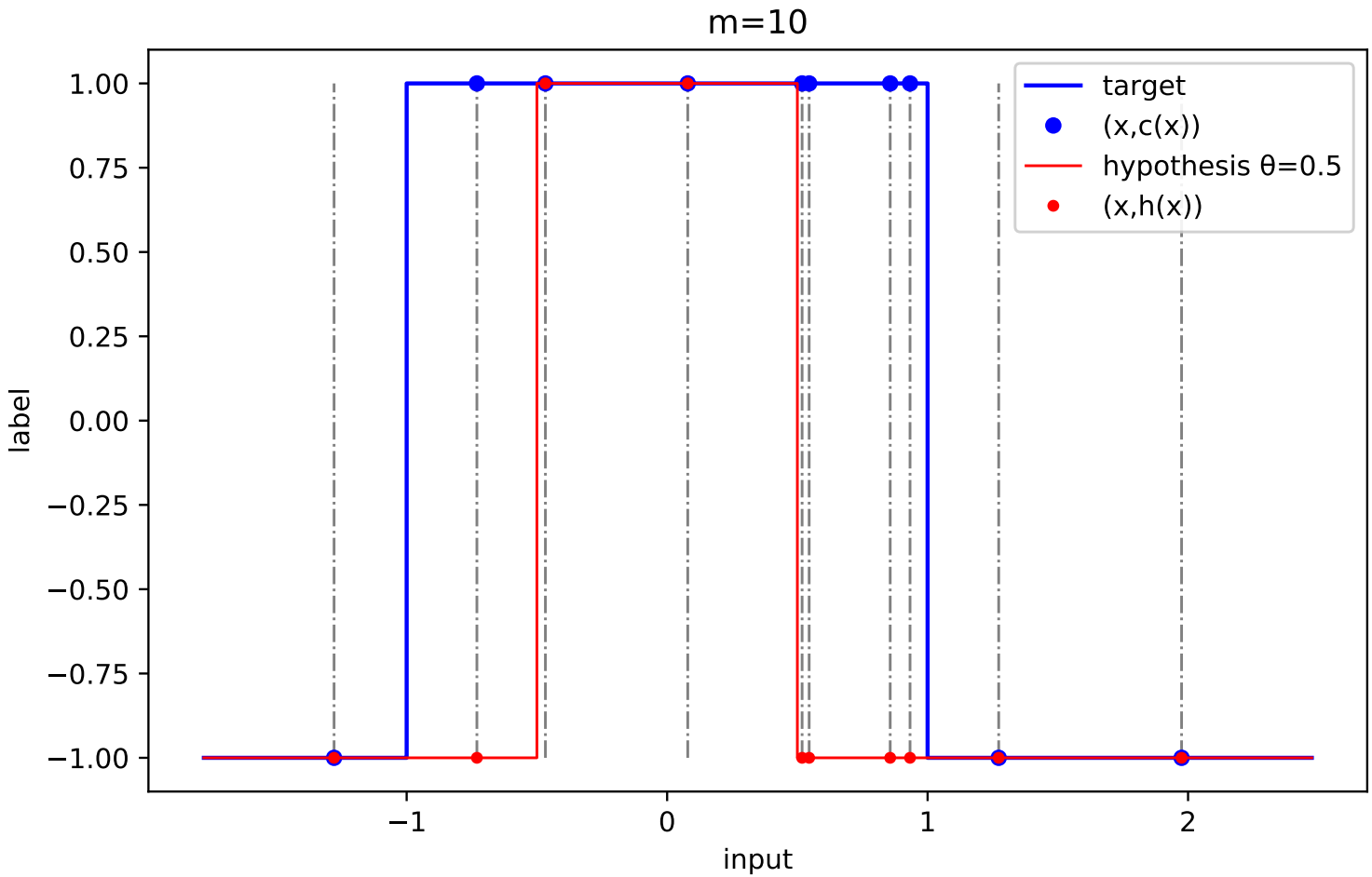

Figure 1: 問題の例.

*解答

$x^2 = y$ とおくことにより問題は次のように置き換えられる.

- 入力 $\mathcal{X}' = \mathbb{R}_{\geq 0}$
- 入力の従う分布 $\mathcal{D}' = \chi^2_1 = \mathrm{Ga}(1/2, 2)$
- 目的関数 $c' : y \mapsto \mathrm{sgn}(1 - y)$
- 仮説 $\mathcal{H}' = \{h'_\theta : y \mapsto \mathrm{sgn}(\theta - y)\}_{\theta \geq 0}$

はじめに経験ラデマッハ複雑度を考える. 損失関数は

$$\begin{aligned} g_\theta(y_i) &= L(h'_\theta(y_i), c'(y_i)) \\ &= \mathbb{I}\{h'_\theta(y_i) \neq c'(y_i)\} \\ &= \mathbb{I}\{\theta \leq y_i < 1 \ \vee \ 1 \leq y_i < \theta\} \end{aligned}$$

$\widehat{\mathfrak{R}}_S(\mathcal{G})$ を計算する際 σ_i が独立であることなどから y_i の順番は問われないし $y_i = y_j (i \neq j)$ となる確率は 0 なので, 以下では $y_1 < y_2 < \cdots < y_m$ として良い.

p, q をそれぞれ $y_p < \theta$, $y_q < 1$ となる最大の添字とすると

$$\begin{bmatrix} \boldsymbol{y}^\top \\ \boldsymbol{g}_S^\top \\ \boldsymbol{\sigma}^\top \end{bmatrix} = \begin{bmatrix} y_1 & \cdots & y_p & y_{p+1} & \cdots & y_q & y_{q+1} & \cdots & y_m \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \sigma_1 & \cdots & \sigma_p & \sigma_{p+1} & \cdots & \sigma_q & \sigma_{q+1} & \cdots & \sigma_m \end{bmatrix}$$

という対応か

$$\begin{bmatrix} \boldsymbol{y}^\top \\ \boldsymbol{g}_S^\top \\ \boldsymbol{\sigma}^\top \end{bmatrix} = \begin{bmatrix} y_1 & \cdots & y_q & y_{q+1} & \cdots & y_p & y_{p+1} & \cdots & y_m \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \sigma_1 & \cdots & \sigma_q & \sigma_{q+1} & \cdots & \sigma_p & \sigma_{p+1} & \cdots & \sigma_m \end{bmatrix}$$

という対応になっている. $\sup_{g \in \mathcal{G}}$ を考えるとき θ を動かすことを考えるが, これは p を動かすのと一緒.

よって $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\frac{1}{m} \sup_p \{ \sum_{i=p+1}^{q} \sigma_i, \sum_{i=q+1}^{p} \sigma_i \} \right]$ となる.

q 以下と $q+1$ 以上は独立で考えられる.

$\longrightarrow$ それぞれの最大値の分布を考える.
$=$ 一次元ランダムウォークの最大位置の分布

q 歩の対称ランダムウォークで最大位置が r である確率は

$$P(q,r) = \frac{1}{2^q} \begin{pmatrix} q \\ \lfloor \frac{q-r}{2} \rfloor \end{pmatrix}$$

証明は略.[1]

例えば $m = 5, q = 2$ の場合

個数	$q+1$ 以上	3	3	1	1
q 以下	最大値	0	1	2	3
2	0	0	1	2	3
1	1	1	1	2	3
1	2	2	2	2	3

このとき

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = (0 \cdot 6 + 1 \cdot 12 + 2 \cdot 10 + 3 \cdot 4)/32/5 = 44/32/5 = 11/40$$

実は経験 RC は q に依存しない.　　$\therefore \mathfrak{R}_5(\mathcal{G}) = 11/40$

よって $\mathfrak{R}_m(\mathcal{G}) = \frac{1}{m}\sum_{r=0}^{m} rP(m,r)$ となる. この計算[1] は $P(m,r)$ が二項分布の pmf の並べ替えという事実を利用すると良い.
答えは

$$\mathfrak{R}_{2k+1}(\mathcal{G}) = \frac{(2k-1)!!}{(2k)!!} - \frac{1}{4k+2} \quad (k = 0, 1, 2, \ldots)$$

$$\mathfrak{R}_{2k}(\mathcal{G}) = \frac{(2k-1)!!}{(2k)!!} - \frac{1}{4k} + \frac{(2k-1)!!}{(2k)!!4k} \quad (k = 1, 2, 3, \ldots)$$

[1] 証明は補足資料を参照.

定理 3.3

定理 3.3

$\mathcal{G}$ を $\mathcal{Z}$ から $[0,1]$ への関数の族とする. 任意の $\delta > 0$ に対して, 少なくとも $1-\delta$ の確率で, m 個の独立同分布標本 S 上, 任意の $g \in \mathcal{G}$ について以下が成り立つ.

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.3}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3.4}$$

定理 3.3 の証明

証明: 任意の標本集合 S と $g \in \mathcal{G}$ に対して $\widehat{\mathbb{E}}_S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$ とおく. また S の関数 Φ を

$$\Phi(S) = \sup_{g\in\mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right) \tag{3.5}$$

とおく.S に対し一つだけ要素が異なる集合 S' を用意する. このとき

$$\begin{aligned} \Phi(S') \quad - \quad & \Phi(S) \\ = \quad & \sup_{g\in\mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] + \frac{g(z_i) - g(z_i')}{m} \right) - \sup_{g\in\mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right) \\ \leq \quad & \sup_{g\in\mathcal{G}} \frac{g(z_i) - g(z_i')}{m} \leq \frac{1}{m} \end{aligned} \tag{3.6}$$

上は S, S' を入れ替え出来るので $|\Phi(S) - \Phi(S')| \leq 1/m$ を得る.

マクダイアミッドの不等式

ここでマクダイアミッドの不等式について掲載する.

マクダイアミッドの不等式

$(X_1, \ldots, X_m) \in \mathcal{X}^m$ を独立な確率変数とする. 関数 $f : \mathcal{X}^m \to \mathbb{R}$ について, ある $c_1, \ldots, c_m > 0$ が存在して

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i$$

が成り立つとき, 任意の $\epsilon > 0$ に対して次の不等式が成り立つ.

$$\mathbb{P}[|f(S) - \mathbb{E}[f(S)]| \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right) \tag{D.16}$$

但し S は $X_1, \cdots, X_m$ の略記.

f に Φ,c_i に $1/m$ を当てはめて (D.16) を適応すると $\epsilon > 0$ を任意として

$$\mathbb{P}[|\Phi(S) - \mathop{\mathbb{E}}_{S}[\Phi(S)]| \geq \epsilon] \leq \exp(-2m\epsilon^2)$$

この式の右辺を $\delta/2$ とおき ϵ について解くと

$$\mathbb{P}\left[\left|\Phi(S) - \mathop{\mathbb{E}}_{S}[\Phi(S)]\right| \geq \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \leq \frac{\delta}{2}$$

すなわち少なくとも $1 - \delta/2$ の確率で

$$\Phi(S) \leq \mathop{\mathbb{E}}_{S}[\Phi(S)] + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3.7}$$

が成り立つ.

次に $\Phi(S)$ の期待値について上界を見積もる.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S}[\Phi(S)] &= \mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]\right)\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\mathop{\mathbb{E}}_{S'}\left[\left(\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]\right)\right]\right] && (3.8)\\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{G}}\left(\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]\right)\right] && (3.9)\\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\left(g(z_i')-g(z_i)\right)\right] && (3.10)
\end{aligned}
$$

ここでの S' は S とは関係がないことに注意.

$\boldsymbol{\sigma}$ を固定する.$S = (z_i)_{i=1}^m, S' = (z'_i)_{i=1}^m$ に対して

$$P(\boldsymbol{\sigma}, S, S') = \left(\prod_{i=1}^m p(z_i)p(z'_i) \right) \sup_{g \in \mathcal{G}} \sum_{i=1}^m \sigma_i \left(g(z'_i) - g(z_i) \right)$$

を考える. $T = (w_i)_{i=1}^m, T' = (w'_i)_{i=1}^m$ を

$$(w_i, w'_i) = \begin{cases} (z_i, z'_i) & (\sigma_i = 1) \\ (z'_i, z_i) & (\sigma_i = -1) \end{cases}$$

と定めると $P(\boldsymbol{\sigma}, S, S') = P((1)_{i=1}^m, T, T')$ である.
逆に T, T' から見て $P(\boldsymbol{\sigma}, T, T') = P((1)_{i=1}^m, S, S')$ である.
つまり S, S' について平均をとっていれば $g(z'_i) - g(z_i)$ の符号は任意に変えて良い.

このことから

$$
\begin{aligned}
(3.10) \quad &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S,S'} \left[\sup_{g\in\mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(g(z_i') - g(z_i) \right) \right] && (3.11) \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S'} \left[\sup_{g\in\mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i') \right] \\
&+ \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S} \left[\sup_{g\in\mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(z_i) \right] && (3.12) \\
&= 2 \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S} \left[\sup_{g\in\mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right] = 2\mathfrak{R}_m(\mathcal{G}) && (3.13)
\end{aligned}
$$

(3.7) で $\delta/2$ の代わりに δ を使い, それと (3.13) から (3.3) を得る.

次に (3.4) を証明する. これには (3.7) について $\Phi(S)$ の代わりに $\widehat{\mathfrak{R}}_S(\mathcal{G})$ を適応できることを確かめれば良い.(3.6) と同様に S, S' を互いに一つだけ要素が異なる集合とすると

$$\begin{aligned}\widehat{\mathfrak{R}}_{S'}(\mathcal{G}) - \widehat{\mathfrak{R}}_S(\mathcal{G}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_S + \sigma_i(g(z_i') - g(z_i))}{m} - \sup_{g\in\mathcal{G}} \frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_S}{m}\right] \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\right] = \frac{1}{m}\end{aligned}$$

従ってマクダイアミッドの不等式から少なくとも $1-\delta/2$ の確率で

$$\mathfrak{R}_m(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3.14}$$

が成り立つ.

$\mathbb{P}(\overline{(3.7)} \cup \overline{(3.14)}) \leq \mathbb{P}(\overline{(3.7)}) + \mathbb{P}(\overline{(3.14)}) \leq \delta$ であることから,
少なくとも確率 $1-\delta$ で (3.7) と (3.14) の両方が成り立つ.
従って (3.7),(3.13),(3.14) から

$$\Phi(S) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3.15}$$

が少なくとも確率 $1-\delta$ で成り立ち, これは (3.4) に一致する. □

定理 3.3 においては $\mathcal{G}$ は単なる $\mathcal{Z}$ から $[0,1]$ への関数族だった.

次に $\mathcal{G}$ が仮説集合 $\mathcal{H}$ に関する $0-1$ 損失関数の族である場合を考える.

$\mathcal{G}$ と $\mathcal{H}$ の RC の関係についてみる.

補題 3.4

補題 3.4

$\mathcal{H}$ を $\{-1,1\}$ を値にとる関数族とし $\mathcal{G}$ を $\mathcal{H}$ に関する $0-1$ 損失関数の族, すなわち $\mathcal{G}=\{(x,y)\mapsto \mathbb{I}\{h(x)\neq y\}\mid h\in\mathcal{H}\}$ とする. また $\mathcal{X}\times\{-1,1\}$ の元からなる任意の標本 $S=((x_i,y_i))_{i=1}^m$ について,$S_{\mathcal{X}}$ を空間 $\mathcal{X}$ への射影, すなわち $S_{\mathcal{X}}=(x_i)_{i=1}^m$ とする. このとき $\mathcal{H},\mathcal{G}$ 間の経験 RC について次の関係が成り立つ.

$$\widehat{\mathfrak{R}}_S(\mathcal{G})=\frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H}) \tag{3.16}$$

補題 3.4 の証明

証明:

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\mathbb{I}\{h(x_i)\neq y_i\}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{1-y_ih(x_i)}{2}\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_iy_ih(x_i)\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_ih(x_i)\right] = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H})
\end{aligned}
$$

□

補題に関する補足

(3.16) に対し期待値をとることで $\mathfrak{R}_m(\mathcal{G}) = \frac{1}{2}\mathfrak{R}_m(\mathcal{H})$ を得る.

$\rightarrow$ この関係を使って, 仮説集合 $\mathcal{H}$ の RC の観点から二値分類の汎化誤差の上界を得ることができる.

定理 3.5

定理 3.5 (RC による上界-二値分類)

$\mathcal{H}$ を $\{-1, 1\}$ を値にとる関数族とし $\mathcal{D}$ を入力空間 $\mathcal{X}$ 上の分布とする. 任意の $\delta > 0$ に対して, 少なくとも $1 - \delta$ の確率で, $\mathcal{D}$ から抽出された m 個の標本 S 上, 任意の $h \in \mathcal{H}$ に対して以下が成り立つ.

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{3.17}$$

$$R(h) \leq \widehat{R}_S(h) + \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{3.18}$$

証明: 定理 3.3 と補題 3.4 から直ちに得られる. □

定理 3.5 の補足

(3.18) についてはデータ依存

$\rightarrow$ $\widehat{\mathfrak{R}}_S(\mathcal{H})$ が計算できるのであれば非常に有益

$\rightarrow$ どうやれば経験 RC を計算出来るのか

RC の問題点

$\boldsymbol{\sigma}$ と $-\boldsymbol{\sigma}$ は同様に分布するので

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(x_i)\right] = -\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\inf_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right]$$

$\inf_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)$ の計算は**経験損失最小化問題**と等価

$\rightarrow$　これはいくつかの仮説集合に対して計算量的に困難

$\rightarrow$　RC をより計算し易い組み合わせ測度と結びつける

成長関数の定義

ここから RC が成長関数によって抑えられることを示す.

定義 3.6 (成長関数)

仮説集合 $\mathcal{H}$ に対する成長関数 $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ は次で定義される.

$$\Pi_{\mathcal{H}} : m \mapsto \max_{\{x_1,\ldots,x_m\} \subset X} |\{(h(x_1), \ldots, h(x_m)) \mid h \in \mathcal{H}\}| \tag{3.19}$$

以下では成長関数を GF と書くことがある.

GF の意味

$\Pi_{\mathcal{H}}(m)$=“ $\mathcal{H}$ を使って m 個の点を分類する異なる方法の最大数 ”

異なる分類の仕方のそれぞれを**二分**という

→GF は仮説によって実現可能な二分の個数を数えている

VC との違い

GF は RC と同様に $\mathcal{H}$ についての豊かさの指標を与える

概念	分布依存
VC	有り
GF	なし

GF は真に組み合わせ的な量.

*例題

例題 (同心球)

入力 $\mathcal{X} = \mathbb{R}$, ラベル $\mathcal{Y} = \{-1, 1\}$,
仮説 $\mathcal{H} = \{h_\theta : x \mapsto \mathrm{sgn}(\theta^2 - x^2)\}_{\theta \in \mathbb{R}}$
この成長関数 $\Pi_{\mathcal{H}}(m)$ を計算せよ.

Ex 3.2 の類題. 図は 1 を参照.

*解答

$x_i^2 = y_i$ とおく. 更には $y_1 < y_2 < \cdots < y_m$ として良い.

θ より小さければ $h(y_i) = +1$, そうでなければ $h(y_i) = -1$ となる.

$\longrightarrow$ y_i を $\circ$ とおき θ を $|$ とおいたときの順列と同じ. よって

$$\Pi_{\mathcal{H}}(m) = \binom{m+1}{1} = m + 1$$

マッサールの補題

定理 3.7 (マッサールの補題)

$\mathcal{A} \subset \mathbb{R}^m$ を有限集合とする. $r = \max_{\boldsymbol{x} \in \mathcal{A}} \|\boldsymbol{x}\|_2$ として次が成り立つ.

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\frac{1}{m} \sup_{\boldsymbol{x} \in \mathcal{A}} \sum_{i=1}^{m} \sigma_i x_i \right] \leq \frac{r\sqrt{2 \log |\mathcal{A}|}}{m} \tag{3.20}$$

但し $\boldsymbol{\sigma}$ はラデマッハ変数.

証明には最大不等式 (D.10) とヘフディングの補題 (D.2) を用いる.

最大不等式

最大不等式

$(X_j)_{j=1}^n$ を確率変数とし, 任意の $t > 0$ について, ある $r > 0$ があって $\mathbb{E}[\exp(tX_j)] \leq \exp\left(\frac{t^2 r^2}{2}\right)$ を満たすとき次の不等式が成り立つ.

$$\mathbb{E}\left[\max_{j\in[n]} X_j\right] \leq r\sqrt{2\log n} \tag{D.10}$$

ヘフディングの補題

ヘフディングの補題

X を平均 0 で $[a, b]$ を値にとる確率変数とすると $(a < b)$, 任意の $t > 0$ について次の不等式が成り立つ.

$$\mathbb{E}[\exp(tX)] \leq \exp\left\{\frac{t^2(b-a)^2}{8}\right\} \tag{D.2}$$

補題 3.7 の証明

証明: ヘフディングの補題を用いて最大不等式の条件を満たすことを確認すれば良い.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\exp\left\{t\sum_{i=1}^{m}\sigma_i x_{(j),i}\right\}\right] &= \prod_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma_i}[\exp(t\sigma_i x_{(j),i})] \\
&\leq \prod_{i=1}^{m}\exp\left\{\frac{t^2 x_{(j),i}^2}{2}\right\} \\
&\leq \exp\left\{\frac{t^2\left(\max_{\boldsymbol{x}\in\mathcal{A}}\|\boldsymbol{x}\|_2\right)^2}{2}\right\}
\end{aligned}
$$

□

(D.11) を用いた証明について

FML では次のバージョンの最大不等式を用いた証明を試みている.

最大不等式 2

固定された $j \in [n]$ について,Y_{ij} は平均 0 で $[-r_i, r_i]$ を値にとる独立な確率変数とする ($r_i > 0$). このとき次の不等式が成り立つ.

$$\mathbb{E}\left[\max_{j\in[n]} \sum_{i=1}^{m} Y_{ij}\right] \leq r\sqrt{2\log n} \tag{D.11}$$

但し $r = \sqrt{\sum_{i=1}^{m} r_i^2}$

一見これを使っても (3.20) を証明できそうだが, 実は
$r = \max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_2$ ではなくなる.

FML の証明では $Y_{ij} = \sigma_i x_i$ としているが...

$\rightarrow$ j として x_i そのものが変わるので $r_i = \max_{\boldsymbol{x} \in \mathcal{A}} |x_i|$ としないと (D.11) の条件は満たせない.

$$\max_{\boldsymbol{x} \in \mathcal{A}} \|\boldsymbol{x}\|_2^2 \leq \sum_{i=1}^{m} \max_{\boldsymbol{x} \in \mathcal{A}} |x_i|^2 = \sum_{i=1}^{m} r_i^2 = r^2$$

系 3.8

系 3.8

$\mathcal{G}$ を $\{-1, 1\}$ を値にとる関数族とすると次が成り立つ.

$$\mathfrak{R}_m(\mathcal{G}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(m)}{m}} \tag{3.21}$$

系 3.8 の証明

証明: 固定された $S = (x_i)_{i=1}^m$ に対し $\mathcal{G}_{|S}$ を S についての関数値ベクトル $(g(x_i))_{i=1}^m$ の集合とする. g は $\{-1, 1\}$ を値にとるので関数値ベクトルの二乗ノルムは $\sqrt{m}$ である. マッサールの補題と GF の定義から

$$\begin{aligned}
\mathfrak{R}_m(\mathcal{G}) &= \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{u\in\mathcal{G}_{|S}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i u_i\right]\right] \\
&\leq \mathop{\mathbb{E}}_{S}\left[\frac{\sqrt{2m\log|\mathcal{G}_{|S}|}}{m}\right] \\
&\leq \mathop{\mathbb{E}}_{S}\left[\frac{\sqrt{2m\log\Pi_{\mathcal{G}}(m)}}{m}\right] = \sqrt{\frac{2\log\Pi_{\mathcal{G}}(m)}{m}}
\end{aligned}$$

□

*Ex 3.5

証明の式の 2 行目で Jensen の不等式を使って $\mathbb{E}_S[|\mathcal{G}_{|S}|]$ による上界が得られる.

$$\mathfrak{R}_m(\mathcal{G}) \leq \mathop{\mathbb{E}}_{S}\left[\frac{\sqrt{2m\log|\mathcal{G}_{|S}|}}{m}\right] \leq \sqrt{\frac{2\log\mathbb{E}_S[|\mathcal{G}_{|S}|]}{m}}$$

問題の方では $|\mathcal{G}_{|S}| = \Pi(\mathcal{G}, S)$ と表示

こちらの方がタイト

系 3.9

定理 3.5 と系 3.8 を使うと,GF による汎化誤差の上界を得る.

系 3.9 (GF による汎化誤差の上界)

$\mathcal{H}$ を $\{-1, 1\}$ を値にとる関数族とする. 任意の $\delta > 0$ に対して, 次が少なくとも $1 - \delta$ の確率で, 任意の $h \in \mathcal{H}$ に対して成り立つ.

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{3.22}$$

GF による汎化誤差の上界 2

GF による上界は RC を介さなくても直接得ることができる.

$$\mathbb{P}\left[\left|R(h) - \widehat{R}_S(h)\right| > \epsilon\right] \leq 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right) \tag{3.23}$$

これは (3.22) と定数程度の差しかない.

証明は [2] を参照. 但し一部証明が省略されている.

また追加の条件がある.$(m \geq 2/\epsilon^2)$

*省略箇所の説明

Vapnik らは [2] で (3.23) の証明において

$$\Gamma = \sum_{k \ \ s.t. \ \ |2k/m - p/m| > \epsilon/2} \frac{\binom{p}{k}\binom{2m-p}{m-k}}{\binom{2m}{m}} \leq 2\exp\left\{-\frac{m\epsilon^2}{8}\right\}$$

の証明を省いている. 但し記法は FML に合わせた.

この証明[2] のアイデアから一般にヘフディングの不等式が非復元抽出の場合にも適応できることが分かる.[3]

[2]補足資料を参照.

GF の問題点

GF の計算はいつも便利というわけではない
(定義から任意の m について $\Pi_{\mathcal{H}}(m)$ の計算が必要)

$\rightarrow$VC 次元の導入 (次回)

- 単一スカラに基づいた代替的な仮説の複雑さの測度
- 成長関数の振る舞いに深い関係

*まとめ

RC,GF による汎化誤差の上界をまとめる.

$$R(h) - \widehat{R}(h) - \sqrt{\frac{\log 1/\delta}{2m}} \leq \mathfrak{R}_m(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}}$$

が少なくとも $1-\delta$ の確率で $^\forall h \in \mathcal{H}$ について成り立つ.

*おまけ:同心球クラスの RC と GF

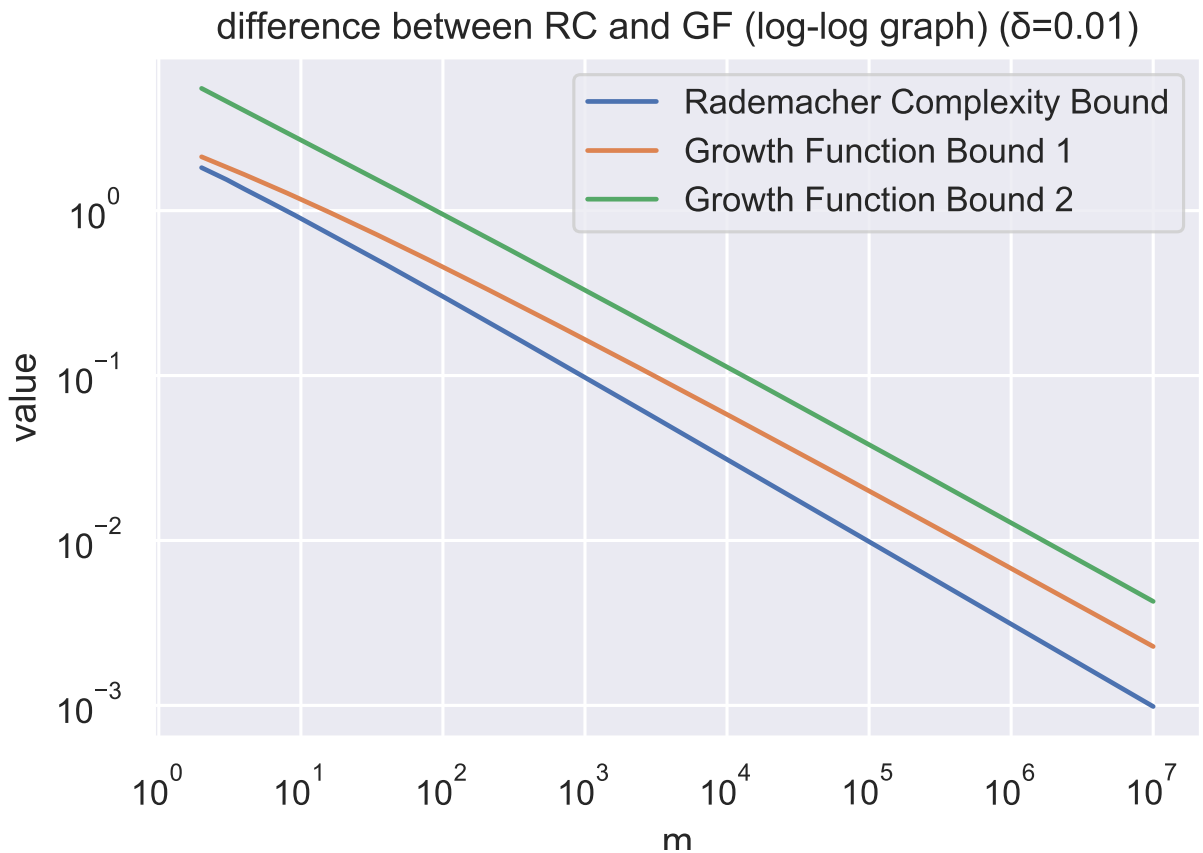

Figure 2: RC 上界と GF 上界の両対数グラフ.$\mathrm{RCB} = \sqrt{\frac{\log 1/\delta}{2m}} + \mathfrak{R}_m(\mathcal{H})$,
$\mathrm{GFB1} = \sqrt{\frac{\log 1/\delta}{2m}} + \sqrt{\frac{2\log \Pi_{\mathcal{H}}(m)}{m}}$, $\mathrm{GFB2} = \sqrt{\frac{8}{m}\log\frac{4\Pi_{\mathcal{H}}(2m)}{\delta}}$, ($\delta = 0.01$ のとき)

参考文献

[1] Kyle Siegrist, “11.6: The Simple Random Walk”, STATISTICS LibreTexts, https://stats.libretexts.org/Bookshelves/Probability_Theory/Book%3A_Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/11%3A_Bernoulli_Trials/11.06%3A_The_Simple_Random_Walk , Aug 10 2020(cited Jun 4 2021)

[2] V.N.Vapnik and A. Ya.Chervonenkis, Translated by B. Seckler “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities”, Theory of Probability and Its Applications, Vol. 16, No. 2, 1971, pp.264-280.

[3] Remi Bardenet and Odalric-Ambrym Maillard, “Concentration inequalities for sampling without replacement”, Bernoulli, Vol. 21, No. 3, 2015, pp.1361-1385.