統計的学習理論読み（Chapter 2）

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 2 章のまとめである.

- 仮説空間の複雑さの指標: VC 次元, ラデマッハ複雑度
  - 時間があればカバリングナンバーも抑えたい（ちょっとだけ入れた）
- 一様大数の法則による汎化誤差の上界評価

がメイントピック

Table of contents

仮説集合の複雑度

仮説集合の複雑度

2.1 VC 次元

問題設定 I

設定 &Notation

- 2 値判別 ($|\mathcal{Y}| = 2$)
- $\mathcal{H} := \{h : \mathcal{X} \longrightarrow \mathcal{Y}\}$: 仮説集合
- 入力 $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$ に対して, $\mathcal{H}$ の元で予測されるラベルの組の集合の要素数を考察:

$$\Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) := \left|\{(h(\boldsymbol{x}_1), ..., h(\boldsymbol{x}_n)) \in \mathcal{Y}^n; h \in \mathcal{H}\}\right|$$

Definition 1 (Growth function, Foundations of Machine Learning Def 3.3)

仮説集合 $\mathcal{H}$ の *growth function* $\hat{\Pi}_{\mathcal{H}} : \mathbb{N} \longrightarrow \mathbb{N}$ は以下で定義される.

$$\forall n \in \mathbb{N}, \hat{\Pi}_{\mathcal{H}}(n) := \max_{\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathcal{X}} \Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$$

$\Pi_{\mathcal{H}}$ の性質

- 定義より

$$\Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \leq |\mathcal{Y}^n| = 2^n$$

- $\Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = 2^n$

$$\iff \forall\{(\boldsymbol{x}_i, y_i)\},\ \exists h \in \mathcal{H} \ \text{ s.t. } \ h(\boldsymbol{x}_i) = y_i$$

  ラベルの組合せを網羅すれば 100% データを分類する仮説が取れる
- 一方, data 数 n が増大するとラベルの組合せが膨大となり, $\mathcal{H}$ の元で網羅できなくなる.
  $\longrightarrow$ この境界を VC 次元という

Definition 2 (VC 次元)

仮説空間 $\mathcal{H}$ の VC 次元は以下で定義される

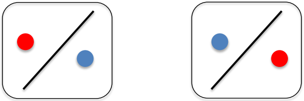
$$VC_{dim}(\mathcal{H}) := \max\{n \in \mathbb{N}; \hat{\Pi}_{\mathcal{H}}(n) = 2^n\}$$

- 仮説空間 $\mathcal{H}$ の元でラベルの組合せを網羅できる最大の data 数が VC 次元
- $\forall n \in \mathbb{N}$, $\exists \boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathcal{X}$ s.t. $\Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = 2^n$ のとき, $VC_{dim}(\mathcal{H}) = \infty$ と定義
- 仮説集合がどんなラベル付にも対応できる $\rightarrow$ ノイズにも fitting する
  cf re-thinking generalization 論文 [ICLR2017]?

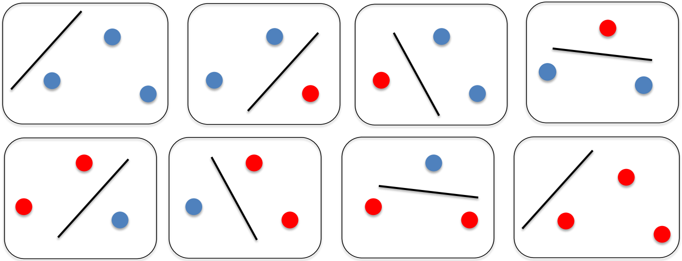
VC 次元, 例

- $\mathcal{H}$: 2 次元直線のとき, $VC_{dim}(\mathcal{H}) = 3$

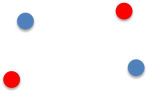
データ数2のとき



データ数3のとき



データ数4のとき



どんな直線でも分離できない
→ ラベルを網羅できない

data 数 n が data の次元 d より大きいとき, growth function は d の多項式オーダーになることを保証

Lemma 1 (Sauer's Lemma (Lemma 2.1))

- $|\mathcal{Y}| = 2$,
- $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$,
- $VC_{dim}(\mathcal{H}) = d$

このとき, $n \geq d$ に対して

$$\Pi_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d = \mathcal{O}(n^d)$$

Sauer's Lemma II

(proof) Thm 3.5 of Foundations of Machine Learning

$$
\begin{aligned}
\Pi_{\mathcal{H}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \underbrace{\leq}_{(\diamond)} \sum_{i=0}^{d} \binom{n}{i} &\leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \\
&\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \\
&= \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \left(\frac{n}{d}\right)^{i} \left(\frac{d}{n}\right)^{i} \\
&= \left(\frac{n}{d}\right)^{d} \underbrace{\sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i}}_{(1+\frac{d}{n})^n} \\
&= \left(\frac{n}{d}\right)^{d} \left(1 + \frac{d}{n}\right)^{n} \\
(\because) \left(1 + \frac{d}{n}\right)^{n} \to e^d \quad \text{as } n \to \infty \to \quad &\leq \left(\frac{n}{d}\right)^{d} e^d \quad \square
\end{aligned}
$$

($\diamond$) の証明 : $n+d$ に関する帰納法で示す.

- $n=1$, $d=0$ or $d=1$ のときは自明
- $n-1$, $d-1$ or d のとき成立つと仮定

Notation

- $S=\{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$: fixed sample set with $\hat{\Pi}_{\mathcal{H}}(m)$ dichotomies
  ($\mathcal{H}$ の元で説明可能なラベル付けの組合せが $\hat{\Pi}_{\mathcal{H}}(m)$ 個存在)
- $G=\mathcal{H}|_S$: domain を S に制限した仮説集合
- $S'=S\backslash\{\boldsymbol{x}_n\}$ として,

$$G_1 = G|_{S'}$$

$$G_2 = G\backslash G_1$$

定義から明らかに $G_1\cup G_2=G$, $G_1\cap G_2=\emptyset$, $|G_1|+|G_2|=|G|$

Sauer's Lemma IV

e.g. $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ のとき, ラベルパターンは 8 通り

Table 1: 8 通りのラベル組合せを 8 つの仮説で実現

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$
h_1	1	1	1
h_2	0	1	1
h_3	1	0	1
h_4	1	1	0
h_5	0	0	1
h_6	0	1	0
h_7	1	0	0
h_8	0	0	0

Table 2: 仮説を $S' = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ 上に制限

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$
$h_1\|_{S'}$	1	1
$h_2\|_{S'}$	0	1
$h_3\|_{S'}$	1	0
$h_5\|_{S'}$	0	0

例えば, $h_1|_{S'} = h_4|_{S'}$ となるが, こういう場合はどちらか一方を G_1 の元とする

これより $G_1 = \{h_1|_{S'}, h_2|_{S'}, h_3|_{S'}, h_5|_{S'}\}$,
$G_2 = \{h_4|_{S'}, h_6|_{S'}, h_7|_{S'}, h_8|_{S'}\}$ とすると $G = G_1 \cup G_2$, $G_1 \cap G_2 = \emptyset$.

- $VC_{dim}(G_1) \le VC_{dim}(G) \le VC_{dim}(\mathcal{H}) \le d$ より,

$$|G_1| \underbrace{\le}_{(\sharp)} \hat{\Pi}_{G_1}(n-1) \underbrace{\le}_{(\sharp\sharp)} \sum_{i=0}^{d} \begin{pmatrix} n-1 \\ i \end{pmatrix}$$

  ここで,
  - $(\sharp)$: by def of growth function
    G_1 の具体形: $G_1 = \{(h(\boldsymbol{x}_1), ..., h(\boldsymbol{x}_{n-1})); h \in \mathcal{H}\}$ であり, この要素数の max を取ったものが growth function だった.
  - $(\sharp\sharp)$: 帰納法の仮定
- さらに, $Z \subset S'$ の取りうるラベルの組合せが G_2 で網羅される ("Z は G_2 で shatter される" という) ならば, $Z \cup \{\boldsymbol{x}_n\}$ は G で shatter される.
  e.g. 先の例で, $S' = \{\boldsymbol{x}_1, \boldsymbol{x}_2\} = Z$ とおくと, Z は
  $G_2 = \{h_4, h_6, h_7, h_8\}$ で shatter され, $S = S' \cup \{\boldsymbol{x}_3\}$ は
  $G = G_1 \cup G_2$ で shatter される

従って

$$VC_{dim}(G_2) \leq VC_{dim}(G) - 1 = d - 1$$

- G_2 が網羅できるラベルの組合せ数は, G が網羅できるラベルの組合せ数より真に小さい.

また, G_1 のときと全く同様の論法で

$$|G_2| \leq \hat{\Pi}_{G_2}(n-1) \leq \sum_{i=0}^{d} \binom{n-1}{i}$$

が成立.

以上の議論より,

$$
\begin{aligned}
|G| = |G_1| + |G_2| &\leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{d} \left\{ \binom{n-1}{i} + \binom{n-1}{i-1} \right\} \\
&= \sum_{i=0}^{d} \binom{n}{i}
\end{aligned}
$$

より (n, d) の場合が示された. □

VC 次元による汎化誤差の一様上界 I

Theorem 1 (Theorem 2.2)

- $\mathcal{H} \subset \{h : \mathcal{X} \to \{+1, -1\}\}$
- $VC_{dim}(\mathcal{H}) = d < \infty$
- *training data :* $(X_i, Y_i) \sim_{i.i.d} \mathcal{D}$
- $0 - 1$ *loss*

$n \geq d$ のとき,

$$P_D\left[\sup_{h \in \mathcal{H}} |R_{err}(h) - \hat{R}_{err}(h)| \leq 2\sqrt{\frac{2d}{n}\log\frac{en}{d}} + \sqrt{\frac{\log 2/\delta}{2n}}\right] \geq 1 - \delta$$

が成立

以下, Thm 2.2 を用いて学習した仮説の汎化誤差を評価
($|\mathcal{H}| = \infty$ なる状況も考える)

VC 次元による汎化誤差の一様上界 II

設定

- $S = \{(X_i, Y_i)\}_{i=1}^{n}$: observed data
- $h_S = \arg\min_{h \in \mathcal{H}} \hat{R}_{err}(h)$: 最小経験誤差を達成する仮説
- $h_0 \in \mathcal{H}$: $\mathcal{H}$ は Bayes rule を含むと仮定

以下は定義から明らか:

$$\hat{R}_{err}(h_S) \leq \hat{R}_{err}(h_0)$$
$$R_{err}(h_0) \leq R_{err}(h_S)$$

Q : h_S の汎化誤差 $R_{err}(h_S)$ のバウンド?

$\longrightarrow$ Thm 2.2 より, 経験誤差 + f(VC 次元, データ数) で押さえられる

VC 次元による汎化誤差の一様上界 III

One of the most important results in learning theory

(by Bottou et al. "Optimization Methods for Large-Scale Machine Learning")

$$
\begin{aligned}
R_{err}(h_S) &\le R_{err}(h_S) + \underbrace{\hat{R}_{err}(h_0) - \hat{R}_{err}(h_S)}_{\ge 0} \\
&= R_{err}(h_0) - R_{err}(h_0) + \hat{R}_{err}(h_0) + R_{err}(h_S) - \hat{R}_{err}(h_S) \\
&\le R_{err}(h_0) + \sup_{h \in \mathcal{H}} |R_{err}(h) - \hat{R}_{err}(h)| \\
&\qquad + \sup_{h \in \mathcal{H}} |R_{err}(h) - \hat{R}_{err}(h)| \\
&= R_{err}(h_0) + 2 \sup_{h \in \mathcal{H}} |R_{err}(h) - \hat{R}_{err}(h)| \\
(Thm 2.2 \to) &\le R_{err}(h_0) + 2 \underbrace{\left(2\sqrt{\frac{2d}{n} \log \frac{en}{d}} + \sqrt{\frac{\log 2/\delta}{2n}} \right)}_{O_p\left(\sqrt{\frac{d}{n} \log \frac{n}{d}}\right)} \quad \text{w.p. } 1 - \delta
\end{aligned}
$$

VC 次元による汎化誤差の一様上界 IV

$$R_{err}(h_S) \leq R_{err}(h_0) + O_p\left(\sqrt{\frac{d}{n}\log\frac{n}{d}}\right)$$

- VC 次元 d fix で data 数 n を増やす $\rightarrow$ 汎化誤差が減る
- data 数 n fix で VC 次元 d を増やす $\rightarrow$ 汎化誤差が増える

VC 次元による汎化誤差の一様上界 V

Example 1 (有限仮説集合)

$|\mathcal{H}| < \infty$ のとき, $VCdim(\mathcal{H})(= d) \le \log_2 |\mathcal{H}|$

(proof) d 個の入力に割り当てられる 2 値ラベルのパターン総数は 2^d.

もし $|\mathcal{H}| < 2^d$ とすると, $\exists y_1, ..., y_d$ s.t. $\forall h \in \mathcal{H},\ \ h(\boldsymbol{x}_i) \neq y_i$ とできる. すなわち, $\mathcal{H}$ の元でラベルパターンを網羅できない. よって,

$$VCdim(\mathcal{H}) = \underbrace{\log_2 2^d}_{=d} \le \log_2 |\mathcal{H}|$$

このとき, 汎化誤差のバウンドは

$$\begin{aligned} R_{err}(h_S) &\le R_{err}(h_0) + O_p\left(\sqrt{\frac{d_{\mathcal{H}}}{n}\log\frac{n}{d_{\mathcal{H}}}}\right) \\ &\le R_{err}(h_0) + O_p\left(\sqrt{\frac{\log_2|\mathcal{H}|}{n}\log\frac{n}{\log_2|\mathcal{H}|}}\right) \end{aligned}$$

Example 2 ($\mathbb{R}^d$ 上の線形判別)

- $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{d+1} \subset \mathcal{X} \times \{+1, -1\}$
- $\mathcal{H} = \{h(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b); \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$: 線形判別器

$A = \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{d+1} \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$ が可逆のとき, $\begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} = A^{-1}\boldsymbol{y}$ とパラメータを取ると, $y_i = h(\boldsymbol{x}_i)$ が成立:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{d+1} \end{pmatrix} = A \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} = \begin{pmatrix} \boldsymbol{w}^\top \boldsymbol{x}_1 + b \\ \vdots \\ \boldsymbol{w}^\top \boldsymbol{x}_{d+1} + b \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x}_1 + b) \\ \vdots \\ \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x}_{d+1} + b) \end{pmatrix} = \begin{pmatrix} h(\boldsymbol{x}_1) \\ \vdots \\ h(\boldsymbol{x}_{d+1}) \end{pmatrix}$$

これより, $VCdim(\mathcal{H}) \geq d + 1$ が言える.

仮説集合の複雑さの upper bound を求めたい

Theorem 2 (Radon's Theorem)

$\forall S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_{d+2}\} \subset \mathbb{R}^d$,
$\exists S_1, S_2$: *a partition of* S *(i.e.* $S_1 \cup S_2 = S$, $S_1 \cap S_2 = \emptyset$*) s.t.*

$$\mathrm{conv}(S_1) \cap \mathrm{conv}(S_2) \neq \emptyset$$

ここで $\mathrm{conv}(A)$ は A の凸包:

$$\mathrm{conv}(A) := \left\{ \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i \,\middle|\, n \in \mathbb{N},\ \sum_{i=1}^{n} \alpha_i = 1,\ \alpha \in [0,1],\ \boldsymbol{x}_i \in A \right\}$$

Radon's Theorem (VC 次元の上界) II

2 値判別問題に対して, Radon's thm を使って VC 次元の上界を計算

- $S_1, S_2 : S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_{d+2}\} \subset \mathbb{R}^d$ の Radon partition
- true label :
$$y_i = \begin{cases} +1 & \text{if } \boldsymbol{x}_i \in S_1 \\ -1 & \text{if } \boldsymbol{x}_i \in S_2 \end{cases}$$
- true label に正答する線形判別器 $h \in \mathcal{H}$ が存在すると仮定:
$$h(\boldsymbol{x}_i) = \begin{cases} +1 & \text{if } \boldsymbol{x}_i \in \mathrm{conv}(S_1) \\ -1 & \text{if } \boldsymbol{x}_i \in \mathrm{conv}(S_2) \end{cases}$$
- しかし, h は $\boldsymbol{x} \in \mathrm{conv}(S_1) \cap \mathrm{conv}(S_2)$ に対してはどちらのラベルも付与してしまい矛盾
$\rightarrow d+2$ 個の入力点のラベル付けは線形判別器では網羅できない
$\rightarrow VCdim(\mathcal{H}) \leq d+1$
- 一方, 線形判別器の VC 次元は $VCdim(\mathcal{H}) \geq d+1$ を満たすから, 両者を合わせると $VCdim(\mathcal{H}) = d+1$ を得る

Proof of Radon's Theorem

$\alpha_1, \ldots, \alpha_{d+2} \in \mathbb{R}$ に関する $d+1$ 個の線形方程式系を考える:

$$\begin{cases} \sum_{i=1}^{d+2} \alpha_i \boldsymbol{x}_i = 0 \\ \sum_{i=1}^{d+2} \alpha_i = 0 \end{cases} \iff \begin{cases} \alpha_1 x_{11} + \cdots \alpha_{d+2} x_{d+2,1} = 0 \\ \alpha_1 x_{12} + \cdots \alpha_{d+2} x_{d+2,2} = 0 \\ \qquad \vdots \\ \alpha_1 x_{1d} + \cdots \alpha_{d+2} x_{d+2,d} = 0 \\ \alpha_1 + \cdots \alpha_{d+2} = 0 \end{cases}$$

$d+2$ 個の未知数に対して方程式の数が $d+1$ であるから, この系は非自明な解 $\beta_1, \ldots, \beta_{d+2}$ を持つ (i.e. $\exists i$ s.t. $\beta_i \neq 0$)

集合 I_1, I_2 をそれぞれ

$$I_1 = \{i \in [d+2] \mid \beta_i > 0\}$$
$$I_2 = \{i \in [d+2] \mid \beta_i \leq 0\}$$

と定めると, $\sum_{i=1}^{d+2} \beta_i = 0$ かつ β の非自明性から,

$$I_1 \neq \emptyset, \quad I_2 \neq \emptyset$$

であり, S_1, S_2 を

$$S_1 = \{\boldsymbol{x}_i \in S \mid i \in I_1\}$$
$$S_2 = \{\boldsymbol{x}_i \in S \mid i \in I_2\}$$

ととると, これらは S の Radon partition をなす (i.e. $S_1 \cup S_2 = S$, $S_1 \cap S_2 = \emptyset$)

Radon's Theorem (VC 次元の上界) V

再び $\sum_{i=1}^{d+2} \beta_i = 0$ より,

$$\sum_{i=1}^{d+2} \beta_i = \sum_{i \in I_1} \beta_i + \sum_{i \in I_2} \beta_i = 0 \Longleftrightarrow \sum_{i \in I_1} \beta_i = -\sum_{i \in I_2} \beta_i$$

が成立. いま, 左辺を β をおくと,

$$\sum_{i=1}^{d+2} \beta_i \boldsymbol{x}_i = \sum_{i \in I_1} \beta_i \boldsymbol{x}_i + \sum_{i \in I_2} \beta_i \boldsymbol{x}_i = 0 \Longleftrightarrow \sum_{i \in I_1} \frac{\beta_i}{\beta} \boldsymbol{x}_i = \sum_{i \in I_2} \frac{-\beta_i}{\beta} \boldsymbol{x}_i$$

かつ, $\frac{\beta_i}{\beta} \geq 0 \, (i \in I_1)$, $\frac{-\beta_i}{\beta} \geq 0 \, (i \in I_2)$ で,

$$\sum_{i \in I_2} \frac{-\beta_i}{\beta} = \sum_{i \in I_1} \frac{\beta_i}{\beta} = \frac{\beta}{\beta} = 1$$

が成立 (β で割って規格化することで凸結合になってる).

凸包の定義から,

$$\mathrm{conv}(S_1) \ni \sum_{i \in I_1} \frac{\beta_i}{\beta} \boldsymbol{x}_i = \sum_{i \in I_2} \frac{-\beta_i}{\beta} \boldsymbol{x}_i \in \mathrm{conv}(S_2)$$

であり, 特に $\frac{\beta_i}{\beta} \boldsymbol{x}_i \in \mathrm{conv}(S_1) \cap \mathrm{conv}(S_2)$ が言えた. □

Example 3 ($VCdim(\mathcal{H}) = \infty$ の例)

$$\mathcal{H} = \{h(\boldsymbol{x}) = \mathrm{sign}(\sin(2\pi\theta\boldsymbol{x})) | \theta \in \mathbb{R}\}$$

仮説集合の複雑度

2.2 ラデマッハ複雑度

ラデマッハ複雑度 I

ある確率分布に基づいて仮説集合の複雑さを測る.

仮説集合: $\mathcal{G} = \{f : \mathcal{X} \longrightarrow \mathbb{R}\}$

Definition 3 (empirical Rademacher complexity)

- $S = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$: *input set*
- $\sigma_i = \pm 1$ *w.p.* $\frac{1}{2}$: *independent r.v.*

このとき, 仮説集合 $\mathcal{G}$ の *empirical Rademacher complexity* は以下で定義される

$$\hat{\mathcal{R}}_S(\mathcal{G}) := \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(\boldsymbol{x}_i)\right]$$

S 上のランダムなラベル付け (x_i, σ_i), $1 \leq i \leq n$ に対して $\mathcal{G}$ の data への平均的適合度を評価している

ラデマッハ複雑度 II

Definition 4 (Rademacher complexity)

$S = \{\boldsymbol{x}_i\}_{i=1}^n \sim D$ のとき, $\hat{\mathcal{R}}_S(\mathcal{G})$ の D に関する期待値

$$\mathcal{R}_n(\mathcal{G}) := \mathbb{E}_{S \sim D}\left[\hat{\mathcal{R}}_S(\mathcal{G})\right]$$

を $\mathcal{G}$ の *Rademacher complexity* という

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質

Theorem 3 (経験ラデマッハ複雑度の性質)

$\mathcal{G}, \mathcal{G}_1, ..., \mathcal{G}_k$: 仮説集合列

1. $\mathcal{G}_i \subset \mathcal{G}_j \Longrightarrow \hat{\mathcal{R}}_S(\mathcal{G}_i) \leq \hat{\mathcal{R}}_S(\mathcal{G}_j)$
2. $\forall c \in \mathbb{R}$, $\hat{\mathcal{R}}_S(c\mathcal{G}) \leq |c|\hat{\mathcal{R}}_S(\mathcal{G})$
3. $\hat{\mathcal{R}}_S(\mathcal{G}) = \hat{\mathcal{R}}_S(\mathrm{conv}(\mathcal{G}))$
4. (Talagrand's lemma)
   $\phi : \mathbb{R} \to \mathbb{R}$: L-Lipschitz $\Longrightarrow \hat{\mathcal{R}}_S(\phi \circ \mathcal{G}) \leq L\hat{\mathcal{R}}_S(\mathcal{G})$
5. (subadditivity)
   $\hat{\mathcal{R}}_S(\sum_{i=1}^k \mathcal{G}_i) \leq \sum_{i=1}^k \hat{\mathcal{R}}_S(\mathcal{G}_i)$
6. $\mathcal{G} \subset \{(\boldsymbol{x}, y) \mapsto f(\boldsymbol{x}, y)\}$ に対して $\mathcal{G}_y = \{x \mapsto f(\boldsymbol{x}, y) \mid f \in \mathcal{G}\}$ とおく
   $\Longrightarrow \hat{\mathcal{R}}_S(\mathcal{G}) \leq \sum_{y \in \mathcal{Y}} \hat{\mathcal{R}}_S(\mathcal{G}_y)$
7. $\mathcal{G} = \{\boldsymbol{x} \mapsto \max\{f_1(\boldsymbol{x}), ..., f_k(\boldsymbol{x})\} \mid f_1 \in \mathcal{G}_1, ..., f_k \in \mathcal{G}_k\}$ とおく
   $\Longrightarrow \hat{\mathcal{R}}_S(\mathcal{G}) \leq \sum_{\ell=1}^k \hat{\mathcal{R}}_S(\mathcal{G}_\ell)$

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 II

Proof

1. $\sup$ の定義から明らか ($\mathcal{G}_i$ より $\mathcal{G}_j$ の方が $\sup$ の範囲が広いから) □
2. $\forall c \in \mathbb{R}$ に対して, σ_i と $\mathrm{sign}(c)\sigma_i$ は同一分布に従う (いずれも等確率で ± 1 を返す). このとき, 以下が成立:

$$
\begin{aligned}
\hat{\mathcal{R}}(c\mathcal{G}) &= \mathbb{E}_{\sigma_1,\dots,\sigma_n}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i c g(\boldsymbol{x}_i)\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i |c| \mathrm{sign}(c) g(\boldsymbol{x}_i)\right] \\
&= |c|\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathrm{sign}(c) g(\boldsymbol{x}_i)\right] \\
&= |c|\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i)\right] = |c|\hat{\mathcal{R}}_S(\mathcal{G}) \quad \square
\end{aligned}
$$

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 III

Proof 続　3. $\mathrm{conv}(\mathcal{G}) = \left\{ \sum_{\ell=1}^{k} \alpha_\ell g_\ell \mid k \in \mathbb{N},\ \alpha_\ell \in [0,1],\ \sum_{\ell=1}^{k} \alpha_\ell = 1,\ g_\ell \in \mathcal{G} \right\}$

より,

$$\begin{aligned}
\sup_{g \in \mathrm{conv}(\mathcal{G})} \sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i) &= \sup_{g_1,\ldots,g_k} \sup_{\alpha_1,\ldots,\alpha_k} \sum_{i=1}^{n} \sigma_i \sum_{\ell=1}^{k} \alpha_\ell g_\ell(\boldsymbol{x}_i) \\
(\text{有限和の順序交換} \to) &= \sup_{g_1,\ldots,g_k} \sup_{\alpha_1,\ldots,\alpha_k} \sum_{\ell=1}^{k} \alpha_\ell \sum_{i=1}^{n} \sigma_i g_\ell(\boldsymbol{x}_i) \\
(\diamond \to) &= \sup_{g_1,\ldots,g_k} \max_{1 \le \ell \le k} \sum_{i=1}^{n} \sigma_i g_\ell(\boldsymbol{x}_i) \\
&= \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i)
\end{aligned}$$

よって両辺で σ について期待値をとれば主張が従う. $\diamond$ は次で示す

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 IV

Proof 続　3. $\diamond$ は, 以下の事実から従う:

$$\sup_{\substack{\alpha_1,\dots,\alpha_k \geq 0 \\ \sum \alpha_\ell = 1}} \sum_{\ell=1}^{k} \alpha_\ell v_\ell = \max_{1 \leq \ell \leq k} v_\ell, \quad \forall \mathbf{v} = (v_1, \dots, v_k)$$

$(\because)$ $(\geq)$ $\hat{\ell} = \arg\max_{\ell} v_\ell$ とおくと, 右辺は $\alpha = (0, \dots, \underbrace{1}_{\hat{\ell}}, \dots, 0)$ なる α のとり方をした場合に相当. 左辺はこのとり方を含めた全ての α で $\sup$ を取っているから明らか.

$(\leq)$

$$\sum_{\ell=1}^{k} \alpha_\ell v_\ell \leq v_{\hat{\ell}} \underbrace{\sum_{\ell=1}^{k} \alpha_\ell}_{=1} = v_{\hat{\ell}}$$

両辺で α について $\sup$ をとれば主張が従う.

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 V

Proof 続

4. $S = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$ に対して

$$u_{n-1}(f) := \sum_{i=1}^{n-1} \sigma_i \phi(f(\boldsymbol{x}_i))$$

とおくと,

$$\begin{aligned}
\hat{\mathcal{R}}(\phi \circ \mathcal{G}) &= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i \phi(f(\boldsymbol{x}_i))\right] \\
&= \frac{1}{n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{G}}\left\{\sum_{i=1}^{n-1}\sigma_i\phi(f(\boldsymbol{x}_i)) + \sigma_n\phi(f(\boldsymbol{x}_n))\right\}\right] \\
&= \frac{1}{n}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{G}}\left\{\sum_{i=1}^{n-1}\sigma_i\phi(f(\boldsymbol{x}_i)) + \sigma_n\phi(f(\boldsymbol{x}_n))\right\}\right]\right]
\end{aligned}$$

と書ける

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 VI

Proof 続

4. sup の定義より, $\forall \varepsilon > 0$, $\exists f^{(+)}, f^{(-)} \in \mathcal{G}$ s.t,

$$\sup_{f \in \mathcal{G}} \{u_{n-1}(f) + \phi(f(\boldsymbol{x}_n))\} \leq u_{n-1}(f^{(\pm)}) \pm \phi(f^{(\pm)}(\boldsymbol{x}_n)) + \varepsilon$$

が成立 (復号同順). いま, $s_n = \mathrm{sign}(f^{(+)}(\boldsymbol{x}_n) - f^{(-)}(\boldsymbol{x}_n))$ とおくと,

$$\begin{aligned}
&\mathbb{E}_{\sigma_n}\left[\sup_{f \in \mathcal{G}} \{u_{n-1}(f) + \sigma_n \phi(f(\boldsymbol{x}_n))\}\right] \\
&\leq \frac{1}{2}\left\{u_{n-1}(f^{(+)}) + \phi(f^{(+)}(\boldsymbol{x}_n)) + u_{n-1}(f^{(-)}) - \phi(f^{(-)}(\boldsymbol{x}_n))\right\} + \varepsilon \\
&\leq \frac{1}{2}\left\{u_{n-1}(f^{(+)}) + u_{n-1}(f^{(-)}) + \underbrace{\phi(f^{(+)}(\boldsymbol{x}_n)) - \phi(f^{(-)}(\boldsymbol{x}_n))}_{\leq L|f^{(+)}(\boldsymbol{x}_n) - f^{(-)}(\boldsymbol{x}_n)|}\right\} \\
&\leq \frac{1}{2}\left\{u_{n-1}(f^{(+)}) + u_{n-1}(f^{(-)}) + L s_n(f^{(+)}(\boldsymbol{x}_n) - f^{(-)}(\boldsymbol{x}_n))\right\}
\end{aligned}$$

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 VII

Proof 続

4.

$$
\begin{aligned}
&\frac{1}{2}\left\{u_{n-1}(f^{(+)}) + u_{n-1}(f^{(-)}) + Ls_n(f^{(+)}(\boldsymbol{x}_n) - f^{(-)}(\boldsymbol{x}_n))\right\} + \varepsilon \\
=&\frac{1}{2}\left\{u_{n-1}(f^{(+)}) + Ls_n f^{(+)}(\boldsymbol{x}_n)\right\} \\
&\qquad\qquad + \frac{1}{2}\left\{u_{n-1}(f^{(-)}) - Ls_n f^{(-)}(\boldsymbol{x}_n)\right\} + \varepsilon \\
\leq&\frac{1}{2}\mathbb{E}_{\sigma_n}\left[\sup_f \{u_{n-1}(f) + \sigma_n Ls_n f(\boldsymbol{x}_n)\}\right] \\
&\qquad\qquad + \frac{1}{2}\mathbb{E}_{\sigma_n}\left[\sup_f \{u_{n-1}(f) + \sigma_n Ls_n f(\boldsymbol{x}_n)\}\right] + \varepsilon \\
=&\mathbb{E}_{\sigma_n}\left[\sup_f \{u_{n-1}(f) + \sigma_n Ls_n f(\boldsymbol{x}_n)\}\right] + \varepsilon
\end{aligned}
$$

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 VIII

Proof 続　4.　上記の不等式が $\forall\varepsilon > 0$ で成立つから, $\varepsilon \searrow 0$ とすると,

$$\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{G}}\left\{u_{n-1}(f) + \sigma_n\phi(f(\boldsymbol{x}_n))\right\}\right] \le \mathbb{E}_{\sigma_n}\left[\sup_{f}\left\{u_{n-1}(f) + \sigma_n L f(\boldsymbol{x}_n)\right\}\right]$$

が成立 (σ_n と $\sigma_n s_n$ が同一の分布を定めることを使う).
次に, $n-1$ 番目に注目して

$$u_{n-2}(f) = \sum_{i=1}^{n-2}\sigma_i\phi(f(\boldsymbol{x}_i)) + \sigma_n L f(\boldsymbol{x}_n)$$

とおき, 同様の議論で

$$\begin{aligned}&\mathbb{E}_{\sigma_{n-1},\sigma_n}\left[\sup_{f\in\mathcal{G}}\left\{u_{n-2}(f) + \sigma_{n-1}\phi(f(\boldsymbol{x}_{n-1}))\right\}\right]\\ \le&\mathbb{E}_{\sigma_{n-1},\sigma_n}\left[\sup_{f}\left\{u_{n-2}(f) + \sigma_{n-1} L f(\boldsymbol{x}_{n-1})\right\}\right]\end{aligned}$$

を得る.

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 IX

Proof 続 4. 以上の手続きを σ_1 まで繰り返すと, 結局

$$\begin{aligned}\hat{\mathcal{R}}_S(\phi \circ \mathcal{G}) &= \frac{1}{n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i\phi(f(\boldsymbol{x}_i))\right] \\ &\leq \frac{L}{n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right] \\ &= L\hat{\mathcal{R}}_S(\mathcal{G})\end{aligned}$$

を得る □

5. $\sup$ の性質

$$\sup(A+B) \leq \sup(A) + \sup(B)$$

から従う □

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 X

Proof 続　6. $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ に対して

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i, y_i)\right]$$

$$= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i \sum_{y\in\mathcal{Y}} f(\boldsymbol{x}_i, y)\mathbf{1}[y = y_i]\right]$$

(sup の性質 $\rightarrow$) $$\leq \frac{1}{n}\sum_{y\in\mathcal{Y}}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i, y)\mathbf{1}[y = y_i]\right]$$

$$= \frac{1}{n}\sum_{y\in\mathcal{Y}}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i, y)\left(\frac{1}{2} + \frac{2\times\mathbf{1}[y = y_i] - 1}{2}\right)\right]$$

$$\leq \frac{1}{2n}\sum_{y\in\mathcal{Y}}\left(\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i, y)\right] + \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i (2\times\mathbf{1}[y = y_i] - 1) f(\boldsymbol{x}_i, y)\right]\right)$$

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 XI

Proof 続　6.

$$
\begin{aligned}
&\frac{1}{2n}\sum_{y\in\mathcal{Y}}\left(\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i, y)\right] + \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^n \sigma_i (2\times\mathbf{1}[y=y_i]-1) f(\boldsymbol{x}_i, y)\right]\right)\\
=\ &\frac{1}{2}\sum_{y\in\mathcal{Y}}\hat{\mathcal{R}}_S(\mathcal{G}_y) + \frac{1}{2}\sum_{y\in\mathcal{Y}}\hat{\mathcal{R}}_S(\mathcal{G}_y)\\
=\ &\sum_{y\in\mathcal{Y}}\hat{\mathcal{R}}_S(\mathcal{G}_y)
\end{aligned}
$$

ここで, 最初の等号では σ_i と $\sigma_i(2\times\mathbf{1}[y=y_i]-1)$ の分布が等しいことを使った. □

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 XII

Proof 続　7. $k = 2$ の場合を示す: $\mathcal{G} = \{\max\{f_1, f_2\} \mid f_1 \in \mathcal{G}_1, f_2 \in \mathcal{G}_2\}$.

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^n \sigma_i \max\{f_1(\boldsymbol{x}_i), f_2(\boldsymbol{x}_i)\}\right]$$

$$= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^n \sigma_i \frac{f_1(\boldsymbol{x}_i)+f_2(\boldsymbol{x}_i)}{2} + \frac{|f_1(\boldsymbol{x}_i)-f_2(\boldsymbol{x}_i)|}{2}\right]$$

$$\left(\uparrow \max\{z_1, z_2\} = \frac{z_1+z_2}{2} + \frac{|z_1-z_2|}{2}\right)$$

$$\leq \frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_1}\sum_{i=1}^n \sigma_i f_1(\boldsymbol{x}_i)\right] + \frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_2}\sum_{i=1}^n \sigma_i f_2(\boldsymbol{x}_i)\right]$$

$$+ \frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^n \sigma_i |f_1(\boldsymbol{x}_i)-f_2(\boldsymbol{x}_i)|\right]$$

$$= \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_1) + \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_2) + \frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^n \sigma_i |f_1(\boldsymbol{x}_i)-f_2(\boldsymbol{x}_i)|\right]$$

経験ラデマッハ複雑度 $\hat{\mathcal{R}}_S(\mathcal{G})$ の性質 XIII

Proof 続　7.

$$\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_1)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_2)+\frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^{n}\sigma_i|f_1(\boldsymbol{x}_i)-f_2(\boldsymbol{x}_i)|\right]$$

$|\cdot|$ は 1-Lipschitz 連続なので, 本定理の 4 より,

$$\begin{aligned}\frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^{n}\sigma_i|f_1(\boldsymbol{x}_i)-f_2(\boldsymbol{x}_i)|\right]&\le\frac{1}{2n}\mathbb{E}_\sigma\left[\sup_{f_1,f_2}\sum_{i=1}^{n}\sigma_i(f_1(\boldsymbol{x}_i)-f_2(\boldsymbol{x}_i))\right]\\&\le\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_1)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_2)\end{aligned}$$

結局,

$$\hat{\mathcal{R}}_S(\mathcal{G})\le\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_1)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_2)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_1)+\frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{G}_2)=\hat{\mathcal{R}}_S(\mathcal{G}_1)+\hat{\mathcal{R}}_S(\mathcal{G}_2)$$

$k\ge3$ の場合は以上を帰納的に繰り返す □

ラデマッハ複雑度と VC 次元の関係

- 2 値判別
- $S = \{\boldsymbol{x}_i\}_{i=1}^n$: input data
- $\mathcal{H} = \{h : \mathcal{X} \to \{+1, -1\}\}$: 仮説集合 with $\mathrm{VCdim}(\mathcal{H}) = d$
- $A = \{(h(\boldsymbol{x}_1), ..., h(\boldsymbol{x}_n)) \in \{+1, -1\}^n \mid h \in \mathcal{H}\}$

このとき, $n \geq d$ ならば,

$$|A| = \Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \leq \underbrace{\max_{\boldsymbol{x}_1, ..., \boldsymbol{x}_n} \Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)}_{growth\ \ function} \underbrace{\leq}_{Sauer} \left(\frac{en}{d}\right)^d$$

が成立. S における $\mathcal{H}$ の経験ラデマッハ複雑度は

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{z}\in A} \sum_{i=1}^n \sigma_i z_i\right] \leq \sqrt{\frac{2d}{n}\log\frac{en}{d}}$$

最後の不等式は Massart's lemma を使った.

ラデマッハ複雑度と VC 次元の関係 II

Lemma 4 (Massart's lemma)

- $A \subset \mathbb{R}^m$: *finite set*
- $r = \max_{\boldsymbol{x} \in A} \|x\|_2$
- $\sigma_1, \ldots, \sigma_m \sim_{i.i.d.} \mathrm{Unif}(\{+1, -1\})$

このとき, 以下が成立

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

x_i として $z_i \in \{+1, -1\}$ ($\|\boldsymbol{z}\| = \sqrt{n}$) をとれば,

$$\frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\boldsymbol{z} \in A} \sum_{i=1}^{n} \sigma_i z_i \right] \leq \frac{\sqrt{n}\sqrt{2 \log |A|}}{n} \leq \sqrt{\frac{2}{n} \log \left(\frac{en}{d} \right)^d} = \sqrt{\frac{2d}{n} \log \left(\frac{en}{d} \right)}$$

がいえる.

ラデマッハ複雑度と VC 次元の関係 III

Proof of Massart's Lemma $\forall t > 0$ に対して,

$$\exp\left\{\mathbb{E}_\sigma\left[t \sup_{\boldsymbol{x}\in A}\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right]\right\} \underbrace{\leq}_{(\diamond)} \mathbb{E}_\sigma\left[\exp\left\{t \sup_{\boldsymbol{x}\in A}\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\}\right]$$

$$\underbrace{\leq}_{(\diamond^2)} \sum_{\boldsymbol{x}\in A}\mathbb{E}_\sigma\left[\exp\left\{t \sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\}\right]$$

$$\underbrace{=}_{(\diamond^3)} \sum_{\boldsymbol{x}\in A}\prod_{i=1}^m \mathbb{E}_{\sigma_i}\left[\exp\{t\sigma_i \boldsymbol{x}_i\}\right]$$

$\diamond$ exp の凸性 + Jensen's inequality ($cvx(\mathbb{E}) \leq \mathbb{E}[cvx]$)

$\diamond^2$ $\sup_{\boldsymbol{x}\in A} \leq \sum_{\boldsymbol{x}\in A}$

$\diamond^3$ 和を exp の外に出して積になった

ラデマッハ複雑度と VC 次元の関係 IV

Proof of Massart's Lemma さらに, Hoeffding's lem より以下が成立.

$$\mathbb{E}_{\sigma_i}\left[\exp\{t\sigma_i \boldsymbol{x}_i\}\right] \leq \exp\left\{\frac{t^2(2\boldsymbol{x}_i)^2}{8}\right\}$$

よって,

$$\begin{aligned}\sum_{\boldsymbol{x}\in A}\prod_{i=1}^{m}\mathbb{E}_{\sigma_i}\left[\exp\{t\sigma_i \boldsymbol{x}_i\}\right] &\leq \sum_{\boldsymbol{x}\in A}\prod_{i=1}^{m}\exp\left\{\frac{t^2(2\boldsymbol{x}_i)^2}{8}\right\} \\ &\leq |A|\exp\{\frac{t^2}{2}\underbrace{\sum_{i=1}^{m}\boldsymbol{x}_i^2\}}_{=r^2} = |A|\exp\left\{\frac{t^2r^2}{2}\right\}\end{aligned}$$

upper bound の対数をとって t で割る:

$$\frac{1}{t}\left(\log|A| + \frac{t^2r^2}{2}\right) = \frac{\log|A|}{t} + \frac{tr^2}{2}$$

ラデマッハ複雑度と VC 次元の関係 V

Proof of Massart's Lemma 最小化した上界を用いて, 以下を得る

$$\exp\left\{\mathbb{E}_\sigma\left[t\sup_{\boldsymbol{x}\in A}\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right]\right\} \le |A|\exp\left\{\frac{t^2r^2}{2}\right\}$$

$$\iff \mathbb{E}_\sigma\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right] \le \frac{\log|A|}{t} + \frac{tr^2}{2}$$

右辺を t について最小化すると,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\log|A|}{t} + \frac{tr^2}{2}\right) = \frac{r^2}{2} - \frac{\log|A|}{t^2} = 0 \iff t^2 = \frac{2\log|A|}{r^2}$$

よって $t = \frac{\sqrt{2\log|A|}}{r}$ とおくと,

$$\mathbb{E}_\sigma\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right] \le \frac{r\sqrt{2\log|A|}}{2} + \frac{r\sqrt{2\log|A|}}{2} = r\sqrt{2\log|A|}$$

より, 両辺を m で割って主張を得る. □

経験ラデマッハ複雑度の例 I : 有限集合

- $\mathcal{G} = \{g_1, ..., g_k\}$: 有限関数集合
- $A = \{g_\ell(z_1), ..., g_\ell(z_n) \in \mathbb{R}^n \mid 1 \leq \ell \leq k\}$ ($\{z_i\}_{i=1}^n$ は fix)
- $1 \leq \forall \ell \leq k$ に対して以下が成立:

$$\|g_\ell\|_\infty = \sup_z |g_\ell(z)| \leq r \left(\Longleftrightarrow \underbrace{\left(\sum_{i=1}^n (g_\ell(z_i))^2\right)^{1/2}}_{=\|G\|,\ G \in A} \leq r \right)$$

このとき,

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \left[\max_{1 \leq \ell \leq k} \frac{1}{n} \sum_{i=1}^n \sigma_i g_\ell(z_i) \right]$$

$$(Massart \rightarrow) \leq \underbrace{\max_{1 \leq \ell \leq k} \left(\sum_{i=1}^n (g_\ell(z_i))^2\right)^{1/2}}_{\leq r} \frac{\sqrt{2 \log |A|}}{n}$$

$$(|\mathcal{G}| = |A| \rightarrow) \leq r \frac{\sqrt{2 \log |\mathcal{G}|}}{n}. \quad \square$$

線形関数集合 $\mathcal{G} = \{\boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} \mid \boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\| \leq \Lambda\}$ の経験ラデマッハ複雑度

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \left[\frac{1}{n} \sup_{\|\boldsymbol{w}\| \leq \Lambda} \sum_{i=1}^n \sigma_i \boldsymbol{w}^\top \boldsymbol{x}_i \right] = \mathbb{E}_\sigma \left[\frac{1}{n} \sup_{\|\boldsymbol{w}\| \leq \Lambda} \boldsymbol{w}^\top \left(\sum_{i=1}^n \sigma_i \boldsymbol{x}_i \right) \right]$$

$$\underbrace{=}_{(\diamond)} \frac{1}{n} \mathbb{E}_\sigma \left[\Lambda \left\| \sum_{i=1}^n \sigma_i \boldsymbol{x}_i \right\| \right]$$

$(\diamond)$ Claim $\sup_{\|\boldsymbol{x}\| \leq r} |\boldsymbol{x}^\top \boldsymbol{y}| = r\|\boldsymbol{y}\|$

$\because$ $(\leq)$ Cauchy-Schwartz 不等式より, $|\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\| \leq r\|\boldsymbol{y}\|$.

$(\geq)$ $\boldsymbol{x} = \frac{r}{\|\boldsymbol{y}\|}\boldsymbol{y}$ ととると, $\|\boldsymbol{x}\| \leq r$ で,

$$|\boldsymbol{x}^\top \boldsymbol{y}| = \left| \left(\frac{r}{\|\boldsymbol{y}\|} \boldsymbol{y} \right)^\top \boldsymbol{y} \right| = r \frac{\|\boldsymbol{y}\|}{\|\boldsymbol{y}\|} \|\boldsymbol{y}\| = r\|\boldsymbol{y}\|$$

が成立 (2 つめの等号は, Cauchy-Schwarz 不等式の等号成立条件 ($\exists \lambda$ s.t. $\boldsymbol{x} = \lambda \boldsymbol{y}$) による). 特に, $|\boldsymbol{x}^\top \boldsymbol{y}| \geq r\|\boldsymbol{y}\|$.

経験ラデマッハ複雑度の例 II : 線形関数集合 II

$$\frac{1}{n}\mathbb{E}_\sigma\left[\Lambda\left\|\sum_{i=1}^n \sigma_i \boldsymbol{x}_i\right\|\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\Lambda\left(\left\|\sum_{i=1}^n \sigma_i \boldsymbol{x}_i\right\|^2\right)^{1/2}\right]$$

$$\underbrace{\le}_{(\diamond)} \frac{\Lambda}{n}\left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i \boldsymbol{x}_i\right\|^2\right]\right)^{1/2} \underbrace{=}_{\diamond^2} \frac{\Lambda}{n}\left(\sum_{i=1}^n \|\boldsymbol{x}_i\|^2\right)^{1/2}$$

($\diamond$) concave function $\sqrt{\cdot}$ に対する Jensen 不等式 ($\mathbb{E}[\sqrt{\cdot}] \le \sqrt{\mathbb{E}[\cdot]}$) による.
($\diamond^2$) $n = 2$ のとき ($n \ge 3$ のときも同様にクロスタームが消える),

$$\begin{aligned}
\mathbb{E}[\|\sigma_1\boldsymbol{x}_1 + \sigma_2\boldsymbol{x}_2\|] &= \mathbb{E}[\|\sigma_1\boldsymbol{x}_1\|^2 + \|\sigma_2\boldsymbol{x}_2\|^2 + \sigma_1\sigma_2\boldsymbol{x}_1^\top\boldsymbol{x}_2] \\
&= \mathbb{E}[\underbrace{\sigma_1^2}_{=1}\|\boldsymbol{x}_1\|^2 + \underbrace{\sigma_2^2}_{=1}\|\boldsymbol{x}_2\|^2 + \sigma_1\sigma_2\boldsymbol{x}_1^\top\boldsymbol{x}_2] \\
&= \|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2 + \mathbb{E}[\sigma_1\sigma_2\boldsymbol{x}_1^\top\boldsymbol{x}_2] \\
(\sigma\text{ の独立性 }\to) &= \|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2 + \underbrace{\mathbb{E}[\sigma_1]}_{=0}\underbrace{\mathbb{E}[\sigma_2]}_{=0}\boldsymbol{x}_1^\top\boldsymbol{x}_2 = \|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2
\end{aligned}$$

経験ラデマッハ複雑度の例 II : 線形関数集合 III

結局,

$$\hat{\mathcal{R}}_S \le \frac{\Lambda}{n}\left(\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2\right)^{1/2}.$$

入力に norm 制約 $\|\boldsymbol{x}_i\| \le r$, $1 \le i \le n$ があるとき, 特に

$$\hat{\mathcal{R}}_S \le \frac{r\Lambda}{\sqrt{n}}$$

が成立.

経験ラデマッハ複雑度の例 III : 線形判別器の集合

- $\mathcal{G} = \{\boldsymbol{x} \mapsto \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b) \mid \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$ の VC 次元は $d+1$ (例 2.2 と Radon の定理より).
- Massart lemma による 2 値判別問題のラデマッハ複雑度と VC 次元の関係 (2.1) より,

$$\hat{\mathcal{R}}_S(\mathcal{G}) \leq \sqrt{\frac{2(d+1)}{n} \log \frac{en}{d+1}}$$

が成立.

経験ラデマッハ複雑度の例 IV : 決定株 I

深さ 1 の決定木. data 点ベクトルの各成分をしきい値 z で分割.

- $\mathcal{X} \subset \mathbb{R}^d$: input space
- $s \in \{+1, -1\}$, $k \in [d]$, $z \in \mathbb{R}$: parameters of decision stumps
- 判別器 (decision stumps) : $h(\boldsymbol{x} \mid s, k, z) := s \times \mathrm{sign}(x_k - z)$
- 仮説集合: $\mathcal{G} = \{h(\boldsymbol{x} \mid s, k, z) \mid s = \pm 1, 1 \le k \le d, z \in \mathbb{R}\}$

経験ラデマッハ複雑度を定義より書き下すと,

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{s,k,z}\sum_{i=1}^{n}\sigma_i h(\boldsymbol{x} \mid s, k, z)\right]$$

<u>observation</u>

- 決定株では, <u>軸毎に $2(n+1)$ 通りのラベルの割り当て方が存在</u> ?
- 全体としては高々 $2(n+1)d$ 通りのラベルの割り当て方を考えれば良い

経験ラデマッハ複雑度の例 IV : 決定株 II

$A \subset \{+1, -1\}^n$: stumps で S に割り当てられる binary vectors

$\Longrightarrow |A| \leq 2(n+1)d$

このとき,

$$\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{s,k,z}\sum_{i=1}^n \sigma_i h(x_i \mid s,k,z)\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{(h_1,\ldots,h_n)\in A}\sum_{i=1}^n \sigma_i h_i\right]$$

$$(\text{Massart } \rightarrow) \leq \sqrt{\frac{2}{n}\log(2(n+1)d)}$$

仮説集合の複雑度

2.3 一様大数の法則

<u>Goal :</u> Thm 2.2 の証明

Theorem 5 (一様大数の法則)

- $\mathcal{G} \subset \{f : \mathcal{Z} \to [a, b]\}$
- $Z_1, \ldots, Z_n, Z \sim_{i.i.d.} D$

このとき, $\forall \delta \in (0, 1)$,

$$\Pr_{D^n}\left[\sup_{g \in \mathcal{G}} \left\{ \mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right\} \leq 2\mathcal{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}\right] \geq 1 - \delta$$

が成立 (同様の *bound* が $\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}[g(Z)]$ に対しても成立). 特に, 以下が成立.

$$\Pr_{D^n}\left[\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right| \leq 2\mathcal{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log \frac{2}{\delta}}{2n}}\right] \geq 1 - \delta$$

一様大数の法則の証明 I

まず必要な補題 (Azuma’s inequality, McDiarmid’s inequality) を用意

Lemma 2 (Azuma’s inequality)

- X_i, Z_i, V_i : r.v. $(1 \leq i \leq n)$
- $V_i = V(X_1, ..., X_i)$ s.t. $\mathbb{E}[V_i \mid X_1, ..., X_{i-1}] = 0$
- $Z_i = Z(X_1, ..., X_{i-1})$ s.t. $\exists c_1, ..., c_n$, $Z_i \leq V_i \leq Z_i + c_i$

このとき, $\forall \varepsilon > 0$,

$$\Pr\left(\sum_{i=1}^{n} V_i \geq \varepsilon\right) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right\}$$
$$\Pr\left(\sum_{i=1}^{n} V_i \leq -\varepsilon\right) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right\}$$

が成立.

一様大数の法則の証明 II

Proof　$S_k = \sum_{i=1}^{k} V_i$ とおく. 任意の $t > 0$ に対して,

$$
\begin{aligned}
\Pr(S_n \geq \varepsilon) &= \Pr\left(e^{tS_n} \geq e^{t\varepsilon}\right) \\
\text{(Markov inequality} \rightarrow) &\leq \frac{1}{e^{t\varepsilon}} \mathbb{E}\left[e^{tS_n}\right] = \frac{1}{e^{t\varepsilon}} \mathbb{E}\left[e^{tS_n + tV_n}\right] = \frac{1}{e^{t\varepsilon}} \mathbb{E}\left[e^{tS_n} e^{tV_n}\right] \\
&= \frac{1}{e^{t\varepsilon}} \mathbb{E}_{X_1,\ldots,X_{n-1}}[e^{tS_{n-1}} \underbrace{\mathbb{E}_{X_n}\left[e^{tV_n} \mid X_1, \ldots, X_{n-1}\right]}_{\leq e^{t^2 c_n^2/8}\ (Hoeffding)}] \\
&\leq \frac{1}{e^{t\varepsilon}} \mathbb{E}_{X_1,\ldots,X_{n-1}}[e^{tS_{n-1}}] e^{t^2 c_n^2/8} \\
&= \frac{1}{e^{t\varepsilon}} \mathbb{E}_{X_1,\ldots,X_{n-1}}[e^{tS_{n-2} + tV_{n-1}}] e^{t^2 c_n^2/8} \\
&\leq \frac{1}{e^{t\varepsilon}} \mathbb{E}_{X_1,\ldots,X_{n-2}}[e^{tS_{n-2}}] e^{t^2 \sum_{i=n-1}^{n} c_i^2/8} \\
&\quad \cdots \\
&\leq \frac{1}{e^{t\varepsilon}} e^{t^2 \sum_{i=1}^{n} c_i^2/8} = \exp\left\{\frac{1}{8}\sum_{i=1}^{n} c_i^2 t^2 - \varepsilon t\right\}
\end{aligned}
$$

一様大数の法則の証明 III

Proof　最右辺の $\exp$ の中身を t について最小化すると,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{8}\sum_{i=1}^{n} c_i^2 t^2 - \varepsilon t\right) = \frac{1}{4}\sum_{i=1}^{n} c_i^2 t - \varepsilon = 0$$

$$\iff t = \frac{4\varepsilon}{\sum_{i=1}^{n} c_i^2}$$

これを $\exp$ の中身に代入すると,

$$\Pr(S_n \geq \varepsilon) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right\}$$

もう一方も同様. □

一様大数の法則の証明 IV

Lemma 3 (McDiarmid's inequality)

- $X_1, ..., X_n$: $\mathcal{X}$-*valued independent r.v.*
- $f : \mathcal{X}^n \to \mathbb{R}$ に対して, $\exists c_1, ..., c_n$ s.t. $\forall \boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}'_i \in \mathcal{X} \ \ (1 \leq i \leq n)$,

$$|f(\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_n) - f(\boldsymbol{x}_1, ..., \boldsymbol{x}'_i, ..., \boldsymbol{x}_n)| \leq c_i$$

このとき, 以下が成立:

$$\Pr\left(f(X_1, ..., X_n) - \mathbb{E}[f(X_1, ..., X_n)] \geq \varepsilon\right) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right\}$$

$$\Pr\left(f(X_1, ..., X_n) - \mathbb{E}[f(X_1, ..., X_n)] \leq -\varepsilon\right) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right\}$$

一様大数の法則の証明 V

Proof　$f(S) = f(X_1, ..., X_n$ とおき, $V_1, ..., V_n$ を

$$V_k = \mathbb{E}[f(S) \mid X_1, ..., X_k] - \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}]$$

とする (ただし $V_1 = \mathbb{E}[f(S) \mid X_1] - \mathbb{E}[f(S)]$ とする).

Claim 1
V_k は *Azuma's inequality* の仮定を満たす.

$(\because)$

- 定義より, V_k は $X_1, ..., X_k$ の関数
- 条件付き期待値の性質から,

$$\begin{aligned}&\mathbb{E}[V_k \mid X_1, ..., X_{k-1}] \\ =&\mathbb{E}[\mathbb{E}[f(S) \mid X_1, ..., X_k] - \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}] \mid X_1, ..., X_{k-1}] \\ =&0\end{aligned}$$

一様大数の法則の証明 VI

- f に対する仮定より,

$$
\begin{aligned}
&\sup_{\boldsymbol{x}} \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}, \boldsymbol{x}] - \inf_{\boldsymbol{x}'} \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}, \boldsymbol{x}'] \\
=& \sup_{\boldsymbol{x}, \boldsymbol{x}'} \{\mathbb{E}[f(S) \mid X_1, ..., X_{k-1}, \boldsymbol{x}] - \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}, \boldsymbol{x}']\} \\
\leq& c_i
\end{aligned}
$$

このとき,

$$
\begin{aligned}
Z_k &= \inf_{\boldsymbol{x}} \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}, \boldsymbol{x}] - \mathbb{E}[f(S) \mid X_1, ..., X_k] \\
&\leq \mathbb{E}[f(S) \mid X_1, ..., X_{k-1}, X_k] - \mathbb{E}[f(S) \mid X_1, ..., X_k] \\
&= V_k \leq Z_k + \underbrace{c_k}_{\geq \ \sup V_k}
\end{aligned}
$$

が成立つので, V_k は Azuma's inequality の仮定を満たす. 以上より, $\sum_{i=1}^{n} V_i = f(S) - \mathbb{E}[f(S)]$ に対して Azuma's inequality を適用すれば OK.

一様大数の法則の証明 VII

Proof of Theorem 2.7

$$A(z_1, ..., z_n) = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}[g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i)] \right\}$$

とおく. このとき,

$$\begin{aligned}
&A(z_1, ..., z_n) - A(z_1, ..., z')\\
=&\sup_{g \in \mathcal{G}} \left\{ \mathbb{E}[g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i)] \right\} - \sup_{f \in \mathcal{G}} \left\{ \mathbb{E}[f(Z) - \frac{1}{n} \sum_{i=1}^{n-1} f(z_i) + f(z_{n'})] \right\}\\
=&\sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{G}} \left\{ \mathbb{E}[g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i)] - \mathbb{E}[f(Z) + \frac{1}{n} \sum_{i=1}^{n-1} f(z_i) + f(z_{n'})] \right\}\\
\leq&\sup_{g \in \mathcal{G}} \left\{ \mathbb{E}[g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(z_i)] - \mathbb{E}[g(Z) + \frac{1}{n} \sum_{i=1}^{n-1} g(z_i) + g(z_{n'})] \right\}
\end{aligned}$$

一様大数の法則の証明 VIII

$$
\begin{aligned}
& \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}[g(Z) - \frac{1}{n}\sum_{i=1}^{n} g(z_i)] - \mathbb{E}[g(Z) + \frac{1}{n}\sum_{i=1}^{n-1} g(z_i) + g(z_{n'})] \right\} \\
= & \sup_{g \in \mathcal{G}} \frac{1}{n}(g(z') - g(z_n)) \\
\leq & \frac{b-a}{n} \qquad (\because)\ g(z'), g(z) \in [a, b]
\end{aligned}
$$

が成立. 同様に,

$$
A(z_1, ..., z_{n-1}, z') - A(z_1, ..., z_n) \leq \frac{b-a}{n}
$$

も成立つ. 合わせて,

$$
|A(z_1, ..., z_n) - A(z_1, ..., z')| \leq \frac{b-a}{n}
$$

を得る.

一様大数の法則の証明 IX

McDiarmid's inequality より, $\varepsilon > 0$ に対して

$$\Pr\left(A(Z_1, ..., Z_n) - \mathbb{E}[A(Z_1, ..., Z_n)] \leq \varepsilon\right) \geq 1 - \exp\left\{-\frac{2\varepsilon^2}{n \times \frac{(b-a)^2}{n^2}}\right\}$$

が成立するので, 特に $\delta = \exp\left\{-\frac{2\varepsilon^2}{\frac{1}{n}(b-a)^2}\right\}$ とおくと,

$$\log\delta = -\frac{2n\varepsilon^2}{(b-a)^2} \iff \varepsilon^2 = (b-a)^2 \times \frac{\log\frac{1}{\delta}}{2n}$$

$$\because \quad \varepsilon = (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

となるので,

$$\Pr\left(A(Z_1, ..., Z_n) - \mathbb{E}[A(Z_1, ..., Z_n)] \leq (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right) \geq 1 - \delta$$

一様大数の法則の証明 X

次に, $\mathbb{E}[A(Z_1, ..., Z_n)]$ を評価する.

$Z_1, ..., Z_n, Z'_1, ..., Z'_n \sim_{i.i.d.} P_Z$ とすると, 以下が成立.

$$
\begin{aligned}
&A(Z_1, ..., Z_n) \\
(\text{標本平均の不偏性} \to) &= \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}_{Z'_1, ..., Z'_n} \left[\frac{1}{n} \sum_{i=1}^{n} g(Z'_i) \right] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right\} \\
(\text{和の sup} \leq \text{sup の和} \to) &\leq \mathbb{E}_{Z'_1, ..., Z'_n} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (g(Z'_i) - g(Z_i)) \right]
\end{aligned}
$$

Fact 6

1. $g(Z'_i) - g(Z_i)$ と $g(Z_i) - g(Z'_i)$ は同一分布に従う (対称性)
2. $\sigma_i = \begin{cases} +1 & w.p.\ \frac{1}{2} \\ -1 & w.p.\ \frac{1}{2} \end{cases}$ とすると, $\sigma_i(g(Z'_i) - g(Z_i))$ と $g(Z'_i) - g(Z_i)$ は同一分布に従う

一様大数の法則の証明 XI

Fact より,

$$
\begin{aligned}
&\mathbb{E}_{\sigma,Z}[A(Z_1,...,Z_n)] \\
&\leq \mathbb{E}_\sigma\left[\mathbb{E}_{Z'_1,...,Z'_n}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i(g(Z'_i)-g(Z_i))\right]\right] \\
&\leq \underbrace{\mathbb{E}_{Z'}\left[\mathbb{E}_\sigma \sup \frac{1}{n}\sum \sigma_i g(Z'_i)\right]}_{=\mathcal{R}_n(\mathcal{G})} + \underbrace{\mathbb{E}_Z\left[\mathbb{E}_\sigma \sup \frac{1}{n}\sum \sigma_i g(Z_i)\right]}_{=\mathcal{R}_n(\mathcal{G})} = 2\mathcal{R}_n(\mathcal{G})
\end{aligned}
$$

これを (2.5) 式に代入すると, 確率 $1-\delta$ で以下が成立.

$$
\begin{aligned}
&\sup_{g\in\mathcal{G}}\left\{\mathbb{E}_Z[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right\} - \mathbb{E}[A(Z_1,...,Z_n)] \leq (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \\
\iff &\sup_{g\in\mathcal{G}}\left\{\mathbb{E}_Z[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right\} \leq 2\mathcal{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \qquad \square
\end{aligned}
$$

一様大数の法則の証明 XII

(Proof of Theorem 2.2)

- $\mathcal{H} \subset \{h : \mathcal{X} \to \{+1, -1\}\}$, VCdim($\mathcal{H}$) $= d$
- $\mathcal{G} = \{(\boldsymbol{x}, y) \mapsto \mathbf{1}[h(\boldsymbol{x}) \neq y] \mid h \in \mathcal{H}\}$

とする. このとき,

$$\Pi_{\mathcal{G}}((\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)) = \Pi_{\mathcal{H}}(\boldsymbol{x}_1, ..., x_n)$$

より, VCdim($\mathcal{G}$) = VCdim($\mathcal{H}$) $= d$ が成立. よって (2.1) と一様大数の法則から, $n \geq d$ のとき,

$$\begin{aligned}\sup_{h \in \mathcal{H}} |R_{err}(h) - \hat{R}_{err}(h)| \leq & 2\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\ \leq & 2\sqrt{\frac{2d}{n} \log \frac{en}{d}} + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \quad \square\end{aligned}$$

一様大数の法則の応用: 2 値判別の例

- 有限仮説集合 $\mathcal{H} \subset \{h : \mathcal{X} \to \{+1, -1\}\}$, $h_0 \in \mathcal{H}$
- $\mathcal{G} = \{(\boldsymbol{x}, y) \mapsto \mathbf{1}[h(\boldsymbol{x}) \neq y] \mid h \in \mathcal{H}\}$

このとき, $|\mathcal{G}| = |\mathcal{H}|$ だから, 例 2.4 (有限集合のラデマッハ複雑度) より,

$$\mathcal{R}_n(\mathcal{G}) \leq \sqrt{\frac{2 \log |\mathcal{H}|}{n}}$$

一様大数の法則より,

$$\max_h |R_{err}(h) - \hat{R}_{err}(h)| \leq 2\sqrt{\frac{2 \log |\mathcal{H}|}{n}} + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \quad w.p.\ 1 - \delta$$

が成立. probabilistic order で書くと,

$$R_{err}(h_S) \leq R_{err}(h_0) + \mathcal{O}_p\left(\sqrt{\frac{\log |\mathcal{H}|}{n}}\right) \quad \square$$

仮説集合の複雑度

補足: カバリングナンバー

カバリングナンバー

ラデマッハ複雑度を上から bound する量

Definition 5 (ε-cover)

$\boldsymbol{x}_{1:n} = \{\boldsymbol{x}_i\}_{i=1}^n$ を点集合, $V \subset \mathbb{R}^n$ とする. 任意の $f \in \mathcal{H}$ に対して, $v \in V$ が存在して,

$$\left(\frac{1}{n}\sum_{i=1}^n |v_i - f(\boldsymbol{x}_i)|^p\right)^{1/p} \leq \varepsilon$$

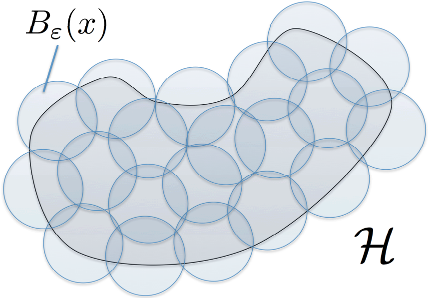
を満たすとき, V を $\mathcal{H}$ の p-次 ε-cover と呼ぶ

Definition 6 (covering number)

$\mathcal{H}$ の p-次 *covering number* は以下で定義される

$$\mathcal{N}_p(\varepsilon, \mathcal{H}, n) = \sup_{\boldsymbol{x}_{1:n}} \min\{|V| \mid V : \mathcal{H} \text{ の } \boldsymbol{x}_{1:n} \text{ 上の } p\text{-次 } \varepsilon\text{-cover}\}$$

カバリングナンバーによる Rademacher Complexity の上界



Theorem 1

$\mathcal{F} \ni f : \mathcal{X} \to [-1, 1]$ とする. このとき,

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \inf_{\varepsilon} \sqrt{\frac{2 \log \mathcal{N}_1\left(\varepsilon, \mathcal{F}, x_{1:n}\right)}{n}} + \varepsilon$$

カバリングナンバーによる Rademacher Complexity の上界

(Proof of Theorem) 半径 ε と minimal cover V を 1 つ固定する.
$U_\varepsilon(v) = \{f \in \mathcal{F} \mid f : \varepsilon\text{-covered by } v\}$ とする. このとき,
$\cup_{v \in V} U_\varepsilon(v) = \mathcal{F}$ より以下が成立.

$$
\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i f\left(\boldsymbol{x}_i\right)\right)\right] \\
&= \mathbb{E}\left[\sup_{v \in V} \sup_{f \in U_\varepsilon(v)} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i f\left(\boldsymbol{x}_i\right)\right)\right] \\
&= \mathbb{E}\left[\sup_{v \in V} \sup_{f \in U_\varepsilon(v)} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i v_i + \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(f\left(\boldsymbol{x}_i\right) - v_i\right)\right)\right] \\
&\leq \mathbb{E}\left[\sup_{v \in V} \frac{1}{n} \sum_{i=1}^{n} \sigma_i v_i\right] + \mathbb{E}\left[\sup_{v \in V} \sup_{f \in U_\varepsilon(v)} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(f\left(\boldsymbol{x}_i\right) - v_i\right)\right]
\end{aligned}
$$

カバリングナンバーによる Rademacher Complexity の上界

(Proof of Theorem つづき) ヘルダー不等式を右辺第 2 項に適用:

$$\mathbb{E}\left[\sup_{v\in V}\sup_{f\in U_\varepsilon(v)}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(f\left(\boldsymbol{x}_i\right)-v_i\right)\right]\leq\mathbb{E}\left[\sup_{v\in V}\sup_{n\in U_\varepsilon(v)}\frac{1}{n}\sum_{i=1}^{n}\left|f\left(\boldsymbol{x}_i\right)-v_i\right|\right]$$
$$\leq\varepsilon$$

また, Massart の補題を第 1 項に適用:

$$\begin{aligned}\mathbb{E}\left[\sup_{v\in V}\frac{1}{n}\sum_{i=1}^{n}\sigma_i v_i\right]&\leq\frac{\sup_{v\in V}\|v\|_2\sqrt{2\log|V|}}{n}\\&\leq\sqrt{\frac{2\log|V|}{n}}\\&=\sqrt{\frac{2\log\mathcal{N}_1\left(\varepsilon,\mathcal{F},x_{1:n}\right)}{n}}\end{aligned}$$

二行目は, $v_i\in[-1,1]$, $i=1,...,n$ から従う. 以上より, 定理の主張が示された.

カバリングナンバーによる Rademacher Complexity の上界

Corollary 1

$\mathcal{F} \ni f : \mathcal{X} \to [-1, 1]$ とする. このとき,

$$\mathfrak{R}_n(\mathcal{F}) \le \inf_{\varepsilon} \sqrt{\frac{2 \log \mathcal{N}_1(\varepsilon, \mathcal{F}, n)}{n}} + \varepsilon$$

実際には, covering のスケール ε に関して積分をしたバウンドが用いられる

$\to$ Dudley 積分, Chaining

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.