機械学習とデータマイニングの基礎（大阪大学）

PAC学習とERM学習法

Matthew J. Holland

大阪大学 産業科学研究所

目次

1. 今回の趣旨
2. PAC 学習の概要
3. 有限モデルにおける汎化性能
4. 無限モデルにおける汎化性能
5. まとめ

目次

1. 今回の趣旨
2. PAC 学習の概要
3. 有限モデルにおける汎化性能
4. 無限モデルにおける汎化性能
5. まとめ

今回の趣旨

今回は「学習法の性能評価」を切り口としてある種の「一致性」の定義から出発．

この「性能」を保証できる代表格の学習法として，ERM を見ていく．

モデルの要素が有限であるか無限であるかで，証明の技法が大きくことなるため，
それぞれの状況での性能保証を探求していく．

特に二値分類の場合，無限モデルにおける ERM 学習法とモデルの「VC 次元」は
深い関係にあり，その基本概念や有名な知見を伝授することを目的とする．

目次

1. 今回の趣旨
2. PAC 学習の概要
3. 有限モデルにおける汎化性能
4. 無限モデルにおける汎化性能
5. まとめ

理想と現実（作図用）

- 世の中に流通するデータはノイジーで矛盾などもある．

  → すべてのデータで完璧な成績は望めない．
- どのようなデータが手に入るかも不確実である．

  → 確実にうまく学習する保証は厳しい．

学習法の評価（「標本外の汎化」篇）

最低限の性能とは？（直感）

学習データさえ十分にあれば，十分に良い成績をほぼ確実に出してほしい．

定式化

- n 個の学習データ $\mathbf{Z}_n \coloneqq (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$ に基づく $\mathbf{Z}_n \mapsto \mathrm{H}_n$ を扱う．
- 広義のリスク関数 $\mathrm{R}(\cdot)$ と比較水準 $\mathrm{R}^*_{\mathsf{con}} \leq \mathrm{R}(\cdot)$ で評価する．

学習法の「一致性」(consistency[1])

$$\forall \varepsilon > 0, \quad \lim_{n \to \infty} \mathbf{P}\left\{\mathrm{R}(\mathrm{H}_n) - \mathrm{R}^*_{\mathsf{con}} > \varepsilon\right\} = 0 \tag{1}$$

[1] 確率変数の期待値を推定する問題における標本平均の弱い一致性が special case である．

学習法の評価（「標本外の汎化」篇）

PAC 学習 [2]

$$\underbrace{\textbf{P}\text{robably}}_{\geq 1-\delta}\ \underbrace{\textbf{A}\text{pproximately}}_{\leq \mathrm{R}^*_{\mathsf{con}}+\varepsilon}\ \underbrace{\textbf{C}\text{orrect}}_{\mathrm{R}(\cdot)}$$

PAC(n, ε, δ) 条件

$$\mathbf{P}\left\{\mathrm{R}(\mathrm{H}_n) - \mathrm{R}^*_{\mathsf{con}} > \varepsilon\right\} \leq \delta \tag{2}$$

- 標本数 $n \in \mathbb{N}$，精度 $\varepsilon > 0$，確度 $1 - \delta$（ただし $0 < \delta < 1$）とする．
- 当然，$\mathrm{R}(\cdot)$ と $\mathrm{R}^*_{\mathsf{con}}$ の性質によって，この条件の強度が大きく変わる．
- この (2) は学習法 $\mathbf{Z}_n \mapsto \mathrm{H}_n$ の性質である．

[2] 有名なのは Valiant [10] とその拡張版 [9, 11]．名称は Angluin [1], Angluin and Laird [2] が初見の模様．

学習法の評価（「標本外の汎化」篇）

一致性と PAC 条件

$$\forall (\varepsilon, \delta),\ \exists n < \infty,\ \mathsf{PAC}(n, \varepsilon, \delta) \text{ が成立} \iff (1) \tag{3}$$

標本複雑度（sample complexity）

$$N^*_{\varepsilon,\delta} := \min \{1 \le n \le \infty : \mathsf{PAC}(n, \varepsilon, \delta) \text{ が成立}\} \tag{4}$$

もう少し強い要求[3]

- $N^*_{\varepsilon,\delta} = \mathcal{O}(\mathsf{poly}(1/\varepsilon, 1/\delta))$
- H_nの計算量 $= \mathcal{O}(\mathsf{poly}(1/\varepsilon, 1/\delta))$（ただし $n \ge N^*_{\varepsilon,\delta}$）

[3]多項式のレートは必然ではないが，典型的ではある（Kearns and Vazirani [7] を参照）．

目次

1. 今回の趣旨
2. PAC 学習の概要
3. 有限モデルにおける汎化性能
4. 無限モデルにおける汎化性能
5. まとめ

簡単な問題設定から出発する

表記

- $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$（データの属する集合）
- $h : \mathcal{X} \to \mathcal{Y}$（意思決定の候補）
- $\mathcal{H}$（候補の全体集合）
- $\mathsf{Z} = (\mathsf{X}, \mathsf{Y}) \sim \mu$（評価用データ）
- $\mathsf{Z}_i = (\mathsf{X}_i, \mathsf{Y}_i)$，$i \in [n]$（学習データ）

簡単な問題設定から出発する

事前知識

データの関係性について，以下のことは既知とする.[4]

$$\exists h^* \in \mathcal{H},\quad \mathsf{Y} = h^*(\mathsf{X}) \tag{5}$$

この仮定によって明らかなのは，以下の事実である．

$$\mathbf{P}\{h(\mathsf{X}) \ne \mathsf{Y}\} = \mathbf{P}\{h(\mathsf{X}) \ne h^*(\mathsf{X})\} \tag{6}$$

ある種の学習法のアイディアはすぐに思い浮かぶ．

[4]この等式が確率 1 で成り立つという仮定のほうが少し弱い．

簡単な問題設定から出発する

単純な学習法（サル真似）

1. $\mathsf{H}_0 \in \mathcal{H}$ を選ぶ．
2. 以降，各 $t = 1, 2, \ldots$ に対して，以下の手順を行う．
   - $(\mathsf{X}_t, \mathsf{Y}_t)$ を得る．
   - $\mathsf{H}_{t-1}(\mathsf{X}_t) = \mathsf{Y}_t$ なら，$\mathsf{H}_t = \mathsf{H}_{t-1}$ とする．
   - $\mathsf{H}_{t-1}(\mathsf{X}_t) \neq \mathsf{Y}_t$ なら，$\mathsf{H}_t(\mathsf{X}_t) = \mathsf{Y}_t$ を満たす $\mathsf{H}_t \in \mathcal{H}$ を選ぶ．

最悪時に試す候補の数
$|\mathcal{H}| - 1$ である．
（無限モデル $|\mathcal{H}| = \infty$ なら，終了しないかもしれない）

簡単な問題設定から出発する

統計的な側面（サル真似） 1

- 先述の「サル真似」学習法（$n+1$ 回の更新）の出力：

$$\mathsf{H}_n^{\mathsf{con}} \coloneqq \mathsf{H}_{n+1} \tag{7}$$

- はっきりとわかっていること：

$$\mathsf{H}_n^{\mathsf{con}}(\mathsf{X}_i) = \mathsf{Y}_i, \quad \forall\, i \in [n] \tag{8}$$

- まだわからないこと：
  事象「$\mathbf{P}\left\{\mathsf{H}_n^{\mathsf{con}}(\mathsf{X}) \neq \mathsf{Y}\right\} > \varepsilon$」の生起確率．

簡単な問題設定から出発する

統計的な側面（サル真似） 2

ε の水準では「不合格」の候補を集めてみよう．

$$\mathcal{H}_{\mathsf{bad}}(\varepsilon) \coloneqq \left\{h \in \mathcal{H} : \mathbf{P}\left\{h(\mathsf{X}) \neq \mathsf{Y}\right\} > \varepsilon\right\} \tag{9}$$

直感
$\mathsf{H}_n^{\mathsf{con}} \in \mathcal{H}_{\mathsf{bad}}(\varepsilon)$ はあり得るが，全問正解なので可能性が低い（はず）．

簡単な問題設定から出発する

統計的な側面（サル真似） 3

$$\begin{aligned}
\mathbf{P}\left\{\mathsf{H}_n^{\mathsf{con}} \in \mathcal{H}_{\mathsf{bad}}(\varepsilon)\right\} &\leq \mathbf{P}\left[\bigcup_{h \in \mathcal{H}_{\mathsf{bad}}(\varepsilon)} \left\{h(\mathsf{X}_1) = \mathsf{Y}_1, \ldots, h(\mathsf{X}_n) = \mathsf{Y}_n\right\}\right] \\
&\leq \sum_{h \in \mathcal{H}_{\mathsf{bad}}(\varepsilon)} \mathbf{P}\left\{h(\mathsf{X}_1) = \mathsf{Y}_1, \ldots, h(\mathsf{X}_n) = \mathsf{Y}_n\right\} \\
&= \sum_{h \in \mathcal{H}_{\mathsf{bad}}(\varepsilon)} \prod_{i=1}^{n} \mathbf{P}\left\{h(\mathsf{X}_i) = \mathsf{Y}_i\right\} \\
&= |\mathcal{H}_{\mathsf{bad}}(\varepsilon)| \left(\mathbf{P}\left\{h(\mathsf{X}) = \mathsf{Y}\right\}\right)^n, \quad h \in \mathcal{H}_{\mathsf{bad}}(\varepsilon) \\
&\leq |\mathcal{H}_{\mathsf{bad}}(\varepsilon)| \left(1 - \varepsilon\right)^n
\end{aligned}$$

簡単な問題設定から出発する

統計的な側面（サル真似）4

$$\mathbf{P}\left\{\mathbf{P}\left\{\mathsf{H}_n^{\mathsf{con}}(\mathsf{X}) \neq \mathsf{Y}\right\} > \varepsilon\right\} \leq \mathbf{P}\left\{\mathsf{H}_n^{\mathsf{con}} \in \mathcal{H}_{\mathsf{bad}}(\varepsilon)\right\} \leq |\mathcal{H}_{\mathsf{bad}}(\varepsilon)|\,(1-\varepsilon)^n \leq |\mathcal{H}|\,(1-\varepsilon)^n \leq |\mathcal{H}|\exp(-n\varepsilon)$$

サル真似と PAC 学習

$\mathrm{R}(h) = \mathbf{P}\left\{h(\mathsf{X}) \neq \mathsf{Y}\right\}$，$\mathrm{R}^*_{\mathsf{con}} = \mathrm{R}(h^*) = 0$ とおけば，標本複雑度のバウンドを得る．

$$N^*_{\varepsilon,\delta} \leq \frac{\log(|\mathcal{H}|) + \log(1/\delta)}{\varepsilon} \tag{10}$$

この論法の弱点

いつでも「全問正解が可能である」という前提．

これを外すと，議論が途中で止まってしまう．

簡単な問題設定から出発する（作図用）

ノイジーな関係性 [5]

$$\mathsf{Y} = h^*(\mathsf{X})\mathsf{U}, \qquad \mathsf{Y} = h^*(\mathsf{X}) + \mathsf{U}, \qquad \mathsf{Y} = h^*(\mathsf{X} + \mathsf{U}), \text{ などなど}\cdots$$

X と Y の関係がノイジーであれば，「矛盾」はあり得るし，

普通のモデルでは表現しきれないことが多々ある．

[5] 特に仮定を置かず，PAC(n, ε, δ) で $\mathrm{R}^*_{\mathsf{con}} = \inf_{h\in\mathcal{H}} \mathrm{R}(h)$ とする設定を Agnostic PAC という．

ERM 学習法

サル真似の特徴

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\mathsf{H}_n^{\mathsf{con}}(\mathsf{X}_i) \neq \mathsf{Y}_i\right\} = 0 \tag{11}$$

その一般化（$\ell(\cdot)$ は自由）

$$\mathsf{H}_n^{\mathsf{ERM}} \in \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h; \mathsf{Z}_i) \tag{12}$$

ERM 学習法 [6]

$$\underbrace{\mathbf{E}\text{mpirical}}_{(\mathsf{Z}_1,\ldots,\mathsf{Z}_n)}\ \underbrace{\mathbf{R}\text{isk}}_{\mathbf{E}[\mathsf{L}]}\ \underbrace{\mathbf{M}\text{inimization}}_{\min_{h\in\mathcal{H}}}$$

[6] 統計的学習理論の中枢にあたる学習法．Vapnik and Chervonenkis [12] は画期的である．

ERM 学習法

論法を点検する

ここで「全問正解」を「ERM 解」に緩和してみよう．

元の議論の途中にあった事象を以下のように置き換えざるを得ない．

$$\{h(\mathsf{X}_1) = \mathsf{Y}_1, \ldots, h(\mathsf{X}_n) = \mathsf{Y}_n\}$$

$$\downarrow$$

$$\{h(\mathsf{X}_1) = \mathsf{H}_n^{\mathsf{ERM}}(\mathsf{X}_1), \ldots, h(\mathsf{X}_n) = \mathsf{H}_n^{\mathsf{ERM}}(\mathsf{X}_n)\}$$

$\mathsf{H}_n^{\mathsf{ERM}}(\cdot)$ が全学習データに依存するため，個別の事象の独立性が失われてしまう．

ERM 学習法

表記

$$\mathsf{L}(h) \coloneqq \ell(h;\mathsf{Z}), \quad \mathrm{R}(h) = \mathbf{E}_\mu \mathsf{L}(h), \quad \mathsf{R}_n(h) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell(h;\mathsf{Z}_i) \tag{13}$$

ひとつのアイディア

$\mathsf{Z}_1, \ldots, \mathsf{Z}_n$ と Z が独立同分布ならば，以下の論法はできそう．

- ERM 学習法で $\mathsf{R}_n(\cdot)$ を小さくする．
- 各 $h \in \mathcal{H}$ に対して，$\mathsf{R}_n(h) \approx \mathrm{R}(h)$ を示す．
- ERM 学習法では $\mathrm{R}(\cdot)$ も小さくなることを示す．

肝心の近似の良し悪しについて，もう少し突き詰めていこう．

ERM 学習法

集中不等式（Chebyshev 型）[7]

$$\mathbf{P}\left\{|\mathsf{R}_n(h) - \mathrm{R}(h)| > \varepsilon\right\} \le \frac{\mathbf{E}(\mathsf{R}_n(h) - \mathrm{R}(h))^2}{\varepsilon^2} = \frac{\mathrm{var}_\mu \mathsf{L}(h)}{\varepsilon^2 n} \tag{14}$$

集中不等式（Hoeffding 型）[8]

$\mathbf{P}\{a \le \mathsf{L}(h) \le b\} = 1$ が成り立つ場合，以下のことが言える．

$$\mathbf{P}\left\{|\mathsf{R}_n(h) - \mathrm{R}(h)| > \varepsilon\right\} \le 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right) \tag{15}$$

[7] Markov の不等式から容易に導き出すことができる．
[8] Hoeffding [6] の技法による．Boucheron et al. [4] には明快な証明が収録されている．

ERM 学習法

議論の流れ [9]

例として先ほどの (15) を借用すると，以下の不等式が出せる．

$$\begin{aligned}
\mathbf{P}\left\{|\mathsf{R}_n(\mathsf{H}_n^{\mathsf{ERM}}) - \mathrm{R}(\mathsf{H}_n^{\mathsf{ERM}})| > \varepsilon\right\} &\le \mathbf{P}\left\{\sup_{h\in\mathcal{H}} |\mathsf{R}_n(h) - \mathrm{R}(h)| > \varepsilon\right\} \\
&= \mathbf{P}\left[\bigcup_{h\in\mathcal{H}} \{|\mathsf{R}_n(h) - \mathrm{R}(h)| > \varepsilon\}\right] \\
&\le \sum_{h\in\mathcal{H}} \mathbf{P}\left\{|\mathsf{R}_n(h) - \mathrm{R}(h)| > \varepsilon\right\} \\
&\le 2|\mathcal{H}| \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)
\end{aligned}$$

[9] ここで，ERM の性質を何も使っていないので，任意の学習法にも通じる議論．

ERM 学習法

ERM 全般についてわかっていること

確率 $1-\delta$ 以上で成り立つリスクバウンド．

$$\mathrm{R}(\mathsf{H}_n^{\mathsf{ERM}}) \le \mathsf{R}_n(\mathsf{H}_n^{\mathsf{ERM}}) + (b-a)\sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)} \tag{16}$$

更には，以下の便利な不等式が使える．

$$\begin{aligned}
\mathrm{R}(\mathsf{H}_n^{\mathsf{ERM}}) - \mathrm{R}_{\mathcal{H}}^* &= \mathrm{R}(\mathsf{H}_n^{\mathsf{ERM}}) - \mathsf{R}_n(\mathsf{H}_n^{\mathsf{ERM}}) + \mathsf{R}_n(\mathsf{H}_n^{\mathsf{ERM}}) - \mathrm{R}_{\mathcal{H}}^* \\
&\le 2\sup_{h\in\mathcal{H}} |\mathsf{R}_n(h) - \mathrm{R}(h)|
\end{aligned} \tag{17}$$

結果として，以下の不等式も導き出せる（これも確率 $1-\delta$ 以上）．

$$\mathrm{R}(\mathsf{H}_n^{\mathsf{ERM}}) - \mathrm{R}_{\mathcal{H}}^* \le 2(b-a)\sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)} \tag{18}$$

ERM 学習法

ノイズ有りの任意 ERM 学習法（(18) の再掲）

$$\mathrm{R}(\mathrm{H}_n^{\mathrm{ERM}}) - \mathrm{R}_{\mathcal{H}}^* \leq 2(b-a)\sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}$$

（確率 $1-\delta$ 以上）

ノイズ無しの「サル真似」と比べて（ゼロイチ損失）

$$\mathrm{R}(\mathrm{H}_n^{\mathrm{con}}) \leq \frac{1}{n}\log\left(\frac{|\mathcal{H}|}{\delta}\right) \tag{19}$$

（確率 $1-\delta$ 以上）

目次

1. 今回の趣旨
2. PAC 学習の概要
3. 有限モデルにおける汎化性能
4. 無限モデルにおける汎化性能
5. まとめ

無限モデルにおける汎化性能

この論法の限界
$|\mathcal{H}| = \infty$ の場合，先述の (16) と (19) はいずれも破綻する（保証として無意味）．

新たな切り口
この $\mathcal{H}$ の表現力を単なる「候補の数」とするのはいささか乱暴なので，
表現力の指標をもう少し工夫してから議論を再建してみよう．

モデル表現力の新しい指標

問題設定を少し狭める
$\mathcal{Y} = \{-1, +1\}$ とおく．つまり二値分類に限定する．

直感
二値ラベルを割り当てる「自由度」が高いほど，表現力が高い．

モデル表現力の新しい指標

識別能力の指標
ここで n 個の入力データを $\mathbf{x}_n \coloneqq (x_1, \ldots, x_n)$ と記す．

$$\mathrm{shatter}(n; \mathcal{H}, \mathcal{X}) \coloneqq \max_{\mathbf{x}_n \in \mathcal{X}^n} |\{\underbrace{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}}_{\text{dichotomies of } \mathcal{H}}\}| \tag{20}$$

要点

- $\mathcal{H}$ によって割り当てられる二値ラベルの組み合わせの数．
- $1 \leq \mathrm{shatter}(n) \leq 2^n$ は常に成り立つ．
- ある m で $\mathrm{shatter}(m) < 2^m$ なら，任意の $n > m$ でも $\mathrm{shatter}(n) < 2^n$．
- 名称として shatter coefficient や growth function などと呼ばれる．

モデル表現力の新しい指標（作図用）

例：実数直線状の有限区間
$\mathrm{shatter}(3) < 2^3$

モデル表現力の新しい指標（作図用）

例：二次元平面における識別線
$\mathrm{shatter}(4) < 2^4$

モデル表現力の新しい指標（作図用）

例：$\mathbb{R}^d$ における長方形
$\mathrm{shatter}(2d+1) < 2^{2d+1}$

モデル表現力の新しい指標

VC 次元の定義 [10]

$$\mathrm{VC}(\mathcal{H};\mathcal{X}) := \max\{1 \le n \le \infty : \mathrm{shatter}(n;\mathcal{H},\mathcal{X}) = 2^n\} \tag{21}$$

先ほどの例

- 実数直線状の有限区間：$\mathrm{VC} = 2$
- 二次元平面における識別線：$\mathrm{VC} = 3$
- $\mathbb{R}^d$ における長方形：$\mathrm{VC} = 2d$

[10] **V**apnik-**C**hervonenkis の研究成果に由来 [12]．入門書として Devroye et al. [5, Ch. 12–14] は傑作．

モデル表現力の新しい指標

面白い性質 [11]
任意の $n > 2\,\mathrm{VC}(\mathcal{H};\mathcal{X})$ に対して，以下が成り立つ．

$$\mathrm{shatter}(n;\mathcal{H},\mathcal{X}) \le \left(\frac{en}{\mathrm{VC}(\mathcal{H};\mathcal{X})}\right)^{\mathrm{VC}(\mathcal{H};\mathcal{X})} \tag{22}$$

また，$\mathrm{shatter}(n) \le n^{\mathrm{VC}} + 1$ は常に成り立つ．

つまり，$\mathrm{shatter}(\cdot)$ は「指数増加」と「多項式増加」のどれかしかあり得ない．

[11] Devroye et al. [5, Thm. 13.3] を参照．1970 年代までに様々な分野でほぼ同時に明らかになった．

モデル表現力の新しい指標

無限モデルの性能保証にどう役立つ？（直感）

$\mathrm{R}_n(h) \approx \mathrm{R}(h)$ をナイーブに全 $h \in \mathcal{H}$ で抑えると $|\mathcal{H}|$ 依存は不可避

↓

$\mathcal{H}$ が無限でも，二値分類の仕方は有限であるため，
一致する候補同士は同一視しても良い

↓

$|\mathcal{H}|$ を $\mathrm{shatter}(n;\mathcal{H},\mathcal{X})$ に置き換える

モデル表現力の新しい指標

無限モデルの性能保証にどう役立つ？ [12]
ゼロイチ損失の下，以下のことが言える（証明はすべて割愛）．

$$\mathbf{P}\left\{\sup_{h\in\mathcal{H}}|\mathrm{R}_n(h) - \mathrm{R}(h)| > \varepsilon\right\} \le 8\,\mathrm{shatter}(n;\mathcal{H},\mathbb{R}^d)\exp(-n\varepsilon^2/32) \tag{23}$$

さらに，(17) を使うと，以下のこともわかる．

$$\mathbf{P}\left\{\mathrm{R}(\mathsf{H}_n^{\mathsf{ERM}}) - \mathrm{R}^*_{\mathcal{H}} > \varepsilon\right\} \le 8\,\mathrm{shatter}(n;\mathcal{H},\mathbb{R}^d)\exp(-n\varepsilon^2/128) \tag{24}$$

[12] Devroye et al. [5, Thm. 12.6] を参照．

モデル表現力の新しい指標

VC 次元と ERM の性能

VC 次元が有限である $\implies$ 無限モデルでも ERM 学習法の性能保証が得られる

ERM 学習法の平均的な性能 [13]

$$\mathbf{E}\left[\mathrm{R}(\mathrm{H}_n^{\mathrm{ERM}}) - \mathrm{R}_{\mathcal{H}}^*\right] \leq 16\sqrt{\frac{\log(8\mathrm{e}) + \log(\mathrm{shatter}(n;\mathcal{H},\mathbb{R}^d))}{n}} \tag{25}$$

[13] Devroye et al. [5, Cor. 12.1] を参照.

モデル表現力の新しい指標

ERM 学習法の標本複雑度 [14]

$$N_{\varepsilon,\delta}^* \leq \max\left\{\mathcal{O}\left(\frac{\mathrm{VC}}{\varepsilon^2}\log\left(\frac{\mathrm{VC}}{\varepsilon^2}\right)\right), \mathcal{O}\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)\right\} \tag{26}$$

ノイズ無しの「サル真似」と比べて

$$N_{\varepsilon,\delta}^* \leq \frac{\log(|\mathcal{H}|) + \log(1/\delta)}{\varepsilon} \tag{27}$$

[14] Devroye et al. [5, Cor. 12.3] を参照.

PAC 学習可能性

学習可能性 [15]

任意の分布 μ の下，$N_{\varepsilon,\delta}^* < \infty$ を満たすような学習法が存在すれば，
その $\mathcal{H}$ を「学習可能 (learnable)」という．

VC 次元との関係（二値分類の場合） 1

先述の (26) からすぐにわかるのは，

$\mathcal{H}$ の VC 次元が有限である $\implies$ $\mathcal{H}$ は学習可能である

ということである．

[15] 大いに研究されてきた．成熟した理論を Shalev-Shwartz et al. [8] はうまくまとめている．

PAC 学習可能性

VC 次元との関係（二値分類の場合） 2

さらに面白いのは，その逆である．任意の学習法に対して，

$$\mathrm{VC}(\mathcal{H};\mathcal{X}) = \infty \implies N_{\varepsilon,\delta}^* = \infty \text{ となる } \mu \text{ が存在する.}^{16}$$

つまり「学習可能性は VC 次元の有限性によって完全に特徴づけられる」．

（これは「**学習理論の基本定理**」と呼ばれることがある）

[16] Anthony and Bartlett [3, Thm. 5.2] を参照.

目次

1. 今回の趣旨
2. PAC 学習の概要
3. 有限モデルにおける汎化性能
4. 無限モデルにおける汎化性能
5. まとめ

まとめ

- 推定量の「一致性」という概念は学習法にも拡張できる．
- 特にデータや計算のコストを考慮した一致性の探求は PAC 学習．
- 従来の学習理論は $\mathrm{R}(h) = \mathbf{E}_\mu \mathrm{L}(h)$ が基軸となっている．
- 計算量を捨象し，ERM 学習法の統計的な性質（リスク上界を求めるなど）は統計的学習理論の中心トピック．
- 有限モデルと無限モデルにおける ERM の解析は手段が大きく異なるが，基本的な議論のアイディアは共通している．
- 有限モデルでも無限モデルでも，R_n をめぐる「集中不等式」が登場する．
- VC 次元は二値分類問題の「解きやすさ」を特徴づけている．

参考文献

[1] Angluin, D. (1988). Queries and concept learning. *Machine Learning*, 2(4):319–342.

[2] Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, 2(4):343–370.

[3] Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

[4] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

[5] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

[6] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

[7] Kearns, M. J. and Vazirani, U. V. (1994). *An Introduction to Computational Learning Theory*. MIT Press.

[8] Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670.

[9] Valiant, L. G. (1984a). Deductive learning. *Philosophical Transactions of the Royal Society of London, Series A*, 312(1522):441–446.

[10] Valiant, L. G. (1984b). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

[11] Valiant, L. G. (1985). Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 560–566. Morgan Kaufmann.

[12] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.