統計的学習理論読み（Chapter 1）

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

導入 I

- 機械学習の文脈でよく見る
  - モデルが汎化する
  - モデルの汎化性能が高い
- 予測損失 (期待損失) が小さいことと定義される
- practical にはテスト誤差 (学習データとは独立に取得したテストデータで評価した誤差) で汎化性能を評価している

素朴な疑問 (学習理論が答えようとしていること)

- 上記の方法はどのように正当化されているのか?
- 経験損失最小化でなぜ予測損失を小さくできるのか?
- 経験損失と予測損失にどんなギャップがあるのか?

導入 II

Vapnik の思想 [Vapnik, 98]

Nothing is more practical than a good theory.

理論に基づいたアルゴリズム

- カーネル法 (サポートベクターマシン)
- ブースティング (アダブースト)...

このセミナーでは [4] を読んで機械学習の理論的側面に親しみたい.

本スライドは [4] の第 1 章のまとめである.

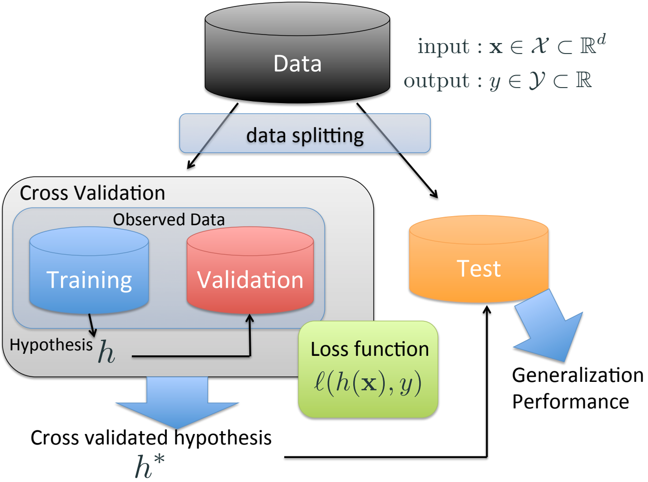
Table of contents

統計的学習理論の枠組み

統計的学習理論の枠組み

1.1 問題設定

問題設定 I (p. 1~3)


Data
input : $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$
output : $y \in \mathcal{Y} \subset \mathbb{R}$
data splitting
Cross Validation
Observed Data
Training
Validation
Hypothesis h
Test
Loss function
$\ell(h(\mathbf{x}), y)$
Generalization
Performance
Cross validated hypothesis
h^*

問題設定 II 判別問題 (§1.1.1)

- $|\mathcal{Y}| < \infty$ のとき, input data から label を予測する.
  - $|\mathcal{Y}| = 2$: 2 値判別 (e.g. 迷惑メール分類 $\mathcal{Y} = \{\text{“}spam\text{”}, \text{“}nonspam\text{”}\}$)
  - $|\mathcal{Y}| \geq 3$: 多値判別
- 判別問題における loss function (0-1 loss)

$$\ell(\hat{y}, y) = \mathbb{1}[\hat{y} \neq y] = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{otherwise} \end{cases}$$

$$\left(= \begin{cases} \ell_y & \text{if } y \neq \hat{y} \\ 0 & \text{otherwise} \end{cases} \right)$$

損失が真ラベルに依存する場合

問題設定 III 回帰問題 (§1.1.2)

- $\mathcal{Y} = \mathbb{R}$ のとき input から output を予測 (e.g. 株価や電力需要の予測)
- 回帰問題の loss function (squared loss)

$$\ell(\hat{y}, y) = |\hat{y} - y|^2$$

問題設定 IV ランキング問題 (§1.1.3)

- 3 つ組 data $(\boldsymbol{x}, \boldsymbol{x}', y) \in \mathcal{X}^2 \times \mathcal{Y}$ を観測

$$y = \begin{cases} +1 & \text{if } \boldsymbol{x} \succ \boldsymbol{x}' \\ -1 & \text{if } \boldsymbol{x} \prec \boldsymbol{x}' \end{cases}$$

- 以下のような仮説 $h : \mathcal{X} \to \mathbb{R}$ を学習

$$\boldsymbol{x} \succ \boldsymbol{x}' \Rightarrow h(\boldsymbol{x}) > h(\boldsymbol{x}')$$
$$\boldsymbol{x} \prec \boldsymbol{x}' \Rightarrow h(\boldsymbol{x}) \leq h(\boldsymbol{x}')$$

- ランキング問題の loss function (0-1 loss)

$$\ell(\hat{h}, y) = \begin{cases} 1 & \text{if } y(h_1 - h_2) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

ここで $h_1 = h(\boldsymbol{x})$, $h_2 = h(\boldsymbol{x}')$, $\hat{h} = (h_1, h_2) \in \mathbb{R}^2$.
$0 - 1$ 損失の下でランキング問題は判別として扱える.

統計的学習理論の枠組み

1.2 予測損失と経験損失

予測損失と経験損失 I

Definition 1 (予測 (期待) 損失)

test data (X, Y) の従う分布 $\mathcal{D}$ の下での仮説 h の予測損失を以下で定義

$$R(h) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h(X), Y)]$$

Example 1 (0-1 loss)

0-1 loss の予測損失 (期待判別誤差) は

$$R_{err}(h) = \Pr[h(X) \neq Y] = \mathbb{E}[\mathbb{1}[h(X) \neq Y]]$$

学習の目標

data の真の分布が未知なため直接計算不可能な期待損失を観測 data のみを用いて小さくする

予測損失と経験損失 II

Definition 2 (経験損失)

$\{(X_i, Y_i)\}_{i=1}^n$: *observed data*
仮説 h の経験損失を以下で定義

$$\hat{R}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

<u>経験分布による表現</u>

$\hat{\mathcal{D}}$: 経験分布 i.e. $(X, Y) \sim \mathcal{D} \Longleftrightarrow \Pr[(X, Y) = (X_i, Y_i)] = \frac{1}{n}$
とするとき,

$$\hat{R}(h) = \mathbb{E}_{(X,Y)\sim\hat{\mathcal{D}}}[\ell(h(X), Y)]$$

予測損失 $R(h)$ と経験損失 $\hat{R}(h)$ の違いは期待値を真の分布 $\mathcal{D}$ で取るか, 経験分布 $\hat{D}$ で取るかの違い

予測損失と経験損失 III

Fact 1

$(X_i, Y_i) \sim \mathcal{D}$ *(identically distributed)*

$$\implies \mathbb{E}[\hat{R}(h)] = R(h)$$

i.e. $\hat{R}$ は R の不偏推定量.

$(\because)$ $\mathcal{D}^n$: (X_i, Y_i), $i = 1, \ldots, n$ の joint distribution とするとき,

$$\mathbb{E}_{\mathcal{D}^n}[\hat{R}(h)] = \mathbb{E}_{\mathcal{D}^n}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_{\mathcal{D}}[\ell(h(X_i), Y_i)]}_{R(h)} = R(h) \quad \square$$

予測損失と経験損失 IV

経験損失は予測損失の不偏推定量: $\mathbb{E}[\hat{R}(h)] = R(h)$

- 上の事実は data の独立性を仮定していない. 独立性があると, さらに一致性が示せる(大数の弱法則) :

Proposition 1

$(X_i, Y_i) \sim_{i.i.d.} \mathcal{D}$ のとき, $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{Pr}_{\mathcal{D}^n}[|\hat{R}(h) - R(h)| > \varepsilon] = 0$$

- 様々な学習問題は, 予測損失 R の最小化が目標 (分布 $\mathcal{D}$ が未知なので R も未知)
$\longrightarrow$ 代理として経験損失 $\hat{R}$ の最小化を通して R を小さくする

統計的学習理論の枠組み

1.3 ベイズ規則とベイズ誤差

ベイズ規則とベイズ誤差 I

Definition 3 (Bayes error / Bayes rule)

- ℓ : *loss* 関数
- $\mathcal{H}_{all}$: 可測関数全体

のとき, *Bayes error* は予測誤差の最小値を達成する仮説:

$$\text{Bayes error} := \inf_{h \in \mathcal{H}_{all}} R(h)$$

また, *Bayes error* を達成する仮説 h_0 を *Bayes rule* という *i.e.*

$$R(h_0) = \text{Bayes error}$$

ベイズ規則とベイズ誤差 II

Bayes rule を具体的に求めてみる.

- $\ell(\hat{y}, y)$: loss 関数
- P : test distribution

とするとき,

$$R(h) = \mathbb{E}_{(X,Y)\sim P}[\ell(\hat{y}, y)] = \mathbb{E}_X\left[\mathbb{E}_Y\left[\ell(\hat{y}, y)|X\right]\right]$$

$(\because)$

$$\begin{aligned}\mathbb{E}_X[\underbrace{\mathbb{E}_Y\left[\ell(h(\boldsymbol{x}), y)|X\right]}_{(\diamond)}] &= \int_{\mathcal{X}}\left\{\int_{\mathcal{Y}}\ell(h(\boldsymbol{x}), y)dP(y|x)\right\}dP(x)\\ &= \int_{\mathcal{X}\times\mathcal{Y}}\ell(h(\boldsymbol{x}), y)dP(x, y)\\ &= R(h) \quad \square\end{aligned}$$

積分の単調性から $(\diamond)$ を小さくする h を選べば予測損失も小さくなる

Example 1.1 判別問題

- 0-1loss を用いると,

$$(\diamond) = \sum_{y \in \mathcal{Y}} \ell(h(X), Y) P(Y = y|X) = 1 - P(Y = h(X)|X)$$

  より,

$$h_0(X) = \arg\max_{y \in \mathcal{Y}} P(Y = y|X)$$

  が予測誤差を最小にする仮説 (input に対して最も出現確率の大きなラベルを出力)

- このときの Bayes error は

$$R^* = 1 - \mathbb{E}_X \left[\max_{y \in \mathcal{Y}} P(Y = y|X) \right]$$

Example 1.2 回帰問題

- 2 乗 loss を用い, Y の分散を $V[Y]$ とおくと,

$$
\begin{aligned}
\mathbb{E}_Y[\ell(h,Y)] &= \mathbb{E}[h^2 - 2hY + Y^2] \\
&= \mathbb{E}[h^2] - 2\mathbb{E}[hY] + \mathbb{E}[Y^2] + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\
&= h^2 - 2h\mathbb{E}[Y] + \mathbb{E}[Y]^2 + \underbrace{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}_{V[Y]} \\
&= (h - \mathbb{E}[Y])^2 + V[Y]
\end{aligned}
$$

第 1 項を最小にする h が Bayes rule

- このとき, Bayes error は

$$
\begin{aligned}
R^* = R(h_0) &= \mathbb{E}_X[\underbrace{\mathbb{E}_Y[\ell(h_0(X),Y)|X]}_{V[Y|X]}] \\
&= \mathbb{E}[V[Y|X]]
\end{aligned}
$$

条件付き分散が一定値 σ^2 ならば, Bayes error も σ^2

Example 1.3 ランキング問題 I

ランキングを 2 値判別として定式化すると, 仮説空間が

$$\mathcal{H} = \{\mathrm{sign}(h(\boldsymbol{x}) - h(\boldsymbol{x}'))\}$$

なる形の関数空間に制限される.
$\to$ 2 値判別の Bayes rule からランキングの Bayes rule は構成できない
$\to$ data 分布に仮定をおき, Bayes rule を特徴づける

設定

- input を $(\boldsymbol{x}_+, \boldsymbol{x}_-) \in \mathcal{X}^2$ とおき, 常に $\boldsymbol{x}_+ \succ \boldsymbol{x}_-$, $y = +1$ とする
- もし $(\boldsymbol{x}, \boldsymbol{x}', -1)$ なる data があれば $(\boldsymbol{x}', \boldsymbol{x}, +1)$ と変換
- $\boldsymbol{x}_+ \sim_{i.i.d.} \mathcal{D}_+$, $\boldsymbol{x}_- \sim_{i.i.d.} \mathcal{D}_-$ とし, ランキング関数 $h : \mathcal{X} \to \mathbb{R}$ を学習

Example 1.3 ランキング問題 II

Definition 4 (true positive rate / false positive rate)

しきい値 $a \in \mathbb{R}$ に対して,

$$TP_h(a) := \mathbb{E}_{\boldsymbol{x}_+ \sim \mathcal{D}_+}[\mathbb{1}[h(\boldsymbol{x}_+) > a]]$$
$$FP_h(a) := \mathbb{E}_{\boldsymbol{x}_- \sim \mathcal{D}_-}[\mathbb{1}[h(\boldsymbol{x}_-) > a]]$$

- $TP_h(a)$: しきい値 a において *positive sample* を正しく *positive* と判定出来ている割合.
- $FP_h(a)$: しきい値 a において *negative sample* を誤って *positive* と判定している割合.

$a \in \mathbb{R}$ に対して, $(FP_h(a), TP_h(a)) \in [0, 1]^2$

Example 1.3 ランキング問題 III

Definition 5 (ROC curve)

$a \to \infty$ とするとき, $(FP_h(a), TP_h(a))$ は $(0,0) \to (1,1)$ と動く. その軌跡の描く曲線を *ROC curve* という

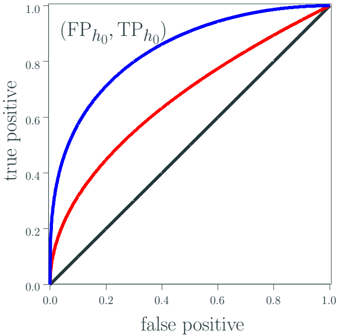


Figure 1: "統計的学習理論" 図 1.3 より抜粋

- AUC : ROC curve と $(1,0)$ で囲まれる領域の面積
- ランダムな仮説 (TP=FP, 45 度直線) は $AUC = 0.5$
- AUC が大きいほど TP が大きいので良い

期待損失と AUC との関係

0-1 loss の下で,

$$
\begin{aligned}
R(h) &= 1 - \mathbb{E}_{\boldsymbol{x}_\pm \sim \mathcal{D}_\pm}\left[\mathbb{1}[h(\boldsymbol{x}_+) - h(\boldsymbol{x}_-) > 0]\right] \\
&= 1 - \mathbb{E}_{\boldsymbol{x}_- \sim \mathcal{D}_-}\Big[\underbrace{\mathbb{E}_{\boldsymbol{x}_+ \sim \mathcal{D}_+}\left[\mathbb{1}[h(\boldsymbol{x}_+) > h(\boldsymbol{x}_-)]\right]}_{TP_h(h(\boldsymbol{x}_-))}\Big] \\
&= 1 - \underbrace{\mathbb{E}_{\boldsymbol{x}_- \sim \mathcal{D}_-}\left[TP_h(h(\boldsymbol{x}_-))\right]}_{AUC(h)} \\
&= 1 - AUC(h)
\end{aligned}
$$

よって $\boldsymbol{x}_+ \perp\!\!\!\perp \boldsymbol{x}_-$ のとき,

- $h_0 = \arg\max AUC(h)$
- Bayes error $= 1 - AUC(h_0)$

統計的学習理論の枠組み

1.4 学習アルゴリズムの性能評価

Definition 6 (学習アルゴリズム)

学習アルゴリズムは観測データ集合から仮説集合への *map* :

$$
\begin{aligned}
\mathcal{A} :& 2^{\mathcal{X}\times\mathcal{Y}} \longrightarrow \mathcal{H} \\
& S \mapsto \mathcal{A}(S) = h_S
\end{aligned}
$$

ここで, $S = \{(X_i, Y_i)\}_{i=1}^{n}$

$\mathcal{A}$ の性能の評価指標

1. 予測損失の学習データに関する期待値をとる:

$$\mathbb{E}_{S\sim\mathcal{D}^n}[R(h_S)]$$

$\longrightarrow$ $\mathcal{A}$ の平均的な性能を評価

2. 汎化誤差の分布を評価:
Bayes error を $R^* = \inf R(h)$ とおく. $\varepsilon > 0$ と $\delta \in (0,1)$ に対して

$$\Pr[R(h_S) - R^* < \varepsilon] > 1 - \delta$$

が成り立つとする.
$\to$ 十分大きい確率 $1-\delta$ に対して ε を十分小さく取れれば
Bayes error に近い予測損失を達成する仮説が求まる

Fact 2 (評価指標 1 と 2 の関係)

$$P_{S\sim\mathcal{D}^n}[R(h_S) - R^* \geq \varepsilon] \leq \frac{\mathbb{E}_{S\sim\mathcal{D}^n}[R(h_S)] - R^*}{\varepsilon}$$

- 予測損失と *Bayes error* の差が ε 以上である確率は, 予測損失の期待値と *Bayes error* の差で上から抑えられる

($\because$) Markov's inequality :

$$P(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}, \quad a > 0$$

より, $|X| = R(h_S) - R^*$, $a = \varepsilon$ とおくと直ちに従う $\square$

学習アルゴリズムの性能評価 IV

Definition 7 (統計的一致性)

$\forall \mathcal{D}$: *distribution*, $\forall \varepsilon > 0$ に対して, 学習アルゴリズム $\mathcal{A} : S \mapsto h_S$ が統計的一致性をもつ

$$:\Longleftrightarrow \lim_{n \to \infty} P_{S \sim \mathcal{D}^n}[R(h_S) - R^* \leq \varepsilon] = 1$$

“data が多ければ最適な仮説を達成する” という良い学習アルゴリズムの性質

統計的学習理論の枠組み

1.5 有限仮説集合を用いた学習

予測判別誤差 (0-1 loss の汎化誤差) の評価 I

問題設定

- 2 値判別問題 (ℓ : 0-1 loss)
- 有限仮説集合: $\mathcal{H} := \{h_1, ..., h_T\}, \quad h_t : \mathcal{X} \to \{+1, -1\}$
- 学習データ: $S = \{(X_i, Y_i)\}_{i=1}^n, \quad (X_i, Y_i) \sim_{i.i.d.} P$

このとき, 学習アルゴリズムとして経験判別誤差を最小にする仮説を出力するものを考える:

$$
\begin{aligned}
\mathcal{A} :& 2^{\mathcal{X}\times\mathcal{Y}} \to \mathcal{H} \\
& S \mapsto \mathcal{A}(S) = h_S = \arg\min_{h\in\mathcal{H}} \underbrace{\hat{R}_{err}(h)}_{\frac{1}{n}\sum_{i=1}^n \ell(h(X_i),Y_i)}
\end{aligned}
$$

分布 P の下での 0-1 loss に関する Bayes rule を h_0 とする (一般に $h_0 \notin \mathcal{H}$)

予測判別誤差 (0-1 loss の汎化誤差) の評価 II

予測判別誤差と Bayes error の gap

$$R_{err}(h_S) - R_{err}(h_0)$$

を評価.

いま, $h_{\mathcal{H}} := \underset{h \in \mathcal{H}}{\arg\min}\, R_{err}(h)$ とおくと以下が成立:

- $\underbrace{R_{err}(h_0)}_{\text{全可測関数で min}} \le \underbrace{R_{err}(h_{\mathcal{H}})}_{\mathcal{H}\text{ 内で min}} \le R_{err}(h_S)$
- $\hat{R}_{err}(h_S) \le \hat{R}_{err}(h_{\mathcal{H}})$

予測判別誤差 (0-1 loss の汎化誤差) の評価 III

$$
\begin{aligned}
& R_{err}(h_S) - R_{err}(h_0) \\
&= R_{err}(h_S) - \hat{R}_{err}(h_S) + \hat{R}_{err}(h_S) - R_{err}(h_{\mathcal{H}}) + R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) \\
&\leq R_{err}(h_S) - \hat{R}_{err}(h_S) + \hat{R}_{err}(h_{\mathcal{H}}) - R_{err}(h_{\mathcal{H}}) + R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) \\
&\leq \max_h |\hat{R}_{err}(h) - R_{err}(h)| + \max_h |\hat{R}_{err}(h) - R_{err}(h)| + R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) \\
&= 2 \max_h |\hat{R}_{err}(h) - R_{err}(h)| + R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) \quad - (\diamond)
\end{aligned}
$$

ここで $(\diamond)$ の第 1 項に Hoeffding's inequality を使う

Lemma 1 (Hoeffding's inequality)

Z : [0,1]-valued r.v. で $Z_1, \ldots, Z_n \sim_{i.i.d.} P_Z$ のとき, $\varepsilon > 0$,

$$
P\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2}
$$

予測判別誤差 (0-1 loss の汎化誤差) の評価 IV

Hoeffding's inequality の Z として $\mathbb{1}[h(X) \neq Y]$ を取ると,

$$
\begin{aligned}
P\left[2 \max_{h \in \mathcal{H}} |\hat{R}_{err}(h) - R_{err}(h)| \geq \varepsilon\right] &\leq \sum_{h \in \mathcal{H}} \underbrace{P\left[|\hat{R}_{err}(h) - R_{err}(h)| \geq \frac{\varepsilon}{2}\right]}_{\leq 2e^{-2n\varepsilon^2/4}} \\
&\leq 2|\mathcal{H}|e^{-n\varepsilon^2/2}
\end{aligned}
$$

ここで, $\delta = 2|\mathcal{H}|e^{-n\varepsilon^2/2}$ とおくと, 学習データ S が given の下で

$$
P\left[R_{err}(h_S) - R_{err}(h_0) \leq R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) + \sqrt{\frac{2}{n} \log \frac{2|\mathcal{H}|}{\delta}}\right] \geq 1 - \delta
$$

が成立.

予測判別誤差 (0-1 loss の汎化誤差) の評価 V

$$P\left[R_{err}(h_S) - R_{err}(h_0) \leq R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) + \sqrt{\frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}}\right] \geq 1-\delta$$

$(\because)$

$$\begin{aligned}\delta = 2|\mathcal{H}|e^{-n\varepsilon^2/2} &\iff \frac{\delta}{2|\mathcal{H}|} = e^{-n\varepsilon^2/2} \\ &\iff \log\frac{\delta}{2|\mathcal{H}|} = \frac{-n\varepsilon^2}{2} \\ &\iff \varepsilon^2 = \frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}\end{aligned}$$

より,

$$\begin{aligned} & P\left[2\max_{h\in\mathcal{H}}|\hat{R}_{err}(h) - R_{err}(h)| \geq \varepsilon\right] \leq 2|\mathcal{H}|e^{-n\varepsilon^2/2} \\ \iff & P\left[2\max_{h\in\mathcal{H}}|\hat{R}_{err}(h) - R_{err}(h)| \leq \sqrt{\frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}}\right] \geq 1-\delta\end{aligned}$$

予測判別誤差 (0-1 loss の汎化誤差) の評価 VI

($\diamond$) の第 1 項を上の評価で置き換えると,

$$R_{err}(h_S) - R_{err}(h_0) \leq \sqrt{\frac{2}{n} \log \frac{2|\mathcal{H}|}{\delta}} + R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) \quad w.p.\ 1 - \delta$$

が言える $\square$

- 仮説集合 $\mathcal{H}$ が Bayes rule を含むとき ($h_{\mathcal{H}} = h_0$ のとき):

$$\begin{aligned} & R_{err}(h_{\mathcal{H}}) - R_{err}(h_0) = 0 \\ \Longrightarrow \quad & R_{err}(h_S) \longrightarrow R_{err}(h_0) \ \text{ as } n \to \infty \end{aligned}$$

- 確率オーダー表記 (cf 例 2.1):

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

i.e. $\lim_{z\to\infty} \limsup_{n\to\infty} P[|R_{err}(h_S)|/\sqrt{\log |\mathcal{H}|/n} > z] = 0$

近似誤差と推定誤差 I

Definition 8 (近似誤差 (bias) / 推定誤差 (variance) 分解)

評価式

$$R_{err}(h_S) - R_{err}(h_0) \le \sqrt{\frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}} + R_{err}(h_{\mathcal{H}}) - R_{err}(h_0)$$

において, 近似誤差 *(bias)* と推定誤差 *(var)* を以下で定義.

$$bias_{\mathcal{H}} := R_{err}(h_{\mathcal{H}}) - R_{err}(h_0)$$

$$var_{\mathcal{H}} := \sqrt{\frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}}$$

- bias はモデルが外れている (Bayes rule を含まない) ことで生じる誤差 (一般に $h_0 \notin \mathcal{H}$ より $bias_{\mathcal{H}} \ge 0$)
- var は学習データ (サンプルサイズ) に由来するばらつき

近似誤差と推定誤差 II

bias-variance trade-off

仮説空間の増大列 $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_M$, $|\mathcal{H}_M| < \infty$ に対して

$$bias_{\mathcal{H}_1} \geq \cdots \geq bias_{\mathcal{H}_M}, \quad var_{\mathcal{H}_1} \leq \cdots \leq var_{\mathcal{H}_M}$$

- 仮説空間が広いほど Bayes rule に近い仮説が手に入りやすい
- サンプルサイズを止めて $\mathcal{H}$ を広げるとばらつきが増大
- サンプルサイズが十分大
  $\Rightarrow$ 大きな $\mathcal{H}$ でも var は bias に対して大きくない
- サンプルサイズが小さい
  $\Rightarrow$ var は $\mathcal{H}$ の大きさの影響を受けやすい

近似誤差と推定誤差 III

予測誤差を小さくする仮説集合 $\mathcal{H}_{\hat{m}}$ として, 以下を満たすものが良さそう

$$\hat{m} = \underset{1 \leq m \leq M}{\arg\min} \left[bias_{\mathcal{H}_m} + var_{\mathcal{H}_m} \right]$$

- bias が data 分布に依存するため上手い基準ではない
  $\rightarrow$ 正則化

正則化 I

アイデア: 大きな仮説集合から仮説を選ぶことに対してペナルティを課す

Definition 9 (ペナルティ関数)

仮説集合の増大列 $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_M$. $\Phi : \mathcal{H}_m \to \mathbb{R}_{\geq 0}$ が仮説 h に対するペナルティ関数

$:\Longleftrightarrow m_1 < m_2$ に対して, $h \in \mathcal{H}_{m_1}$, $h' \in \mathcal{H}_{m_2} \backslash \mathcal{H}_{m_1} \Rightarrow \Phi(h) \leq \Phi(h')$

Example 2 (大きい仮説集合ほどペナルティも大きい)

$\mathcal{H}_0 = \emptyset$ として, $0 < w_1 < \cdots < w_M$ に対して

$$\Phi(h) = \sum_{m=1}^{M} w_m \mathbb{1}[h \in \mathcal{H}_m \backslash \mathcal{H}_{m-1}]$$

正則化 II

正則化付き経験誤差最小化

$$\min_{h \in \mathcal{H}_M} \hat{R}_{err}(h) + \lambda \Phi(h)$$

- 想定する最大の仮説空間で最適化を実行
- λ の決め方:
  - data 数に依存させ, 適切なオーダーで $\lambda_n \to 0$ as $n \to \infty$ とする
  - クロスバリデーション

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), 2015.