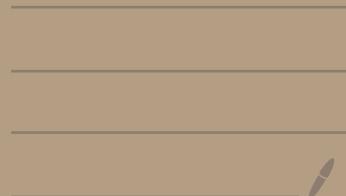


# 統計的學習理論

---

## Chapter 1

### 統計的學習理論，粹種



## □ Notation

$\mathbb{R}$ : 実数の集合,  $\mathbb{R}_{\geq 0}$ : 非負実数の集合,  
 $\mathbb{N}$ : 自然数の集合.

$$\|\cdot\|: 2 - 1 \text{ ノルム}$$

$$\|\cdot\|_1: 1 - 1 \text{ ノルム}$$

$$\|\cdot\|_\infty: \infty - 1 \text{ ノルム}$$

$\|\cdot\|_H$ : 再生核 Hilbert 空間  $H$  上の内積から誘導されたノルム

$|S|$ : 集合  $S$  の要素数

$$\mathbb{I}[A] = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases} : \text{定義関数}$$

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases} : \text{符号関数}$$

$$\text{lerr}(\hat{y}, y) = \mathbb{I}[\hat{y} \neq y] = 0 - 1 \text{ 損失}$$

$$\phi_{\text{hinge}}(m) = \max \{1 - m, 0\} : \text{ヘンジ損失}$$

$$\Phi_p(m) = \begin{cases} 1 & \text{if } m \leq 0 \\ 1 - m/p & \text{if } 0 \leq m \leq p \\ 0 & \text{if } m \geq p \end{cases}$$

$y f(x)$ : 判別関数  $f: X \rightarrow \mathbb{R}$  ( $\infty$  可能)  
 $(x, y)$  における 2 値マシン.

$$\text{mrg}(f; x, y) = f(x, y) - \max_{y': y' \neq y} f(x, y')$$

多個マシン

損失・誤差 (経験損失・経験誤差 ( $\hat{\eta} \in \{0, 1\}$  の  $\hat{\eta}$ ))

$$R_{err}(f) = E[\mathbb{1}[\gamma \neq \text{sign}(f(x))]] : \text{予測判別誤差}$$

$$R_{err}^*(f) = \inf_{f: \text{可測}} R_{err}(f) : 0 - 1 \text{ 損失の} \hat{\eta} \text{ もとでのベイズ誤差}.$$

… ベイズ規則  $h_0$  が存在するとき.

$$R_{err}^* = R_{err}(h_0) \text{ が成立.}$$

予測損失

予測  $\phi$ -損失 (2値)

$$R_\phi(f) = E[\phi(\gamma f(x))]$$

予測  $\psi$ -損失 (多値)

$$R_\psi(f) = E[\psi(f; x, \gamma)] = E[\psi(f(x), \gamma)]$$

経験マシン判別誤差

$$\text{2値: } \hat{R}_{err, \rho}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\gamma_i f_i(x_i) < \rho]$$

$$\text{多値: } \hat{R}_{mrg, \rho}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\text{mrg}(f; x_i, \gamma_i) < \rho]$$

予測重み - マシン判別誤差

$$\text{2値: } \hat{R}_{\bar{\omega}, \rho}(f) = E[\bar{\omega}_\rho(\gamma f(x))]$$

$$\text{多値: } \hat{R}_{\bar{\omega}, \rho}(f) = E[\bar{\omega}_\rho(\text{mrg}(f; x, \gamma))]$$

# 1. 統計的学習理論の枠組

⑥ 統計的学習理論を展開するための(必要となる)用語、概念(2つ以上)。

## 1.1. 内込設定

石倉論議、統計学

- 過去の経験と将来の出来事との間に「継続つながり」を表現

→ 過去と将来の間に、関連や何を目標にするかによって、未来の内込設定を考える。

… 石倉論議の内込問題と定式化。

## 「学習する」

(定式化された問題を解いて、観測された情報から仮想立て、情報を抽出し、利用するにて、

推定、予測とほぼ同義

AIがリスニングの観点、(二重点)が置かれる。

## ④ 用語

### データ (data)

- 観測によつて得られる情報。
- 一貫性の仮説的観察。

⑥ 過去の観測によつて得られた元になる情報(元データ)。  
将来得られる情報(次データ)と用語を使う。

### 学習データ (training data)

・(広義) 学習(2通りある)データ

・(狭義) 仮説(10データ)を決定するための直感用いるもの

## ・検証データ (Validation data)

- ・実験結果、性能を検証するためのデータ。
- ・主に学習アルゴリズム（= 簡単な正則化）の適切性を確かめるために用いられる。

## ・観測データ (observed data)

- ・観測されたデータ。
- ・（通常）実験データ + 検証データ。
- ・検証データを用いて模型の適合度を評価する。

## ・テストデータ (test data)

- ・学習アルゴリズムの予測精度を評価するためのデータ。
- ・（実験データ解説）

学習データ、検証データ：すでに観測され、利用可能。  
テストデータ：将来に観測されるデータ。

～）テストデータに対して高い予測精度を達成する方が  
モデルにおける大半は目標。

- ・（学習実験）

データと実験データ、検証用データ、テストデータは分離する。  
学習 検証 テスト

## ・入力データ (input data), 出力データ (output data), ラベル (label)

- ・データが入出力の組で表される。入力部分と入力データ、  
出力部分と出力データとなる。  
 $x$ : 入力,  $y$ : 出力とする。入出力データは  $(x, y)$  となる。  
（ $\oplus$ ）

$X$ : 入力空間  $y$   
(input space)

- ・出力データが有限集合の値をとること、つまり値をラベルとする。

$$|y|=2 \Rightarrow 2 \text{ラベル}$$

$$|y| \geq 3 \Rightarrow 3 \text{ラベル}$$

## ・仮説 (hypothesis)

- ・入力空間から出力集合へ, 肉叔.  $\sim h: X \rightarrow Y$
- ・仮説, 集合を仮説集合 (hypothesis set) と呼ぶ.

① 実験アルゴリズムは, 学習データ / 観測データを仮説集合から肉叔を見出すことをします.

## ・判別器 (classifier), 判別肉叔 (discriminant function)

有限集合 ( $n$ 個) と  
仮説.

判別器を記述する方法の中  
実数値 / ベクトル値 肉叔.

e.g. 出力が 2 価格ベリ

① 判別肉叔を用いたとき.

判別肉叔:  $f: X \rightarrow \mathbb{R}$

仮説集合をモデル化する  
が簡単になります.

判別器:  $h: X \rightarrow \{-1, 1\}$

$$h(x) = \text{sign}(f(x))$$

## ・損失肉叔 (loss function)

- ・出力値と予測結果の間, 誤差を測る肉叔.
- ・損失肉叔の値が大きいほど, 誤差や損失が大きい.

## 機械学習 分類, 主観因子

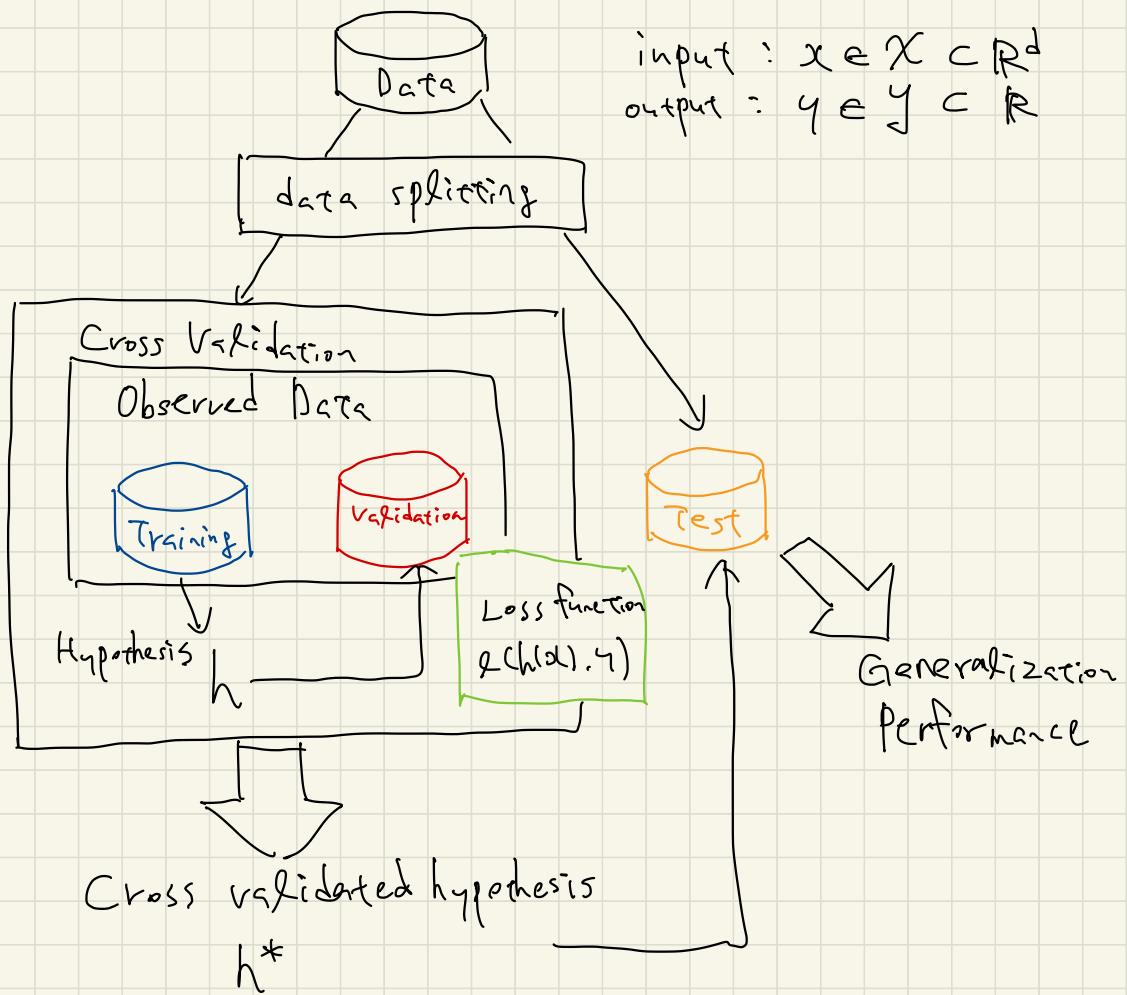
・観測データから 仮説を学習するアルゴリズム を設計する.

## 統計的学習理論 (statistical learning theory)

- ・統計的学習理論は、得られた仮説の予測精度を評価し、性能を向上させた上で、精緻な予測理論の構成.

以下、統計的学習理論で扱う代表的な問題設定を紹介する.

向量設置，以之  
 (參考：松井，「系統的統計理論」輪読セミナー)



# 1.1.1. 判別問題

## \* 判別問題 (classification problem)

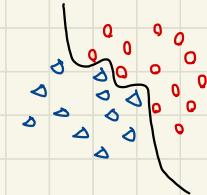
出力が有限集合  $\{1, 2, \dots, K\}$  のラベルに値を持つとき、入力データから対応するラベルを予測する問題。

$$|\gamma| = 2$$

## 2値判別問題 (binary classification problem)

$$|\gamma| \geq 3$$

## 多値判別問題 (multiclass classification problem)



SVM

$$\gamma - \gamma^* = \gamma$$

判別問題の例：迷惑メール分類。

入力データ： $X$ -メールテキストデータ

出力ラベル：迷惑メール ("spam") , 非迷惑メール ("non-spam")

$\hat{y}$ ：予測ラベル,  $y$ ：真ラベル

損失関数： $0-1$  損失

$$l_{err}(\hat{y}, y) = \begin{cases} 1 & \hat{y} \neq y \\ 0 & \hat{y} = y \end{cases} \quad (1.1)$$

損失が真ラベル  $y$  に依存する場合もあす。

E.g. クレジットカード会社が顧客の購買履歴から将来の支払いが可能かどうかを判別する場合。

→ 誤りによく被験子損失が用ひる。

$\Rightarrow 0-1$  損失を拡張。

$$l(\hat{y}, y) = \begin{cases} 0 & \hat{y} = y \\ 1 & \hat{y} \neq y \end{cases}$$

… 通常では  $l$  の値を定めることで、目的に見合った損失を小さくするよう工夫する。

### 1.1.2. 回帰問題.

#### 回帰問題 (regression problem)

出力が実数値をとるとき、入力データから出力を予測する問題.

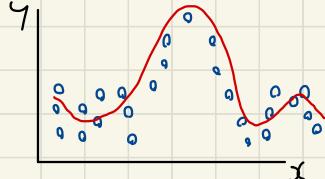
e.g. 株価や電力需要の予測.

出力と完全に一致する値を予測することは通常でない。

~) 2乗損失 (squared loss) がよく用いられる。

$\hat{y}$ : 予測値,  $y$ : 真の出力,

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



④ 適当な統計モデルを設定し、2乗損失のもとで最適な係数を学習する。

→ 2乗結果を用いて将来の入力  $x$  に対する出力  $y$  の値を予測する。

### 1.1.3. ランキング問題.

2つの入力  $x, x' \in \mathcal{X}^2$  に対し  $(x, x') \in \mathcal{X}^2$  とする,

$\begin{cases} y = +1 : x_1 \text{ 方が } x' \text{ より大きい} \\ y = -1 : x_1 \text{ が } x' \text{ より小さい} \end{cases}$

が観測されたとする。(入力  $x, x', y$  が得られる)

#### ランクニグ問題 (ranking problem)

$\begin{cases} h(x) > h(x') : x_1 \text{ 方が } x' \text{ より大きい} \\ h(x) \leq h(x') : x_1 \text{ が } x' \text{ より小さい} \end{cases}$

よろづや係数  $h: \mathcal{X} \rightarrow \mathbb{R}$  の定義.

入力  $x, x' \in \mathcal{X}^2$ ,  $h_1 = h(x)$ ,  $h_2 = h(x')$  とする.

$\hat{h} = (h_1, h_2) \in \mathbb{R}^2$  と出力ラベル  $y \in \{-1, +1\}$  とする

損失  $\ell(\hat{h}, y)$  は、0-1損失

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

因数誤植と呼ばれるが遠

なるが用いられる。

入力  $(x, x')$  に対するラベル  $y$  と  $h(x) - h(x')$  の符号で予測を決める  
と解釈すれば、損失  $\ell(h, y)$  もってラニキニヤ"肉眼を判断"  
肉眼と比較する。

一般に、複数の入力データ  $(x_1, x_2, \dots)$  が与えられたとき、  
出力  $y$  と入力データ  $(x_1, x_2, \dots)$  に対する判断肉眼を統合  
有可能"らうを考えるのもある。

例：ウェーブ検索、コンピュータによく似た構造、対戦

## 1.2 予測損失と経験損失

統計的学習理論では、この2種類の損失を用いる。

→ これらは関係を調べるために、学習アルゴリズムの予測精度  
などを定量的に評価できる。

① 確率的価値  $E$  による出入力データと確率変数で表す。

$$E_{(x, y) \sim D} [\cdot]$$

- データ  $(x, y)$  が従う確率分布  $D$  のもとで、期待値
- 簡単なたとえ、誤解がないように、 $E[\cdot]$  や  $E[\cdot]$  とする。

### 予測損失 (predictive loss)

損失函数として  $\ell(h, y)$  を用いる。

Def (予測損失)  
 $R(h)$  : 予測損失を

$$R(h) := E_{(x, y) \sim D} [\ell(h(x), y)]$$

で定義。

テストデータ  $(x, y)$  の分布の下で  
予測値  $h(x)$  の損失の期待値

④ (1.1) の  $l_{\text{err}}$  が定まる予測損失を  $R_{\text{err}}(h)$  と表し、  
予測判別誤差 (predictive classification error) と呼ぶ。

⑤ 実質上おける通常の内訳設定では、 $T = 1$ 、分布は未知。  
→ 予測損失を計算できない。

目標

観測  $T = 1$  のみから予測損失をできるだけ小さくする  
仮説を求める。

### 経験損失 (empirical loss)

$T = 1$  ( $X_1, Y_1$ , ...,  $(X_n, Y_n)$ ) が観測されたとき、  
出入力関係を仮説  $h$  で説明する参考。

-Def (経験損失)

$$\sum_{i=1}^n l(h(X_i), Y_i) : \text{観測 } T = 1$$

$\hat{R}(h)$  : 観測  $T = 1$  に対する仮説  $h$  の経験損失

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

予測値  $h(X_i)$  と観測値  $Y_i$  の  
間の損失を平均値。

④ (1.1) の 0-1 損失が定まる経験損失を  $\hat{R}_{\text{err}}(h)$  と表し、  
経験判別誤差 (empirical classification error) と呼ぶ。

出入力空間  $X \times Y$  上の分布に従う分布に従う期待値を用いて、  
経験損失を求める。もとより。

経験分布  $\hat{D}$

$T = 1$  放が  $n$  のとき、確率  $1/n$  で、 $(X_i, Y_i)$  の値を  $x$   
確率変数の従う分布。

経験損失は。

$$\hat{R}(h) = \mathbb{E}_{(x, y) \sim \hat{D}} [l(h(x), y)]$$

② 予測損失と経験損失との違いは、期待値を計算するときの確率分布の違い。

各  $i \sim (X_i, Y_i)$  が同一分布  $D$  に従うとき、経験損失の期待値は、予測損失は一致する。

但し、観測  $i \sim$  同時分布を  $D^n$  とする。  $R(h)$

$$\text{E}_{D^n} [\hat{R}(h)] = \text{E}_{D^n} \left[ \frac{1}{n} \sum_{i=1}^n \ell(h(x_i), Y_i) \right] = \frac{1}{n} \sum_{i=1}^n \text{E}_D [\ell(h(x_i), Y_i)]$$

$$\frac{1}{n} \sum_{i=1}^n \ell(h(x_i), Y_i) \quad (X_1, Y_1), \dots, (X_n, Y_n) \sim D^n \quad = R(h)$$

$$R(h) = \text{E}_D [\ell(h(x_i), Y_i)]$$

$\hat{R}(h)$  は  $R(h)$  の不偏推定量。

③  $i \sim$  独立性は仮定していい。

通常は独立性を仮定し、経験損失と予測損失が成り立つ相対について解析を行なう。独立性が成り立つことを示せ。

観測  $i \sim$  が独立で同一分布  $D$  に従うとき大抵、法则から  $\hat{R}(h)$  が  $R(h)$  に確率収束する。i.e.

分布  $D^n$  のもとで、 $\forall \varepsilon > 0$  は成り立つ。

(33)

$$\lim_{n \rightarrow \infty} \Pr_{D^n} (|\hat{R}(h) - R(h)| > \varepsilon) = 0,$$

が成立する。 $(\Pr_{D^n} : D^n$  のもとでの確率)

④ ここで問題は予測損失を最小化する仮説を求めるといふ問題として定式化せよ。

正確には予測損失の公因は未知だから、近似値として経験損失が分かる。

~経験損失を最小化するとして、適切な仮説を學習で王子を考える。

⇒ 学習した仮説の精度を評価するため、経験損失と予測損失の違いを重視する。

# 7.3 ベイズ規則とベイズ誤差

学習, 目標

予測誤差を最小にする仮説を求める.

→ 予測誤差を最小にする仮説が学習, 理想.

Def 7.7. (ベイズ規則, ベイズ誤差)

Given: 損失関数  $\ell$

任意の可測函数  $h: X \rightarrow Y$  がもつての予測損失の下限

$$\inf_{h: \text{可測}} R(h) = E_{(x, Y) \sim P} [\ell(h(x), Y)] \quad (\text{データ確率分布から定まる})$$

を、損失関数  $\ell$  もつての ベイズ誤差 (Bayes error) という.

下限を達成する仮説が存在する. これを 仮説を  
ベイズ規則 (Bayes rule) という.

ベイズ規則が  $h_0: X \rightarrow Y$  のとき,

$$R(h_0) = \inf_{h: \text{可測}} R(h)$$

が成立する.

ベイズ誤差

損失関数定義

→ データ確率分布から定まる.

以下、予測誤差を最小化する仮説を具体的に示す.

$\ell(\cdot, \cdot)$ : 損失関数

$P$ : データ確率分布

とすると.

$$R(h) = E_{(x, Y) \sim P} [\ell(h(x), Y)]$$

$$= E_x [E_Y [\ell(h(x), Y) | x]] \quad (\because \text{条件的期待値})$$

$$\star E[h(x_1) | x_2 = x_2] := \int h(x_1) f(x_1 | x_2) dx_1$$

$\cdots x_2 = x_2$  とするときまたもって  $h(x_1)$  の条件的期待値

$$E[h(x_1) | x_2]$$

$\cdots x_2$  の確率ベクトル  $x_2$  の確率ベクトル

$$\star E[E[h(x_1) | x_2]]$$

$$= E[h(x_1)]$$

各入力  $X = x$  の下の条件付期待値

$$\mathbb{E}_Y [l(h(x), Y) | x] = \int_Y l(h(x), y) dP(y|x)$$

を最小にするには仮説  $h$  を選べば、予測誤差が最小となる。

~ 内單数  $t_2$  の  $x$  を省略して、

$$\mathbb{E}_Y [l(h, Y)] = \int l(h, y) dP(y)$$

を最小にする  $h \in \mathcal{Y}$  を  $t_2$  の内定を考へる。

### 例 1.1 判別函数

損失関数 : 0-1 損失  $\ell_{\text{err}}(y, \hat{y}) (= \mathbb{1}[y \neq \hat{y}])$

$$\begin{aligned} \sim \mathbb{E}_Y [l(h, Y)] &= \sum_{y \in \mathcal{Y}} l(h, y) \Pr(Y=y) \\ &= \sum_{y \neq h} \Pr(Y=y) = 1 - \Pr(Y=h) \end{aligned}$$

よって、 $h \in \mathcal{Y}$ .

$$h = \operatorname{argmax}_{y \in \mathcal{Y}} \Pr(Y=y)$$

とすると、予測誤差が最小値を得られる。

条件付期待値 (2) 式、(3) 式を用いて、ベイズ規則  $h_0$  (2)

$$h_0(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \Pr(Y=y|x)$$

入力  $x$  が与えられたとき、  
最も出現する確率が大きいラベル  
を予測ラベルとする仮説

0-1 損失のもとで、ベイズ誤差  $R_{\text{err}}^*$  (2).

$$R_{\text{err}}^* = R_{\text{err}}(h_0) = 1 - \mathbb{E}_x \left[ \max_{y \in \mathcal{Y}} \Pr(Y=y|x) \right]$$

## 13) 1.2. 回帰分析

損失函数: 2乗損失  $l(\hat{Y}, Y) = (\hat{Y} - Y)^2$   
 $V[Y]$ :  $Y$  の分散.

$$\begin{aligned}\Rightarrow E_Y[l(h, Y)] &= E_Y[(h - Y)^2] \\ &= E_Y[h^2 - 2hY + Y^2] \\ &= h^2 - 2hE[Y] + E[Y^2] \\ &= (h - E[Y])^2 - (E[Y])^2 + E[Y^2] \\ &= (h - E[Y])^2 + V[Y]\end{aligned}$$

よし、 $h = E[Y]$  とすれば 最小値が得られる.

条件付期待値は  $x$  を考慮して、ベイズ規則は 入力  $x$  における  
条件付期待値

$$h_o(x) = E[Y|x]$$

であることを示す.

$V[Y|x]$ : 入力  $x$  における出力  $Y$  の条件付分散.

ベイズ誤差  $R^*$  は.

$$\begin{aligned}R^* &= R(h_o) = E_x[V[Y|x]] \\ &= E_x \left[ \int (Y - E[Y|x])^2 dP(Y|x) \right]\end{aligned}$$

である.

② 条件付分散  $V[Y|x]$  が 入力  $x$  によらず一定の値  $\sigma^2$  である場合、ベイズ誤差は  $\sigma^2$  である.

### 3) 1.3. ラニキニゲ"問題

ラニキニゲ"問題, 判別問題と(7.9)定式化

入力  $(x, x') \in X^2$  をラベル  $y \in \{-1, +1\}$  に対応させる  
2種類の式, 仮説, 定義は次のとおり.

~) ラニキニゲ" (2用い子) すなはち  $h: X \rightarrow \mathbb{R}$  (2種類の判別器が  
sign  $(h(x) - h(x'))$  と見せる場合に仮説が当たる).

⇒ 2種類の式を用いてベイズ規則から ラニキニゲ"問題における  
ベイズ規則を構成します.

以下,  $\tilde{x} = x_+ - x_-$  分布は仮定をおいて, ラニキニゲ"問題における  
ベイズ規則の特徴を示すを行なう.

入力  $x = (x_+, x_-) \in X^2$  とする.

•  $x_+$  の方が  $x_-$  より大きい.

• ラニキニゲ"を表す出力ラベルは常に  $y = +1$ .

∴ ラニキニゲ"の入力  $\tilde{x} = (x_+, x_-, -1) \in (x'_+, x_-, +1) \subset$   
変換可逆で得られる.

仮定

$x_+, x_- \in X$  は確実に独立して分布  $D_+, D_-$  とする.

⇒ 定義  $\tilde{x} = x_+ - x_-$  が, 入力  $x$  ラニキニゲ"問題の  $h: X \rightarrow \mathbb{R}$  を定義する.

真陽性率 (true positive rate)

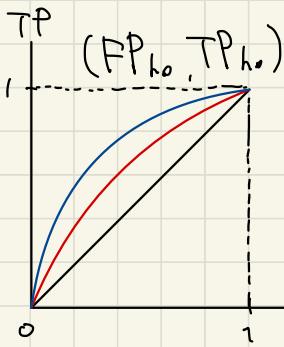
$$TP_h(a) = \mathbb{E}_{x_+ \sim D_+} [\mathbb{1}[h(x_+) > a]]$$

偽陽性率 (false positive rate)

$$FP_h(a) = \mathbb{E}_{x_- \sim D_-} [\mathbb{1}[h(x_-) > a]]$$

- $\forall a \in \mathbb{R}$  ( $\in \mathbb{R}$ )  $(FP_h(a), TP_h(a)) \in [0, 1]^2$ .
- $11^{\circ}$   $x \rightarrow a$  が  $\infty$  から  $-\infty$  まで動くとき  $(FP_h(a), TP_h(a))$  は  $(0, 0)$  から  $(1, 1)$  まで動く。

受信者操作特性曲線 (receiver operating characteristic curve)  
ROC 曲線.



AUC (曲线下面積: area under the curve)

... ROC 曲線の下側の面積

(ROC 曲線で  $TP = 0, FP = 1$  )  
(曲線を下へ延長する)

... AUC(h) と表す

~ 同じ偽陽性率に対する真正陽性率が大きい方が良い。

~ AUC が大きい方が良い。

通常 AUC が 0.5 より大きい状況を考える。

- 仮説  $h$  が人力 ( $\in \mathbb{R}$ ) から出力 ( $\in \mathbb{R}$ ) の標準正規分布 ( $\mathcal{N}(0, 1)$ ) に従う値を返すとき, ROC 曲線は  $45^{\circ}$  の斜線 ( $TP = FP$ ) ( $\approx 1$ ),  $AUC(h) = 0.5$  ( $\approx 72\%$ ).
- ランダムな予測より良い仮説を述べるは,  $AUC(h) > 0.5$  である。
- ランダムな予測より悪い仮説を述べるは,  $AUC(h) < 0.5$  である。

$$\begin{aligned} R(h) &= 1 - \mathbb{E}_{x_{\pm} \sim D_{\pm}} \left[ \mathbb{I}[h(x_+) - h(x_-) > 0] \right] \\ &= 1 - \mathbb{E}_{x_{\pm} \sim D_{\pm}} \left[ \mathbb{E}_{x_{\mp} \sim D_{\mp}} [\mathbb{I}(h(x_+) > h(x_-))] \right] \\ &= 1 - \mathbb{E}_{x_{\pm} \sim D_{\pm}} [TP_h(h(x_{\pm}))] \quad \leftarrow \\ &= 1 - AUC(h) \quad \leftarrow \end{aligned}$$

~  $x_+, x_-$  が独立のとき, ランダムな出力におけるベイズ規則  $h_0$  は,  $AUC$  を最大にする仮説 ( $\approx 72\%$  となる), ( $h_0 = \arg \max AUC(h)$ )

ベイズ誤差 ( $\approx 1 - AUC(h_0)$  となる。

7.4. 学習アルゴリズムの性能評価。

$$A : 2^{X \times Y} \rightarrow \mathcal{H}$$

$\hookrightarrow$  内双  $S \mapsto A(s) = h_{S, x}$

$$S = \left\{ (x_i, y_i) \right\}_{i=1}^n$$

中華人民國(大清)

... 閱覧(測)する集合  $\{x_i\}$  が既述集合への肉故  $S = \{(X_i, Y_i)\}_{i=1}^n$   
 $(H \subset S)$

→ ある学習アルゴリズムを用いて式と左. データ  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  から得られた仮説を  $h_s$  と表す.

…アルゴリズムの内部で乱数を用いること。同じデータを入れても  
異なる仮説が得られることがある。

## 学習アシスタントの性能を評価する方法

損失函数を定めると、学習モデルは仮説  $h_s$  の予測損失と  $R(h_s)$  が定まる。

確率的に得点を立てるまでの頻度データに基づいて、  
予測失敗率を算出する

評價尺度，(2)：期待予測損失

字體アルゴリズムが平均的  
性能を評価

觀測了 7-9 月，分布 D<sup>n</sup> 次肉仔苗的喪失，期待值。

$$\mathbb{E}_{s \sim D} [R(h_s)] \dots \text{「27」}$$

(定義式)

通常、内蔵設定では、予測損失  $R(h_s)$  を計算するとき、テストデーターの分布も  $D$  とする。

他<sup>9</sup> 言平價尺度：予測損失  $R(h_s)$ ，值の分布

$$R^* = \inf_h R(h) = \text{optimal value}$$

$\exists \eta \in \mathbb{R}, \delta \in (0, 1) \wedge \forall \varepsilon > 0 \exists N \in \mathbb{N}$

$$\Pr_{S \sim P_n} (R(hs) - R^* < \varepsilon) > 1 - \delta \quad (1.3)$$

が成り立つとする。

・・・石窟寺の  $|-\delta|$  は如何で、それがどうような値になるかを評価できます。ここで、学習アルゴリズムの性能を評価できます。

この本では、主に確率論的評価式 (1.3) を用いる。

期待予測損失と (1.3) による確率評価式の関係：

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

(∴ 実際の不等式)

① ベイズ誤差は必ず予測損失を達成する仮説を示すことがでます。

→ ここで予測アルゴリズムは、統計的一致性を持つ。

Def 1.2. (統計的一致性)

任意の分布  $D$  と任意の  $\varepsilon > 0$  に対して、

$$\lim_{n \rightarrow \infty} \Pr_{S \sim D^n} (R(h_s) \leq R^* + \varepsilon) = 1 \quad (1.4)$$

が成立すると、学習アルゴリズム  $A: S \mapsto h_s$  は 統計的一致性 (Statistical consistency) をもつという。

② 統計的一致性を持つ學習アルゴリズムを用いれば、データを生成する分布の(実用的事前知識)がなくても、データ数が十分多ければ、最適な仮説を示すことができるのです。

(統計的一致性)の下で、分布  $D$  について考え方範囲を制約する場合もある。

e.g. 2個類別を扱う SVM

入力空間  $\mathcal{X} \subset \mathbb{R}^d$  は  $\mathbb{R}^d$  の子集合を考える。

判別関数  $f$  は予測誤差  $R_{err}(h)$  がベイズ誤差  $R_{bay}$  (2種類) 未満であるような學習アルゴリズムが望むべき場合がある。

③ 近年用いられてきた主な學習アルゴリズムは、統計的一致性を持つこれが証明されています。

… 統計的一致性が保証されていない學習アルゴリズムでも、計算効率が非常に高い場合もある。

7.5. 有限な仮説集合を用いて学習する.

7.5.1. 予測判別誤差, 評価.

仮説集合が有限の場合には、学習された仮説の予測損失を評価可能.

① ここで  $\mathcal{H}$  が  $\mathcal{H} = \{h_1, \dots, h_T\}$ ,  $h_i : \mathcal{X} \rightarrow \{-1, +1\}$

## 問題設定

2種類の問題

・ 有限な仮説集合  $\mathcal{H} := \{h_1, \dots, h_T\}$ ,  $h_i : \mathcal{X} \rightarrow \{-1, +1\}$

・ 実質データ:  $S = \{(x_i, y_i)\}_{i=1}^n$ ,  $(x_i, y_i) \stackrel{i.i.d.}{\sim} P$

→ 程度判別誤差を最小にした仮説を出力する学習アルゴリズムを考える.

$$A : 2^{x \times y} \rightarrow \mathcal{H}$$

$$S \mapsto A(S) = h_s = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_{\text{err}}(h)$$

式の出力  $\sum_{i=1}^n \mathbb{1}(h(x_i), y_i)$

分布  $P$  で  $\mathbb{E}_{(x,y) \sim P} [0-1\text{損失}]$  のベイズ規則を  $h_0$  とする.

( $h_0 \in \mathcal{H}$  かつ  $R_{\text{err}}(h_0) \leq u$ )

学習アルゴリズムが性能を評価可能.

$R_{\text{err}}(h_s)$ : 予測判別誤差

$R_{\text{err}}(h_0)$ : ベイズ誤差

差を評価可能.

$\mathcal{H}$  内で予測判別誤差を最小にする仮説

$$h_{\mathcal{H}} := \underset{h \in \mathcal{H}}{\operatorname{argmin}} R_{\text{err}}(h) \quad \text{となる。以下が成立。}$$

$$\circ R_{\text{err}}(h_0) \leq R_{\text{err}}(h_{\mathcal{H}}) \leq R_{\text{err}}(h_s)$$

全  $h$  可能  
 $\inf$

$\mathcal{H}$  内で min

$$\circ R_{\text{err}}(h_s) \leq R_{\text{err}}(h_{\mathcal{H}})$$

$R_{err}(h_s) \sim R_{err}(h_0)$  の上界を導入して証明.

$$\begin{aligned}
& R_{err}(h_s) \sim R_{err}(h_0) \\
& = R_{err}(h_s) - \widehat{R}_{err}(h_s) + \widehat{R}_{err}(h_s) - R_{err}(h_{2n}) + R_{err}(h_{2n}) - R_{err}(h_0) \\
& \leq \underbrace{R_{err}(h_s) - \widehat{R}_{err}(h_s)}_{\stackrel{(1.5)}{\approx}} + \underbrace{\widehat{R}_{err}(h_{2n}) - R_{err}(h_{2n}) + R_{err}(h_n) - R_{err}(h_0)}_{\text{誤差}} \\
& \leq \max_{h \in \mathcal{H}} |\widehat{R}_{err}(h) - R_{err}(h)| + \max_{h \in \mathcal{H}} |\widehat{R}_{err}(h) - R_{err}(h)| \\
& \quad + R_{err}(h_{2n}) - R_{err}(h_0) \\
& = 2 \max_{h \in \mathcal{H}} |\widehat{R}_{err}(h) - R_{err}(h)| + R_{err}(h_{2n}) - R_{err}(h_0) \quad (1.5)
\end{aligned}$$

したがって、(1.5) が示された。以上で証明が終った。

- Lemma 1.3. ( $\widehat{R}_{err} = \frac{1}{n} \sum_i \mathbb{I}[h(x_i) \neq y_i]$  不等式 (Hoeffding's inequality))

$Z : [0, 1]$  - valued r.v.

$Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} P_Z$  ( $Z$  は同一分布)

$\forall \varepsilon > 0$  は

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

が成り立つ。

(証明は金森 2015)

Lemma 1.3,  $Z \in \mathcal{G}$ ,  $Z = \mathbb{I}[h(x) \neq y]$  とする。

$$\begin{aligned}
& \Pr \left( 2 \max_{h \in \mathcal{H}} \left| \widehat{R}_{err}(h) - R_{err}(h) \right| \geq \varepsilon \right) \\
& \leq \sum_{h \in \mathcal{H}} \Pr \left( \left| \widehat{R}_{err}(h) - R_{err}(h) \right| \geq \frac{\varepsilon}{2} \right) \\
& \leq \sum_{h \in \mathcal{H}} 2e^{-2n(\varepsilon/2)^2} \quad (\because \text{Lemma 1.3}) \\
& = 2|\mathcal{H}| e^{-n\varepsilon^2/2}.
\end{aligned}$$

$$\text{ここで}, \delta = 2|H|e^{-n\varepsilon^2/2} \text{ とする}. \text{ 学習誤差} - \varepsilon \text{ の分布},$$

$$\Leftrightarrow \frac{2|H|}{\delta} = e^{n\varepsilon^2/2} \Leftrightarrow \varepsilon = \sqrt{\frac{2}{n} \log \frac{2|H|}{\delta}}$$

$$\Pr \left( 2 \max_{h \in H} \left| \hat{R}_{err}(h) - R_{err}(h) \right| \leq \varepsilon \right) \geq 1 - \delta$$

$$\Leftrightarrow \Pr \left( 2 \max_{h \in H} \left| \hat{R}_{err}(h) - R_{err}(h) \right| \leq \sqrt{\frac{2}{n} \log \frac{2|H|}{\delta}} \right) \geq 1 - \delta$$

が成立.

ここで (1.5) 式で全ての学習誤差の分布が成立する.

$$R_{err}(h_s) - R_{err}(h_0) \leq 2 \max_{h \in H} \left( \hat{R}_{err}(h) - R_{err}(h) \right) + R_{err}(h_s) - R_{err}(h_0)$$

より.

$$\Pr \left( R_{err}(h_s) - R_{err}(h_0) \leq R_{err}(h_s) - R_{err}(h_0) + \sqrt{\frac{2}{n} \log \frac{2|H|}{\delta}} \right) \geq 1 - \delta$$

(1.6)

が成立する.

仮説集合  $H$  が「ベイズ規則」 $h_0$  を含む ( $h_s = h_0$ ) の仮定は成り立つ.

$$R_{err}(h_s) - R_{err}(h_0) = 0$$

となる. つまり  $n$  分の大部分が  $h_s$  である.  $R_{err}(h_s) \rightarrow R_{err}(h_0)$  の確率オーダーは  $O_p(1)$ .

$$R_{err}(h_s) = R_{err}(h_0) + O_p \left( \sqrt{\frac{\log |H|}{n}} \right)$$

④  $|H| = \infty$  の場合は上記の議論を用いて予測損失の上界を

... ここで、易しい仮説集合 ( $H$ ) によ詳しく述べる.

有限の仮説集合  $H$  の場合は帰着させることができる.

## 1.5.2. 近似誤差と推定誤差

実際、 $\mathcal{H}$  内で  $h_s$  は、仮説集合がベイズ規則  $h_o$  を含むと假定できぬ。一般に  $h_s \neq h_o$

$$Rerr(h_s) - Rerr(h_o) > 0$$

となる。

~ 近似誤差 (approximation error) と推定誤差 (estimation error) を区別。

Def (近似誤差、推定誤差)

評価式 (1.6)

$$Rerr(h_s) - Rerr(h_o) \leq \sqrt{\frac{2}{n} \log \frac{2|\mathcal{H}|}{\delta}} + Rerr(h_{\mathcal{H}}) - Rerr(h_o)$$

(おいた近似誤差と推定誤差は上下で定義)

$$bias_{\mathcal{H}} := Rerr(h_{\mathcal{H}}) - Rerr(h_o)$$

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

- bias : モデルを外れていたときに生じる誤差  $\rightarrow$  一般に  $h_o \notin \mathcal{H}$  时
- var : 実験データ(サンプルサイド)に由来するばらつき,  $bias_{\mathcal{H}} \geq 0$

仮説集合  $\mathcal{H}$  を適切に設定すれば  $h_s$  の予測判別誤差を小さくすることができる。すなはち  $h_s$  が成り立つ。

複数の仮説集合  $\mathcal{H}_1, \dots, \mathcal{H}_m$  が次の包含関係を満たす。

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots \subset \mathcal{H}_m$$

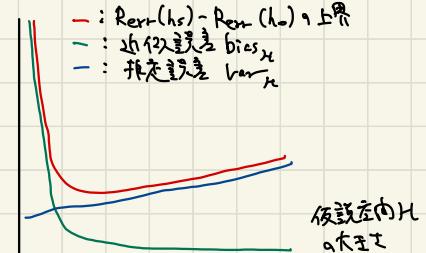
$(f_2 f_2 \dots (f_m | \mathcal{H}_m | < \infty \approx 0))$

$\therefore$  一定の定義から近似誤差と推定誤差は内

$$bias_{\mathcal{H}_1} \geq bias_{\mathcal{H}_2} \geq \dots \geq bias_{\mathcal{H}_m}$$

$$var_{\mathcal{H}_1} \leq var_{\mathcal{H}_2} \leq \dots \leq var_{\mathcal{H}_m}$$

が成立する。



- ・仮説空間が広いほどベイズ規則に近い仮説が手に入りやすくなる。
- ・サニ7°ルサイズを止めても広くならず、バラつきが増大。
- ・サニ7°ルサイズが十分大きくなる  
⇒ 大きなHでもvar(子bias)は対して大きくなる。
- ・サニ7°ルサイズが小さくなる  
⇒ var(子H)が大きくなり、影響を受ける。

予測精度を高める仮説を得るには、bias<sub>H</sub>とvar<sub>H</sub>の和を小さくする仮説集合、すなわち、

$$\hat{m} = \arg \min_m \text{bias}_{H_m} + \text{var}_{H_m} \quad (1.7)$$

しかし、H<sub>m</sub>を用いるのは危険。

but. bias<sub>H<sub>m</sub></sub>が正規分布に従うことはない。

∴ 基準で仮説集合を選ぶ方法は実用的ではない。

⇒ 正則化法

### 1.5.3. 正則化

- ② 大きな仮説集合を用いると推定誤差が大きくなるため、学習結果得られる仮説、予測判別誤差が大きくなる傾向がある。
- ~ 小さな仮説集合で十分に対応できず正規化法
- 大きな仮説集合を使うのがよろしくないが改善。

#### idea

大きな仮説集合から仮説を適切に選ぶことを対して規則化を課す。

複数の仮説集合  $H_1 \subset \dots \subset H_m \in \mathbb{R}$  の学習を行える。

重:  $H_m \rightarrow \mathbb{R}_+$ : 仮説hに対する  $\ell^0 + \|\cdot\|_1$  の和

•  $m_1 < m_2 \Rightarrow$

$$h \in H_{m_1}, h' \in H_{m_2} \setminus H_{m_1} \Rightarrow \text{重}(h) \leq \text{重}(h')$$

∴ 大きな仮説集合による大きな  $\ell^0 + \|\cdot\|_1$  を課す。

13]

$$H_0 = \emptyset, \quad 0 < \omega_1 < \dots < \omega_m (= \text{対応} \cap$$

$$\Psi(h) = \sum_{m=1}^M \omega_m \cdot \mathbb{I}[h \in H_m \setminus H_{m-1}]$$

仮説  $h$  は対応  $\cap$  に  $L_i$  が定義された場合、 $|L_i \cap \Psi(h)|$  と  $\hat{\Psi}(h)$  の値が実際、学習アルゴリズムではよく用いられる。

仮説に対する評価ルールを考慮しながら、経験判別誤差を最小化する  
小数点以下2桁まで行う。

探索範囲:  $H_m$

$$\text{基準: } \min_{h \in H_m} \hat{R}_{\text{err}}(h) + \lambda \cdot \Psi(h)$$

正則化項

~ 経験判別誤差が同じ仮説が複数あるとき、最も小数点以下2桁まで仮説集合に含まれる仮説が選ばれる。

① データが十分大きくなると、大半の仮説集合を用いても  
 判別誤差があまり大きくならない。

→ 決め方

- データ数は十分でない場合、 $\lambda$  を  $\lambda - \frac{1}{n} - \frac{1}{n^2} \rightarrow 0$  as  $n \rightarrow \infty$  とする。
- ログバーナード

→ (1) 複数の理論的評価 (や可)