仮説集合の複雑度 2.1節

参考文献

- 金森, 統計的学習理論, 講談社, 2015.

Table of contents

VC次元

- VC dimension
  - Vapnik and Chervonenkis
  - 2値判別問題のための仮説集合に対して定義される複雑度
    - 多値判別や回帰問題の場合に拡張することも可能.
  - 集合族の組合せ的な性質を捉えるための量
    - 組合せ論などにも応用

- $\mathcal{H}$: 2値判別のための仮説集合
  - $h \in \mathcal{H}$: 仮説 ($h : \mathcal{X} \to \mathcal{Y},\ |\mathcal{Y}| = 2$)

$\{x_1, \dots, x_n\} \subset \mathcal{X}$: 入力の集合 に対して，$\mathcal{Y}^n$の部分集合

$$\{(h(x_1), \dots, h(x_n)) \in \mathcal{Y}^n \mid h \in \mathcal{H}\}$$

の要素数を

$$\Pi_{\mathcal{H}}(x_1, \dots, x_n) = |\{(h(x_1), \dots, h(x_n)) \in \mathcal{Y}^n \mid h \in \mathcal{H}\}|$$

とおく．定義から，

$$\Pi_{\mathcal{H}}(x_1, \dots, x_n) \leq 2^n$$

が成り立つ．

等式$\Pi_{\mathcal{H}}(x_1, \ldots, x_n) = 2^n$ が成り立つとする．
$\rightsquigarrow$ 各x_iにラベル$y_i \in \mathcal{Y}$を割り付けて得られる任意の2値データ
$\{(x_1, y_1), \ldots, (x_n, y_n)\}$に対して，適切に$h \in \mathcal{H}$を選べば

$$h(x_i) = y_i,\ i = 1, \ldots, n$$

とできる．

入力の数nが増えていけば，ラベル付けのパターンが豊富になり，上記の等式が成立しにくくなると考えられる．
その境界となるデータ数nを$\mathcal{H}$のVC次元という．つまり，

$$\mathrm{VCdim}(\mathcal{H}) = \max \left\{ n \in \mathbb{N} \;\middle|\; \max_{x_1, \ldots, x_n \in \mathcal{X}} \Pi_{\mathcal{H}}(x_1, \ldots, x_n) = 2^n \right\}$$

と定義される．

$$\mathrm{VCdim}(\mathcal{H}) = \infty \overset{\text{def}}{\iff} \forall n \in \mathbb{N},\ \exists x_1, \ldots, x_n \in \mathcal{X},\ \Pi_{\mathcal{H}}(x_1, \ldots, x_n) = 2^n.$$

- VC次元の解釈
  - どのようなラベル付けにも対応可能な仮説が存在するようなデータ数の上限
- データ数nが$n \leq \mathrm{VCdim}(\mathcal{H})$のときは学習がうまくいくと考えることもできる．
  - ノイズによってラベルが反転してしまう状況を考えると，必ずしも学習がうまくいくわけではない
  - 仮説集合がどのようなラベル付けにも対応できる = ノイズとして無視すべきデータも学習してしまう
  - 仮説集合の複雑度は，データの複雑さに合わせて適切に設定することが重要

Lemma 2.1 (Sauer's lemma)

2値ラベルに値をとる仮説集合$\mathcal{H}$のVC次元がdのとき，$n \geq d$に対して

$$\max_{x_1, \ldots, x_n \in \mathcal{X}} \Pi_{\mathcal{H}}(x_1, \ldots, x_n) \leq \left(\frac{en}{d}\right)^d \quad (e: \text{ネイピア数})$$

Th 2.2

2値ラベルに値をとる仮説集合$\mathcal{H} \subset \{h : \mathcal{X} \to \{-1, +1\}\}$のVC次元を$d < \infty$とする．学習データ$(X_1, Y_1), \ldots, (X_n, Y_n)$はi.i.d.であるとする．損失として0-1損失を用いると，$n \geq d$のとき，学習データの分布の下で確率$1 - \delta$以上で

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

が成り立つ．

Lemma 2.1の証明は略．Th 2.2の証明は2.3節で示す．

Th 2.2を用いて，推定された仮説の予測判別誤差を評価する．
$\rightsquigarrow$ 有限集合ではない仮説集合が扱える．

- $S=\{(X_1,Y_1),\dots,(X_n,Y_n)\}$: 学習データ
- h_S: 経験判別誤差$\hat{R}_{\mathrm{err}}(h)$の最小化によって得られる仮説
- h_0: ベイズ規則
  - 予測損失$R_{\mathrm{err}}(h)$の下限を達成する仮説
  - $h_0\in\mathcal{H}$を仮定．

$\rightsquigarrow$ 以下の不等式が常に成立

$$\hat{R}_{\mathrm{err}}(h_S)\le\hat{R}_{\mathrm{err}}(h_0),\quad R_{\mathrm{err}}(h_0)\le R_{\mathrm{err}}(h_S)$$

$\rightsquigarrow$ 学習データSの分布の下で，$1-\delta$以上の確率で以下が成立．

$R_{\mathrm{err}}(h_S) \leq \hat{R}_{\mathrm{err}}(h_0) + R_{\mathrm{err}}(h_S) - \hat{R}_{\mathrm{err}}(h_S) \ \ (\because \hat{R}_{\mathrm{err}}(h_0) \geq \hat{R}_{\mathrm{err}}(h_S))$
$\leq R_{\mathrm{err}}(h_0) + |R_{\mathrm{err}}(h_0) + \hat{R}_{\mathrm{err}}(h_0)| + \sup_{h\in\mathcal{H}} |R_{\mathrm{err}}(h) - \hat{R}_{\mathrm{err}}(h)| \ \ (\because (\dagger), (\ddagger))$
$\leq R_{\mathrm{err}}(h_0) + 2\sup_{h\in\mathcal{H}} |R_{\mathrm{err}}(h) - \hat{R}_{\mathrm{err}}(h)| \ \ (\because (\ddagger))$
$\leq R_{\mathrm{err}}(h_0) + 4\sqrt{\frac{2d}{n}\log\frac{en}{d}} + 2\sqrt{\frac{\log(2/\delta)}{2n}} \ \ (\because \mathrm{Th}\ 2.2)$

これより，

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

となる．

$\rightsquigarrow$ 予測判別誤差はデータ数とVC次元の比n/dと関連している．

(†)
$$\hat{R}_{\mathrm{err}}(h_0) = R_{\mathrm{err}}(h_0) - R_{\mathrm{err}}(h_0) + \hat{R}_{\mathrm{err}}(h_0)$$
$$\leq R_{\mathrm{err}}(h_0) + |R_{\mathrm{err}}(h_0) - \hat{R}_{\mathrm{err}}(h_0)|$$

(‡)
$$R_{\mathrm{err}}(h_S) - \hat{R}_{\mathrm{err}}(h_S) \leq |R_{\mathrm{err}}(h_S) - \hat{R}_{\mathrm{err}}(h_0)|$$
$$\leq \sup_{h\in\mathcal{H}} |R_{\mathrm{err}}(h) - \hat{R}_{\mathrm{err}}(h)|$$

以下，VC次元の例を示す.

例 2.1 (有限仮説集合)

有限な仮説集合$\mathcal{H}$に対して

$$\mathrm{VCdim}(\mathcal{H}) \leq \log_2 |\mathcal{H}|$$

が成り立つ．

- d個の入力点に割り当てられるラベルのパターンは2^d通り
  - $|\mathcal{H}| < 2^d$ならば，すべてのラベルの割り当てに対応することができない

$\rightsquigarrow$ 推定された仮説h_Sの予測損失について

$$R_{\mathrm{err}}(h_0) \leq R_{\mathrm{err}}(h_S) \leq R_{\mathrm{err}}(h_0) + O_p\left(\sqrt{\frac{\log |\mathcal{H}|}{n}} \log \frac{n}{\log |\mathcal{H}|}\right)$$

が成立する．

例 2.2 (線形判別器のVC次元)

- $\mathcal{X} = \mathbb{R}^d$: 入力空間
- $\mathcal{H} = \{h(x) = \mathrm{sign}(w^{\mathrm{T}}x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$: 仮説集合 (線形判別器の集合)
- $\{x_1, \ldots, x_{d+1}\} \subset \mathbb{R}^d$: 列ベクトルの集合

これらのベクトルが$\mathbb{R}^d$で一般の位置にあるとき

$$A = \begin{pmatrix} x_1 & \cdots & x_{d+1} \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$$

は可逆な行列となる．

入力x_iにラベル$y_i \in \{+1, -1\}$を割り当てたデータに対して

$$\begin{pmatrix} w \\ b \end{pmatrix} = A^{-1}y,\ y = (y_1, \ldots, y_{d+1})^{\mathrm{T}}$$

とすれば，$h(x_i) = y_i,\ i = 1, \ldots, n$が成立．

実際，

$$y = A\begin{pmatrix} w \\ b \end{pmatrix} = (w^{\mathrm{T}}x_1 + b, \dots, w^{\mathrm{T}}x_{d+1} + b)^{\mathrm{T}}$$

より，$y_i = w^{\mathrm{T}}x_i + b = \mathrm{sign}(w^{\mathrm{T}}x_i + b)$となる．

$\rightsquigarrow \mathrm{VCdim}(\mathcal{H}) \geq d + 1$

以下，VC次元の上界を求めるのに役立つラドンの定理を紹介する．
集合Aの凸包$\mathrm{conv}(A)$を

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

とする．

Th 2.3 (ラドンの定理)

任意の点集合$S = \{x_1, \dots, x_{d+2}\} \subset \mathbb{R}^d$に対して，$S = S \cup S_2$，$S_1 \cap S_2 = \emptyset$かつ$\mathrm{conv}(S_1) \cap \mathrm{conv}(S_2) \neq \emptyset$となるような$S$の分割$S_1, S_2$が存在する．

- S_1, S_2
  - 任意の点集合$\{x_1, \dots, x_{d+2}\} \subset \mathbb{R}^d$に対して，ラドンの定理から定まる分割
- $y = +1$: S_1の点である
- $y = -1$: S_2の点である

このラベル付けに正答する線形判別器$h \in \mathcal{H}$が存在すると仮定する．

$\rightsquigarrow$ hは$\mathrm{conv}(S_1)$の点に$+1$を割り当て，$\mathrm{conv}(S_2)$の点に-1を割り当てる．

$\rightsquigarrow$ $S_1 \cap S_2$で矛盾 $\Rightarrow$ そのようなラベル付けは存在しない

$\rightsquigarrow$ $\mathrm{VCdim}(\mathcal{H}) = d + 1$

例 2.3

線形判別器の集合では，判別器を指定するパラメータの次元とVC次元が一致していた．しかし，常にパラメータの次元とVC次元が一致するわけではない．

- θ: 1次元パラメータ
- $\mathcal{H} = \{h(x) = \mathrm{sign}(\sin(2\pi\theta x)) \mid \theta \in \mathbb{R}\}$: 仮説集合

$\rightsquigarrow \mathrm{VCdim}(\mathcal{H}) = \infty$
$\rightsquigarrow$ 適切に入力点を選べば，その上の任意のラベル付けに対応できる仮説が$\mathcal{H}$の中に存在する．