実データで学ぶ人工知能講座

モデル表現力と性能保証

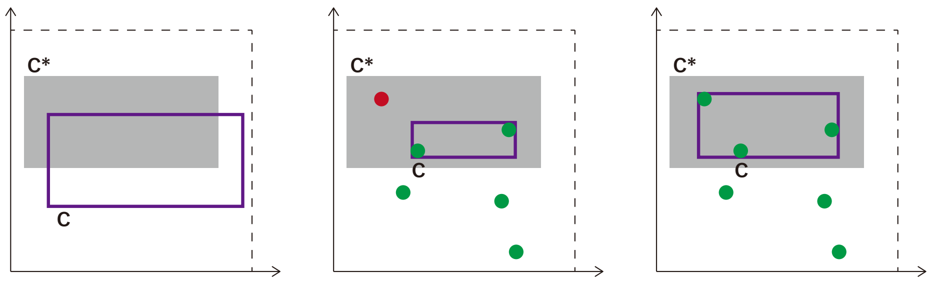
マシュー　ホーランド

Matthew J. Holland

matthew-h@ar.sanken.osaka-u.ac.jp

大阪大学 産業科学研究所 助教

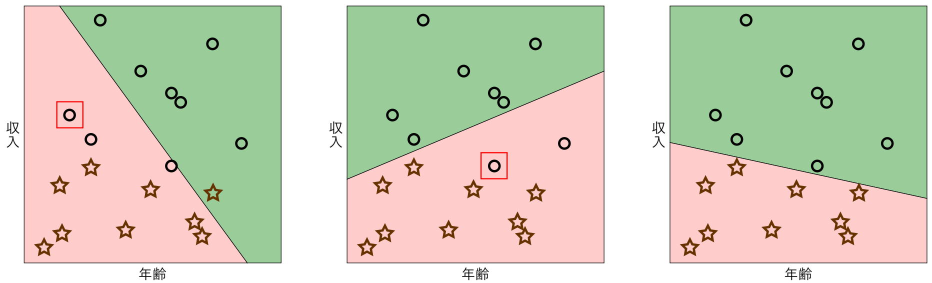
長方形と平面



正例包絡という ERM 学習則ならば, 強い性能保証は示せる.[1]

[1] 本資料の `rectangles.pdf` を参照.

長方形と平面



収束について触れたが，PLA の汎化能力はまだ判然としない.

長方形と平面

正例包絡の性能保証

$$\mathbf{P}\left\{R(\widehat{C}_{\text{fit}}) > \varepsilon\right\} \leq 4\left(1 - \frac{\varepsilon}{4}\right)^n \leq 4\exp\left(-\frac{n\varepsilon}{4}\right).$$

PLA の性能保証

データが線形分離可能な場合, 十分に反復回数を重ねておけば,

$$\widehat{R}(h_{\mathsf{PLA}}) = \frac{1}{n}\sum_{i=1}^{n} I\{h_{\mathsf{PLA}}(X_i) \neq Y_i\} = 0$$

が成り立つ.[2] これ以外はまだわからない.

[2] 本資料の `learning_intro.pdf` を参照.

長方形と平面

正例包絡と PLA の共通点と相違点

- どれも二値識別.
- どれもノイズやモデル誤差がないものとしている.
- どれも ERM 学習則である.
- ただし, 容易に出力が特徴づけられるのは, 前者のみ.

長方形と平面

正例包絡と PLA の共通点と相違点

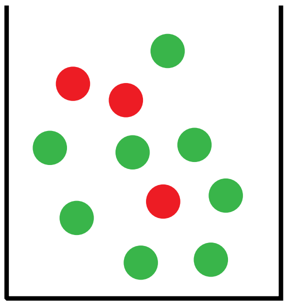
- どれも二値識別.
- どれもノイズやモデル誤差がないものとしている.
- どれも ERM 学習則である.
- ただし, 容易に出力が特徴づけられるのは, 前者のみ.

前者「どのような長方形になる？」
$\rightarrow$ データの関数として的確に説明できる.

後者「どのような平面になる？」
$\rightarrow$ データを分けるような平面. . . ？ 一意に定まらない. . .

より汎用的な解析アプローチ

まずは初歩的なところから始める.[3]

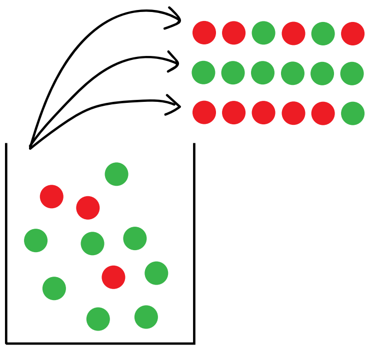


問題の概要
玉の色は赤か緑の 2 種類.
その割合を知りたい.

$$X = I\{\text{red marble}\}$$
$$\mathbf{E}\,X = \mathbf{P}\{\text{red marble}\}$$

[3] この玉と壺の例は Abu-Mostafa et al. (2004) から着想を得た.

より汎用的な解析アプローチ



標本を集める
取る $\rightarrow$ 色を記録する $\rightarrow$ 壺に戻す.
データ $X_1, \ldots, X_n$ を得る.

簡単な統計量

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i = \text{標本の赤の割合}$$

これはランダムなので，$\overline{X} = 0$ も $\overline{X} = 1$ もあり得るが, 本当の割合から大きく離れることは稀であろう.

この直感を数値化できないか？

より汎用的な解析アプローチ

Hoeffding の不等式 [4]

$X_1, \ldots, X_n \in \{0, 1\}$ なので, 以下の不等式が成り立つ.

$$\mathbf{P}\left\{|\overline{X} - \mathbf{E}\,X| > \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 n\right)$$

機械学習の設定に戻る

$X_i = L(h; \boldsymbol{x}_i, y_i) = I\{h(\boldsymbol{x}_i) \neq y_i\}$ とすると,

$$\mathbf{P}\left\{|\widehat{R}(h) - R(h)| > \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 n\right).$$

要注意：あらかじめこの候補 h を固定しないといけない.

[4] Hoeffding は Chebyshev とともに, 本資料の `rectangles.pdf` ではすでに登場している.

より汎用的な解析アプローチ

データ依存性の影響

Hoeffding の不等式は統計的独立性を有する観測データを仮定している.

前スライドの不等式が得られたのは,

$$(X_1, Y_1), \ldots, (X_n, Y_n) \text{ が独立} \implies L(h; X_1, Y_1), \ldots, L(h; X_n, Y_n) \text{ が独立}$$

が成り立つからである.

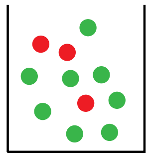


Figure: 候補 h_1 の損失値

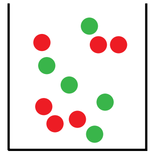


Figure: 候補 h_2 の損失値

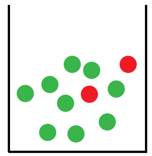


Figure: 候補 h_3 の損失値

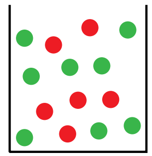
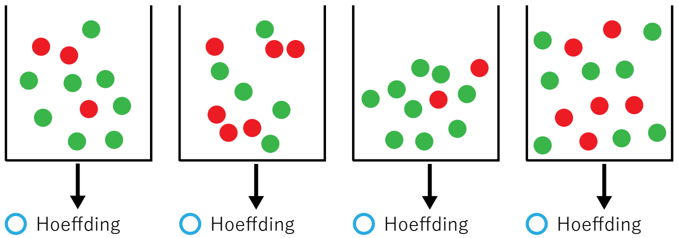


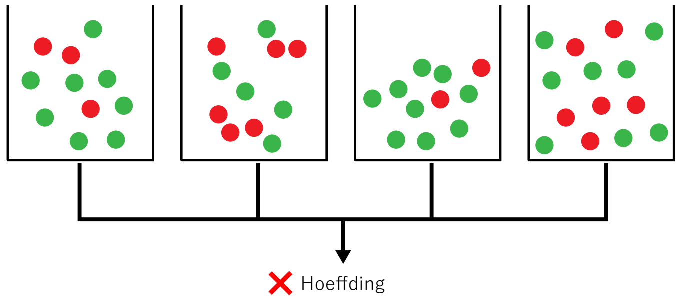
Figure: 候補 h_4 の損失値

より汎用的な解析アプローチ



独立性があれば問題はないが, この前提が崩れると, 何も言えなくなる.

より汎用的な解析アプローチ



独立性があれば問題はないが, この前提が崩れると, 何も言えなくなる.

より汎用的な解析アプローチ

問題点

学習アルゴリズムの出力はデータの全体に依存する.

$$\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \mapsto \widehat{h} \in \mathcal{H}$$

したがって, データ依存の候補となると,

$$\left\{L(\widehat{h}; X_i, Y_i)\right\}_{i=1}^{n} \text{ は独立でない} \implies \text{Hoeffding の不等式が不成立.}$$ [5]

もちろん, 事前にどの候補を返すかはわからないから困る...

[5] 演習では, これを入念に調べていただく課題を用意している.

より汎用的な解析アプローチ

一つの解決策

モデル $\mathcal{H}$ が有限であれば, いわゆる union bound を使うことができる.

任意の学習則の出力 $\widehat{h}$ について,

$$\begin{aligned}
\mathbf{P}\left\{|\widehat{R}(\widehat{h}) - R(\widehat{h})| > \varepsilon\right\} &\leq \mathbf{P} \bigcup_{h \in \mathcal{H}} \left\{|\widehat{R}(h) - R(h)| > \varepsilon\right\} \\
&\leq \sum_{h \in \mathcal{H}} \mathbf{P}\left\{|\widehat{R}(h) - R(h)| > \varepsilon\right\} \\
&\leq 2|\mathcal{H}| \exp\left(-2\varepsilon^2 n\right).
\end{aligned}$$

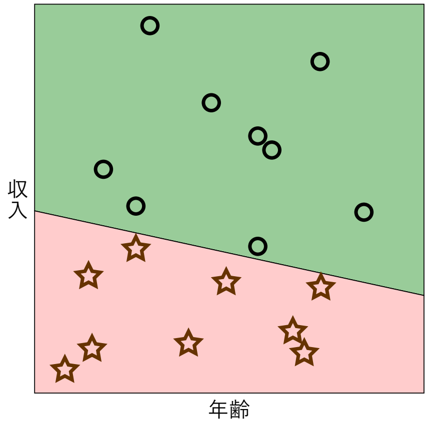
より汎用的な解析アプローチ

換言すると, $1 - \delta$ 以上の確率で,
以下が成り立つ.

$$R(\widehat{h}) \leq \widehat{R}(\widehat{h}) + \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{n}}.$$

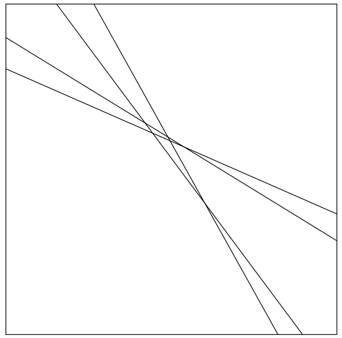
一見して良さそうだが...

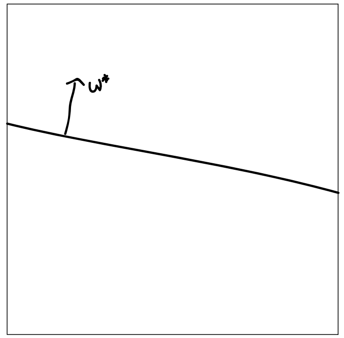
- $\widehat{h}_{\text{PLA}}$ は右辺第 1 項を最小化.
- PLA では $|\mathcal{H}| = \infty$...



モデルの表現力について

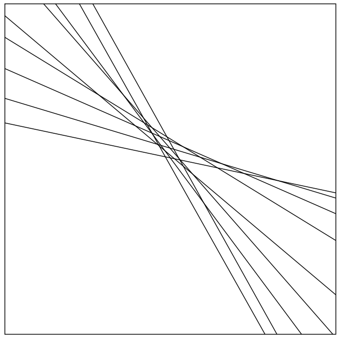
モデルは実世界を模して作られる. 表現力が高い＝多様な現象を説明する.

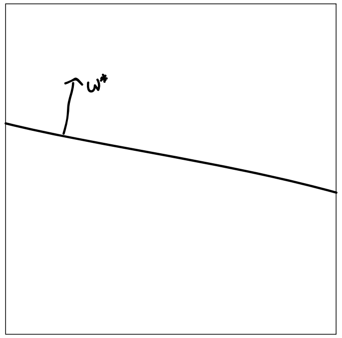




モデルの表現力について

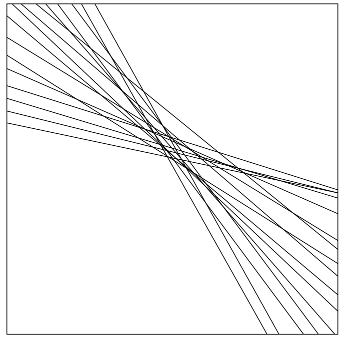
モデルは実世界を模して作られる. 表現力が高い＝多様な現象を説明する.

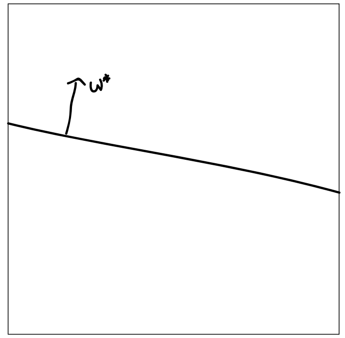




モデルの表現力について

モデルは実世界を模して作られる. 表現力が高い＝多様な現象を説明する.





モデルの表現力について

事前知識がない分, いうまでもなく豊富な選択肢が必要になる.

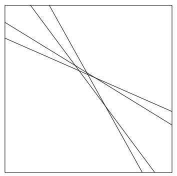


Figure: $\mathcal{H}_1$

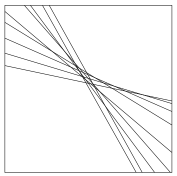


Figure: $\mathcal{H}_2$

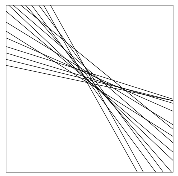


Figure: $\mathcal{H}_3$

しかし，$|\mathcal{H}_1| < |\mathcal{H}_2| < |\mathcal{H}_3|$ なので表現力の対価を払う.

$$\mathbf{P}\left\{|\widehat{R}(\widehat{h}) - R(\widehat{h})| > \varepsilon\right\} \le 2|\mathcal{H}|\exp\left(-2\varepsilon^2 n\right)$$

モデルの表現力について

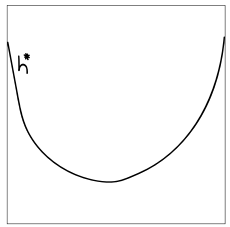


Figure: 真の識別ルール.

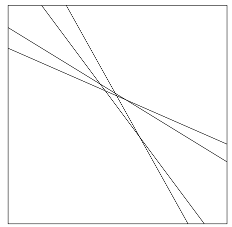


Figure: 線形モデル $\mathcal{H}_{\text{lin}}$.

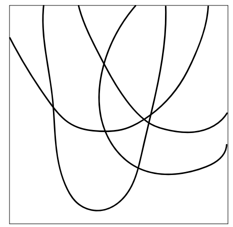


Figure: 非線形モデル $\mathcal{H}_{\text{non}}$.

事前知識がない場合，$\mathcal{H} = \mathcal{H}_{\text{lin}} \cup \mathcal{H}_{\text{non}}$ が無難. しかし，

- $\mathcal{H}$ が大きくなり, 汎化能力の保証が悪化
- 線形・非線形が混在してくると，$\widehat{R}$ の最小化も不透明に...

表現力とモデル設計の考え方

データを貨幣と思えば良い
高級なモデルを使う場合, それ相応の
データがないと, 性能保証はない.

$$R(\hat{h}) \leq \hat{R}(\hat{h}) + \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{n}}$$

解釈に注意
あくまで上界に基づく原則.
それでも過学習は実際によく見られる.

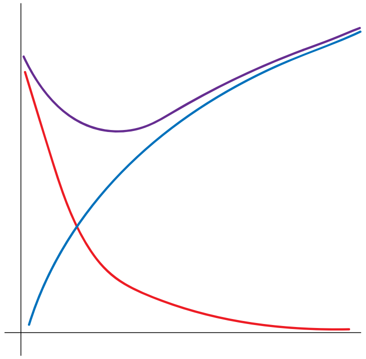


Figure: 横軸は表現力の高さ (右ほど高い).

表現力とモデル設計の考え方

データを貨幣と思えば良い
高級なモデルを使う場合, それ相応の
データがないと, 性能保証はない.

$$R(\hat{h}) \leq \hat{R}(\hat{h}) + \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{n}}$$

解釈に注意
あくまで上界に基づく原則.
それでも過学習は実際によく見られる.

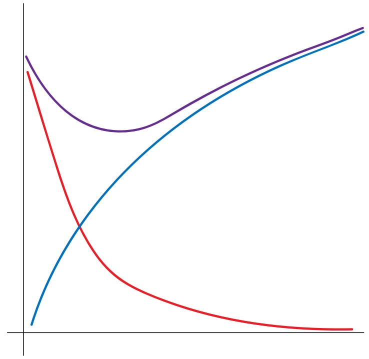


Figure: 横軸は表現力の高さ (右ほど高い).

表現力とモデル設計の考え方

データを貨幣と思えば良い
高級なモデルを使う場合, それ相応の
データがないと, 性能保証はない.

$$R(\hat{h}) \leq \hat{R}(\hat{h}) + \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{n}}$$

解釈に注意
あくまで上界に基づく原則.
それでも過学習は実際によく見られる.

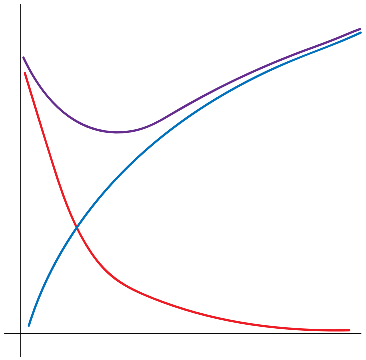


Figure: 横軸は表現力の高さ (右ほど高い).

再び：より汎用的な解析アプローチ

さて，PLA の場合は $|\mathcal{H}| = \infty$ なので, 性能保証はまだ得られていない.

表現力の指標を考える
そもそも,「表現力＝要素の数」という考え方には無理がある.

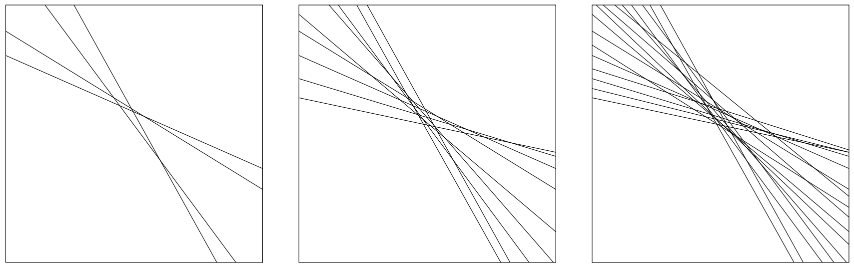
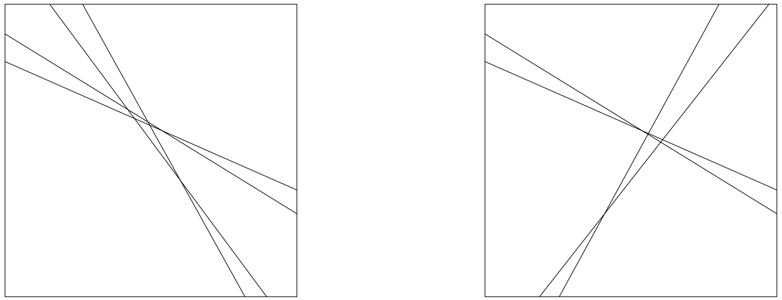


Figure: より緻密に識別できるが, 数が増えても対応できていない領域もある.

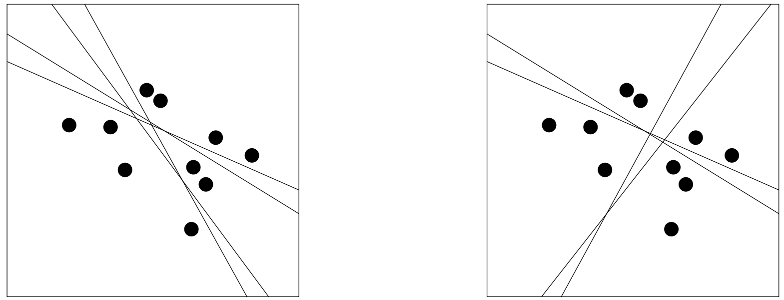
再び：より汎用的な解析アプローチ

新しい表現力の指標

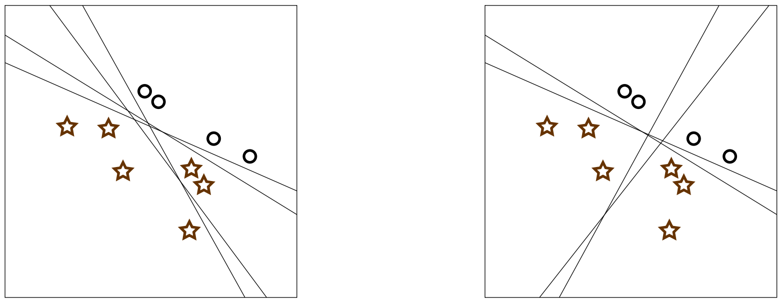
冗長性をなくして, 実質的な表現力を捉えるのは VC 次元である.
VC 次元では, 数ではなく,「表現力＝データから見た識別能力」である.



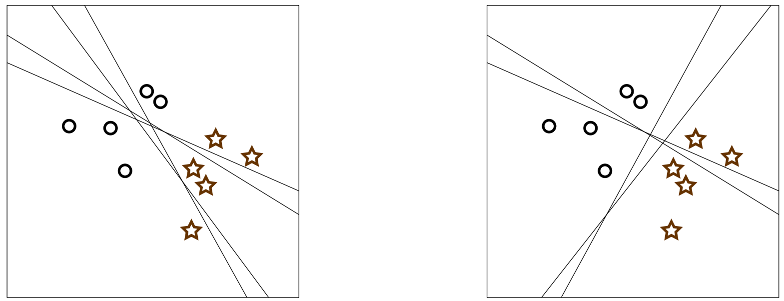
再び：より汎用的な解析アプローチ
新しい表現力の指標
冗長性をなくして, 実質的な表現力を捉えるのは VC 次元である.
VC 次元では, 数ではなく,「表現力＝データから見た識別能力」である.



再び：より汎用的な解析アプローチ
新しい表現力の指標
冗長性をなくして, 実質的な表現力を捉えるのは VC 次元である.
VC 次元では, 数ではなく,「表現力＝データから見た識別能力」である.



再び：より汎用的な解析アプローチ
新しい表現力の指標
冗長性をなくして, 実質的な表現力を捉えるのは VC 次元である.
VC 次元では, 数ではなく,「表現力＝データから見た識別能力」である.



再び：より汎用的な解析アプローチ

具体例から VC 次元の定義を導き出す
平面にはまず，お好きな 1 点 (データ $\{x_1\}$) を描いてください.

- この点を正例とする識別線は描ける？
- この点を負例とする識別線は描ける？

「できた」という場合, 変数 $k \leftarrow 1$ を初期化する.

再び：より汎用的な解析アプローチ

次は先ほどとは別の平面に, 今度は 2 点 $\{x_1, x_2\}$ を描いてください.

- 両方とも正例になる識別線は描ける？
- 両方とも負例になる識別線は描ける？
- それぞれが正例・負例となるような識別線は描ける？

できた場合, 例の変数を一つ増やして $k \leftarrow k+1$ とする $(k=2)$.[6]

[6] もちろん, この 2 点が重なって $x_1 = x_2$ のとき, 上記の識別はできない.

再び：より汎用的な解析アプローチ

次は 3 点 $\{x_1, x_2, x_3\}$ を描いてください.

識別の組み合わせが $2^3 = 8$ 通りある：

- $(+1,+1,+1), (-1,-1,-1)$
- $(+1,+1,-1), (+1,-1,+1), (-1,+1,+1)$
- $(-1,-1,+1), (-1,+1,-1), (+1,-1,-1)$

線を適切に選べば, 上記の正負ラベルの割り振りをすべて実現できる？

できた場合, 例の変数を一つ増やして $k \leftarrow k+1$ とする $(k=3)$.

再び：より汎用的な解析アプローチ

次は 4 点 $\{x_1, x_2, x_3, x_4\}$ を描いてください.

識別の組み合わせが $2^4 = 16$ 通りある：

- $(+1,+1,+1), (-1,-1,-1)$
- $(+1,+1,-1), (+1,-1,+1), (-1,+1,+1)$
- $(-1,-1,+1), (-1,+1,-1), (+1,-1,-1)$

線を適切に選べば, 上記の正負ラベルの割り振りをすべて実現できる？

再び：より汎用的な解析アプローチ

次は 4 点 $\{x_1, x_2, x_3, x_4\}$ を描いてください.

識別の組み合わせが $2^4 = 16$ 通りある：

- $(+1, +1, +1), (-1, -1, -1)$
- $(+1, +1, -1), (+1, -1, +1), (-1, +1, +1)$
- $(-1, -1, +1), (-1, +1, -1), (+1, -1, -1)$

線を適切に選べば, 上記の正負ラベルの割り振りをすべて実現できる？

⋮

再び：より汎用的な解析アプローチ

次は 4 点 $\{x_1, x_2, x_3, x_4\}$ を描いてください.

識別の組み合わせが $2^4 = 16$ 通りある：

- $(+1, +1, +1), (-1, -1, -1)$
- $(+1, +1, -1), (+1, -1, +1), (-1, +1, +1)$
- $(-1, -1, +1), (-1, +1, -1), (+1, -1, -1)$

線を適切に選べば, 上記の正負ラベルの割り振りをすべて実現できる？

⋮

いくら頑張っても, データが 4 点あるとできない.

ここで終了して，このモデルの VC 次元 $:= d_{\mathrm{VC}}(\mathcal{H}) = k = 3$.

再び：より汎用的な解析アプローチ

何のモデル？

先ほどは 2 次元平面に任意の線を描いていたので, モデルは

$$\mathcal{H} = \left\{ x \mapsto \mathrm{sign}(w^\top x - w_0) : (w, w_0) \in \mathbb{R}^3 \right\},$$

つまり平面上のパーセプトロンにほかならない.[7]

もう少し一般的には？

2 次元平面ではなく，d 次元の線形空間のときのパーセプトロンならば,

$$d_{\mathrm{VC}}(\mathcal{H}) = d + 1. \tag{1}$$

[7] 本資料の learning_intro.pdf では具体例とともにパーセプトロンを紹介している.

再び：より汎用的な解析アプローチ

性能評価との関係

導出自体は割愛するが, 任意の学習則について以下の不等式が $1 - \delta$ 以上の確率で成り立つ.

$$R(\widehat{h}) \leq \widehat{R}(\widehat{h}) + \sqrt{\frac{8}{n} \log\left(\frac{4((2n)^{d_{\mathrm{VC}}(\mathcal{H})} + 1)}{\delta} \right)}$$

嬉しいのは, モデルが無限で $|\mathcal{H}| = \infty$ でも, 上記のバウンドが使える.[8]

[8] パーセプトロンの場合は $d_{\mathrm{VC}} = d + 1$ と代入すれば良い.

再び：表現力とモデル設計の考え方

異なる表現力指標でも, 考え方は一緒

$$R(\widehat{h}) \leq \widehat{R}(\widehat{h}) + \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{n}}$$

$$R(\widehat{h}) \leq \widehat{R}(\widehat{h}) + \sqrt{\frac{8}{n}\log\left(\frac{4((2n)^{d_{\mathrm{VC}}(\mathcal{H})}+1)}{\delta}\right)}$$

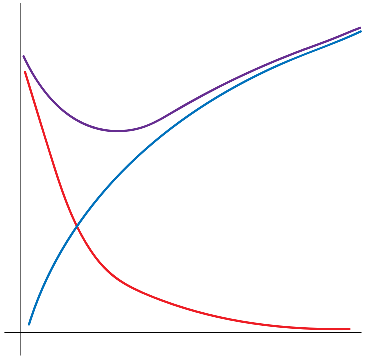


Figure: 横軸は表現力の高さ (右ほど高い).

まとめ

表現力の指標

- 長方形と平面の相違点.
- データの独立性と Hoeffding の不等式.
- 有限モデルの場合の性能保証の導出.
- 識別能力に基づく表現力指標としての VC 次元.

モデルの設計

- 表現力が高くなると, 限られたデータから最良の候補を選ぶことが難しくなる.
- これと連動して, 従来の表現力指標に基づくリスク上界も緩くなる.
- 知見として, 表現力に見合ったデータ数がないと保証が消える.

※関心のある方は付属の演習課題にも取り組んでみてください.

参考文献

Abu-Mostafa, Y., Song, X., Nicholson, A., and Magdon-Ismail, M. (2004). The bin model. Technical report, California Institute of Technology.