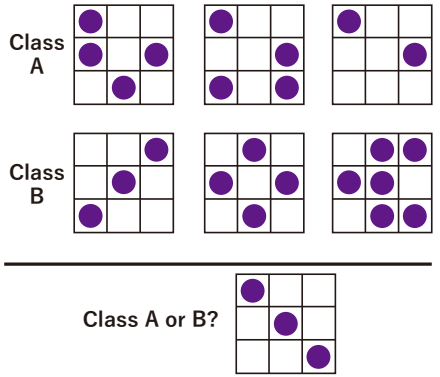
実データで学ぶ人工知能講座

長方形の例からAIの性能保証を考える

マシュー　ホーランド
Matthew J. Holland
matthew-h@ar.sanken.osaka-u.ac.jp

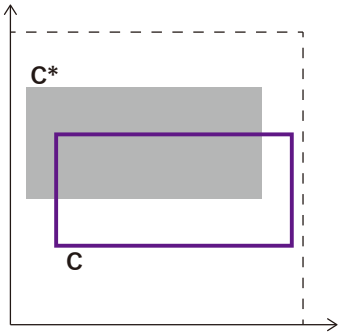
大阪大学 産業科学研究所 助教

学習の可能性？[1]



[1] この例は Abu-Mostafa et al. (2012) による. 画像は自作.

長方形の学習[2]



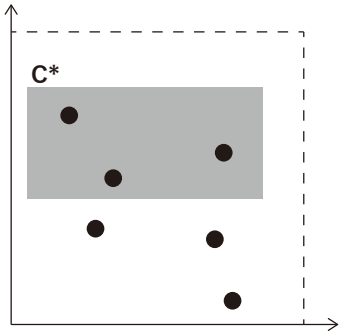
学習課題の概要

- 平面上の長方形 C^* を知りたい.
- データに基づいて, 候補 C を選ぶ.
- 二値識別と表裏一体である.

「ちょうど良い天気」「小太りの体型」など,
色々と現実世界の概念が当てはまる.

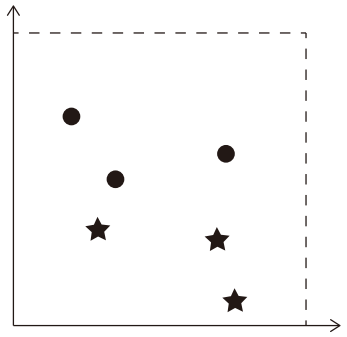
[2] Blumer et al. (1989) が考案した明快な例から着想を得た.

長方形の学習



いうまでもなく C^* は未知である.

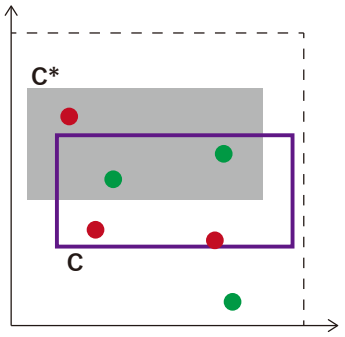
長方形の学習



情報として得られるのは, 各データ点が C^* に入っているかどうか, これだけである.

正例：●
負例：★

長方形の学習



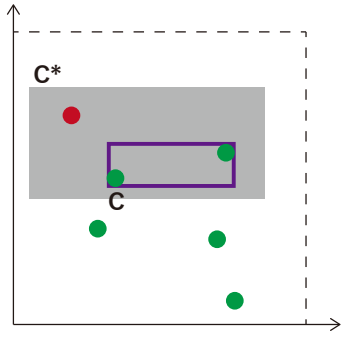
データがあれば, これをフィードバックとして, 今の C よりも良い位置に動かすことができる.

みんなで考えよう. どのように更新する？

重要なのは：

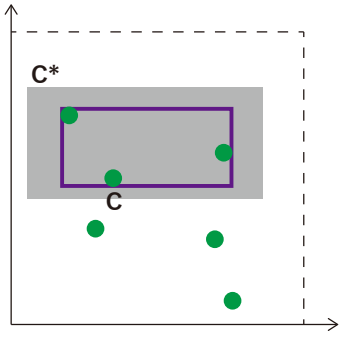
- 計算機でプログラムできること.
- データのみを頼りにすること.

長方形の学習



左図の候補では微妙な気がする...

長方形の学習



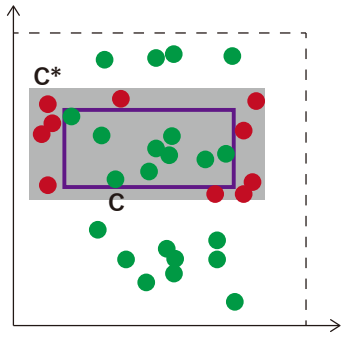
簡明な方法：正例をぴったりと包み込む.

計算機で実装できそう？

アルゴリズム (正例の包絡)

1. 正例を全部かき集める.
2. 横軸の最大値と最小値を記録.
3. 縦軸の最大値と最小値を記録.
4. 極値によって新しい C の境界を決定.

長方形の学習



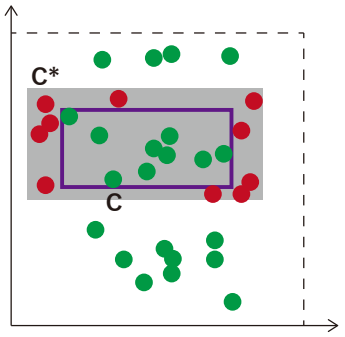
既存データとの整合性よりも, これから入ってくるデータをどう分けるかが重要である.

データの数が増えると, より良い候補が得られそうだが, 実際のところ,

いくらあれば「十分」

といえるだろうか.

長方形の学習



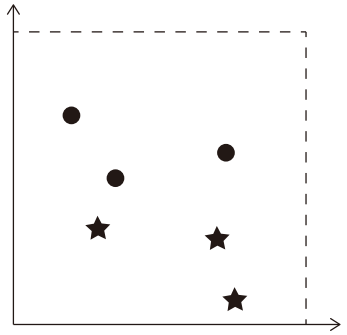
既存データとの整合性よりも, これから入ってくるデータをどう分けるかが重要である.

データの数が増えると, より良い候補が得られそうだが, 実際のところ,

いくらあれば「十分」

といえるだろうか.

学習問題の定式化

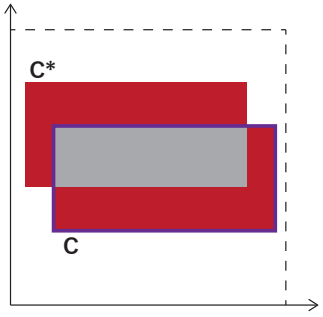


データは下記のように表記する.

- データの確率変数：$(X_1, Y_1), \ldots, (X_n, Y_n)$.
- 観測データ：$(x_1, y_1), \ldots, (x_n, y_n)$.
- 入力データ：$x_i = (x_{i,1}, x_{i,2}) \in [0,1]^2$.
- ラベルづけ：

$$y_i = \begin{cases} +1, & x_i \in C^* \\ -1, & x_i \notin C^* \end{cases}$$

学習問題の定式化



出来栄えをどう評価すれば良い？

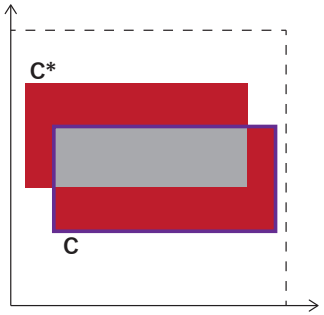
確率モデルを導入する.

$$R(C) := \mathbf{P}\{X \in (C^* \cup C) \setminus (C^* \cap C)\}.$$

たとえ $C \neq C^*$ であっても, 高確率の領域で合致していれば十分であろう.

- 一様分布の場合は，$R(C)$ = 赤い領域の面積.
- 一般には，$R(C)$ は赤い領域に点が入る確率.

学習問題の定式化



データに基づいて選んだ候補を $\widehat{C}$ と書く.

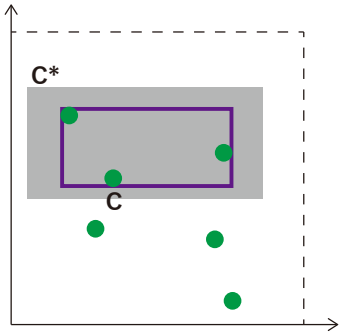
許容範囲を決める
誤差は ε 以下が望ましい.
それを超える確率が δ 以下である.

- $R(\widehat{C}) \leq \varepsilon$
- $\mathbf{P}\{R(\widehat{C}) > \varepsilon\} \leq \delta$

上記の確率 $\mathbf{P}$ は標本 $(X_1, Y_1), \ldots, (X_n, Y_n)$ の分布である.

さて, データ数 n をいくら取れば良い？

正例包絡の誤差解析



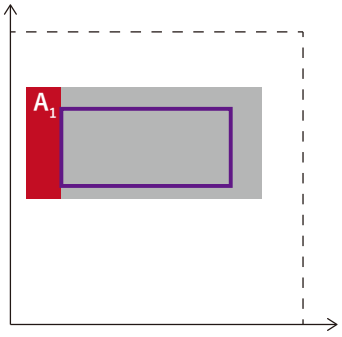
正例包絡の候補を $\widehat{C}_{\text{fit}}$ と書く.

正例が一つもない場合，$\widehat{C}_{\text{fit}} = \emptyset$ とする.

事実：$\widehat{C}_{\text{fit}} \subset C^*$ が常に成り立つ.

また, 左図の「誤差域」はいずれも
$A_1, A_2, A_3, A_4 \subset C^*$.

正例包絡の誤差解析



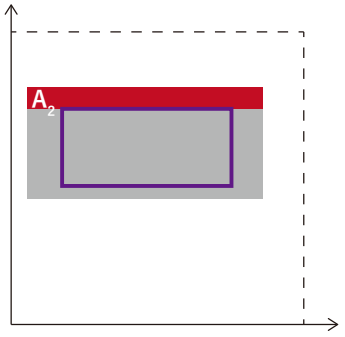
正例包絡の候補を $\widehat{C}_{\text{fit}}$ と書く.

正例が一つもない場合，$\widehat{C}_{\text{fit}} = \emptyset$ とする.

事実：$\widehat{C}_{\text{fit}} \subset C^*$ が常に成り立つ.

また, 左図の「誤差域」はいずれも
$A_1, A_2, A_3, A_4 \subset C^*$.

正例包絡の誤差解析



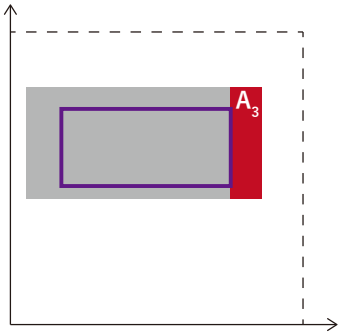
正例包絡の候補を $\widehat{C}_{\text{fit}}$ と書く.

正例が一つもない場合，$\widehat{C}_{\text{fit}} = \emptyset$ とする.

事実：$\widehat{C}_{\text{fit}} \subset C^*$ が常に成り立つ.

また, 左図の「誤差域」はいずれも
$A_1, A_2, A_3, A_4 \subset C^*$.

正例包絡の誤差解析



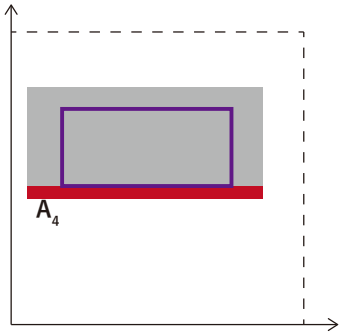
正例包絡の候補を $\widehat{C}_{\text{fit}}$ と書く.

正例が一つもない場合，$\widehat{C}_{\text{fit}} = \emptyset$ とする.

事実：$\widehat{C}_{\text{fit}} \subset C^*$ が常に成り立つ.

また, 左図の「誤差域」はいずれも
$A_1, A_2, A_3, A_4 \subset C^*$.

正例包絡の誤差解析



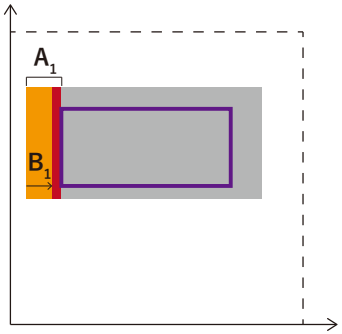
正例包絡の候補を $\widehat{C}_{\text{fit}}$ と書く.

正例が一つもない場合，$\widehat{C}_{\text{fit}} = \emptyset$ とする.

事実：$\widehat{C}_{\text{fit}} \subset C^*$ が常に成り立つ.

また, 左図の「誤差域」はいずれも
$A_1, A_2, A_3, A_4 \subset C^*$.

正例包絡の誤差解析



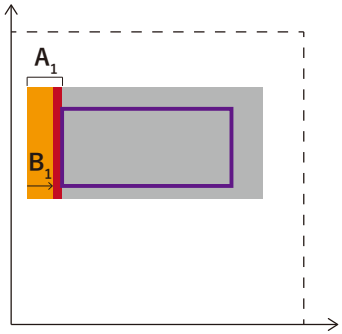
誤差の許容範囲を $0 < \varepsilon$ 以下とする.

そこで「高確率域」B_1, B_2, B_3, B_4 を定める.
左図と下記の式を参照.

$$\mathbf{P}\{X \in B_i\} = \frac{\varepsilon}{4}, \quad i = 1, \ldots, 4.$$

上記の等式を満たせない場合，$B_i = C^*$ にしておけば少なくとも $\mathbf{P}\{X \in B_i\} \leq \varepsilon/4$.

正例包絡の誤差解析



まず, 以下が成り立つ.

$$C^* \setminus \widehat{C}_{\text{fit}} \subseteq \bigcup_{i=1}^{4} A_i. \tag{1}$$

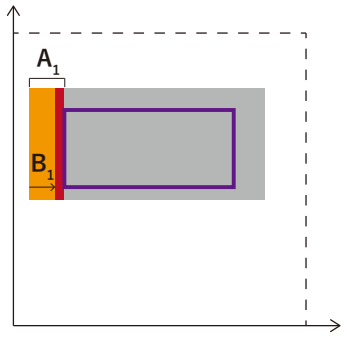
次の包含関係が仮に成り立つとする.

$$A_i \subset B_i, \quad i = 1, \ldots, 4. \tag{2}$$

その場合, 以下の不等式も成り立つ.

$$R(\widehat{C}_{\text{fit}}) \leq \varepsilon. \tag{3}$$

正例包絡の誤差解析

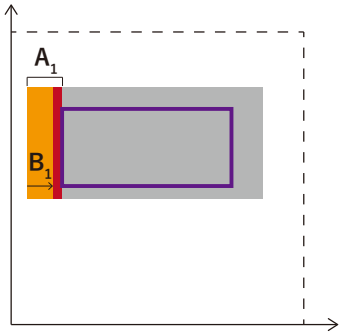


式 (2) から式 (3) を導き出せるのは, 下記の展開からわかる.

$$
\begin{aligned}
R(\widehat{C}_{\text{fit}}) &= \mathbf{P}\left\{X \in C^* \setminus \widehat{C}_{\text{fit}}\right\} \\
&\le \mathbf{P}\left\{X \in \bigcup_{i=1}^{4} A_i\right\}, \text{ by (1)} \\
&\le \mathbf{P}\left\{X \in \bigcup_{i=1}^{4} B_i\right\}, \text{ by (2)} \\
&\le \sum_{i=1}^{4} \mathbf{P}\left\{X \in B_i\right\} \\
&\le 4\,(\varepsilon/4) \\
&= \varepsilon.
\end{aligned}
$$

Note: $(C^* \cup \widehat{C}_{\text{fit}}) \setminus (C^* \cap \widehat{C}_{\text{fit}}) = C^* \setminus \widehat{C}_{\text{fit}}$.

正例包絡の誤差解析



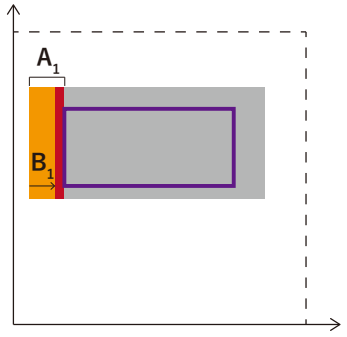
対偶をとると,

$$R(\widehat{C}_{\text{fit}}) > \varepsilon \text{ implies } B_i \subset A_i \text{ for some } i. \qquad (4)$$

言葉でいうと,

ε 以上の誤差を食らった場合,
少なくとも一つの誤差域 A_i が
高確率域 B_i を含む.

正例包絡の誤差解析



これまでの話をデータセット
$(X_1, Y_1), \ldots, (X_n, Y_n)$ に結びつけたい.

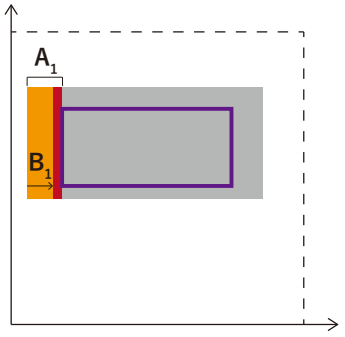
$B_i \subset A_i$ と, 以下のことは等価である.

$$X_j \notin B_i, \text{ for all } j = 1, \ldots, n. \qquad (5)$$

少し考えれば, これは自明である.

なぜなら, 一つでも X_j が B_i に入っていれば, 正例包絡の候補は定義上 B_i と重なるからである.

正例包絡の誤差解析



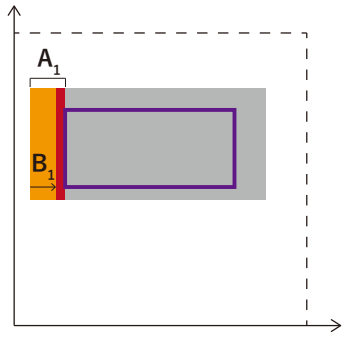
データセットの全点が特定の高確率域から外れることはほぼ起こらない.

$$
\begin{aligned}
\mathbf{P}\left\{B_i \subset A_i\right\} &= \mathbf{P}\left\{X_j \notin B_i, \text{ all } j\right\} \\
&= \left(1 - \frac{\varepsilon}{4}\right)^n.
\end{aligned}
$$

ε を小さくとれば $0 < 1 - \varepsilon/4 < 1$ となる.

したがって, データ数 n の増加によって, 上記の確率が急速に減少する.

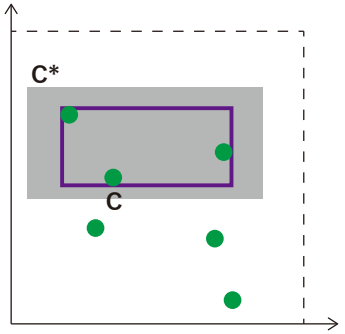
正例包絡の誤差解析



式 (4) に戻ると，ε よりも大きな誤差を被ってしまう確率は高くても

$$\begin{aligned}\mathbf{P}\left\{R(\widehat{C}_{\text{fit}}) > \varepsilon\right\} &\leq \mathbf{P}\left\{B_i \subset A_i, \text{ some } i\right\} \\ &\leq \sum_{i=1}^{4} \mathbf{P}\left\{B_i \subset A_i\right\} \\ &= \sum_{i=1}^{4} \mathbf{P}\left\{X_j \notin B_i, \text{ for all } j\right\}, \text{ by (5)} \\ &= 4\left(1-\frac{\varepsilon}{4}\right)^n.\end{aligned}$$

正例包絡の誤差解析



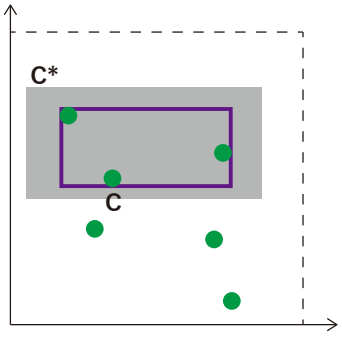
したがって，誤差の許容範囲を ε とすれば，正例包絡の $\widehat{C}_{\text{fit}}$ がこの範囲を超える確率は高くても

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

要点：

- 考えている誤差 $R(\widehat{C}_{\text{fit}})$ は確率であるが，これ自体もランダム (＝確率変数) である．なぜ？
- データの確率分布について，何一つ仮定は置いていない．
- よって，これらの上界は「悲観的」．

正例包絡の誤差解析



数値例を一つ考えよう．許容範囲を以下のようにする．

$$R(\widehat{C}_{\text{fit}}) \leq 0.1.$$

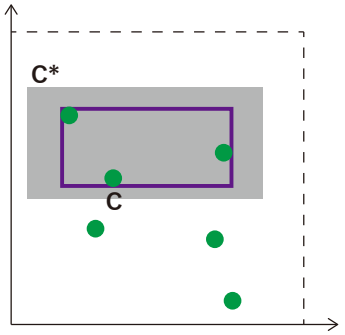
そのとき，もし $n = 10$ ならば，

$$\mathbf{P}\left\{R(\widehat{C}_{\text{fit}}) > 0.1\right\} \leq 4\left(1-\frac{0.1}{4}\right)^{10} \approx 0.78.$$

もし $n = 100$ ならば，

$$\mathbf{P}\left\{R(\widehat{C}_{\text{fit}}) > 0.1\right\} \leq 4\left(1-\frac{0.1}{4}\right)^{100} \approx 0.08.$$

正例包絡の誤差解析



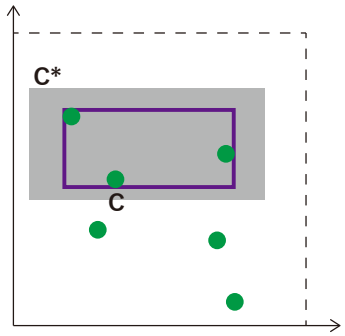
一定の性能を高い確率で約束したいときは？

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

δ 以下の確率に抑えたいなら，以下で十分である．

$$\varepsilon \geq \frac{4\log(4\delta^{-1})}{n} \quad \text{or} \quad n \geq \frac{4\log(4\delta^{-1})}{\varepsilon}.$$

正例包絡の誤差解析



99% の確信度でいえるのは...
もし $n = 10$, 約束できるのは

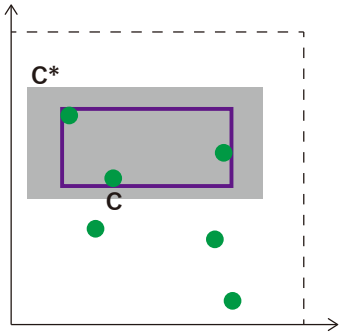
$$R(\widehat{C}_{\text{fit}}) \leq \varepsilon = \frac{4\log(400)}{10} \approx 2.40.$$

もし $n = 100$, 話が良くなる.

$$R(\widehat{C}_{\text{fit}}) \leq \varepsilon = \frac{4\log(400)}{100} \approx 0.24.$$

前者は意味のある約束ができない.
後者では, 効力のある性能保証が付いている.

正例包絡の誤差解析



前もって誤差範囲と確信度を両方とも決めた場合は？

95% の確率で $\varepsilon = 0.1$ 以内の誤差に抑えたいとする. そのとき, 以下を満たせば十分である.

$$n \geq \frac{4\log(4/0.05)}{0.1} \approx 175.$$

ある ε と δ を実現するのに最低必要な n を標本複雑度 (sample complexity) と呼ぶ.

PAC 学習と統計的決定理論

PAC 学習は Valiant (1984) によって考案された.

性能保証は欲しいが, 完璧な性能を確実に担保することは現実的ではない.

- $(1 - \delta)$ の確信度 (**P**robably)
- 最大でも ε 誤差 (**A**pproximately)

PAC: **P**robably **A**pproximately **C**orrect.

ここでいう correct とは, どういうことか.

PAC 学習と統計的決定理論

$Z_1, \ldots, Z_n \sim \mathsf{P}$
無作為にサンプルされたデータの確率変数. 観測値を小文字 z_i で表記.

$h \in \mathcal{H}$
学習課題のほとんどは集合から要素を選ぶことに帰着する.

$L(h; z)$
候補 h に数値的な罰則 (フィードバック) を与える損失関数.

$R(h) := \mathbf{E}_{\mathsf{P}}\, L(h; Z)$
損失関数の期待値をリスクと呼ぶ.

いわゆるリスク最小化の枠組みは Wald (1949) 以前にも遡る.

PAC 学習と統計的決定理論

$\{Z_1, \ldots, Z_n\} \mapsto \widehat{h}$
学習アルゴリズムはデータセットから候補への写像である.

リスクは, 学習機の「汎化性能」を測る典型的な指標である：

- n 個の標本数から候補 $\widehat{h}$ を選ぶ.
- $R(\widehat{h})$ が小さければ, これから入ってくるデータに対して, 平均的に損失値が小さい.

いうまでもなく, 本当に重要なのは損失値の分布である.

汎化性能のより包括的な評価をするには，$Z \sim \mathrm{P}$ からサンプルしたときの $L(\widehat{h}; Z)$ の確率分布を見る必要がある.

PAC 学習と統計的決定理論

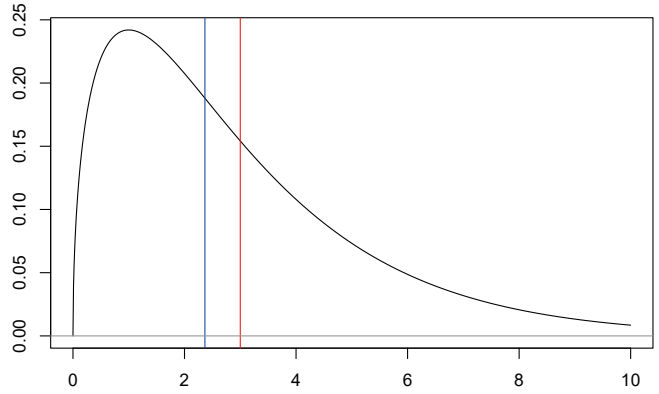


Figure: $R(h) = \mathbf{E}_{\mathrm{P}}\, L(h; Z)$ は損失値の分布を実数値一つで要約したものと捉えれば良い. もちろん代替物も山ほどある.

PAC 学習と統計的決定理論

ここで, ある問題に直面する.

$$\mathrm{P} \text{ は未知} \implies \text{リスク } R \text{ は未知.}$$

計算できないリスクには一体, 何の意味がある？

PAC 学習と統計的決定理論

幸い, 確率論が救いの手を差し伸べてくれる.

- 観測データを得て, 学習則を実行し，$\widehat{h}$ を選ぶ.
- よって，$\widehat{h}$ はデータ標本 $\{Z_1, \ldots, Z_n\}$ に依存する.
- したがって，$R(\widehat{h})$ 自体も確率変数である.

上記を踏まえて，$R(\widehat{h})$ が計算できないなら, 信頼区間を見れば良い.

$$\mathbf{P}\left\{R(\widehat{h}) > \varepsilon(n, \delta)\right\} \leq \delta.$$

ここで先ほどの PAC 学習の概念が鮮明に見えてくる.

PAC 学習と統計的決定理論

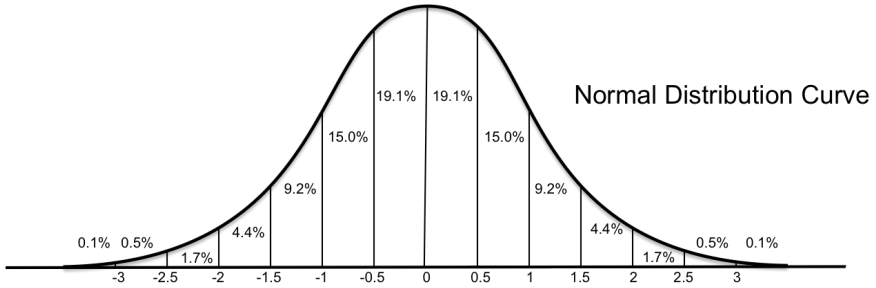
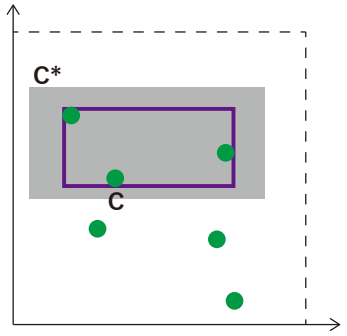


Figure: 古典統計学からで重宝されてきた信頼区間といえば, 鐘形の正規分布が思い浮かぶ. 機械学習では, 損失値の分布の形状は知らないが, それでもその「裾」について明確に論じることは可能.[3]

[3]画像出処：Confidence Intervals. Brilliant.org. Retrieved 19:26, July 5, 2019, from https://brilliant.org/wiki/confidence-intervals/.

PAC 学習と統計的決定理論



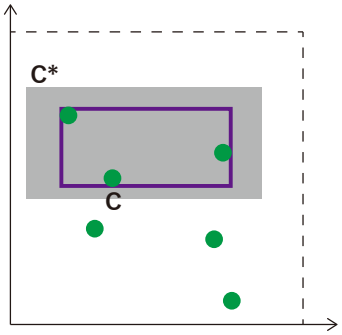
P が未知なら, 信頼区間なんて議論しようもないのでは...？

自明ではないが, これは可能である.

なぜこれが言える？
長方形の学習の議論ですでに示したから.

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

長方形の例を見つめなおす



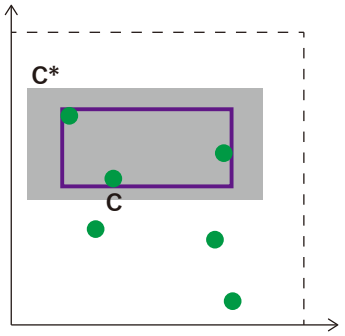
対応関係は明確：

- $\mathcal{H} \leftrightarrow$ 縦横軸と平行な辺を持つ長方形.
- $h \leftrightarrow$ 長方形 C.
- $Z \leftrightarrow$ 入出力のペア (X, Y).

今後の議論を明確にすべく, 候補 C を識別器として扱う.

$$h_C(x) = \begin{cases} +1, & \text{if } x \in C \\ -1, & \text{if } x \notin C \end{cases}$$

長方形の例を見つめなおす



性能評価の指標も明らかな対応関係が見いだせる.

- $L(h; z) \leftrightarrow$ 下記の損失関数.

$$L(C; z) = \begin{cases} 1, & \text{if } h_C(x) \neq y \\ 0, & \text{if } h_C(x) = y. \end{cases}$$

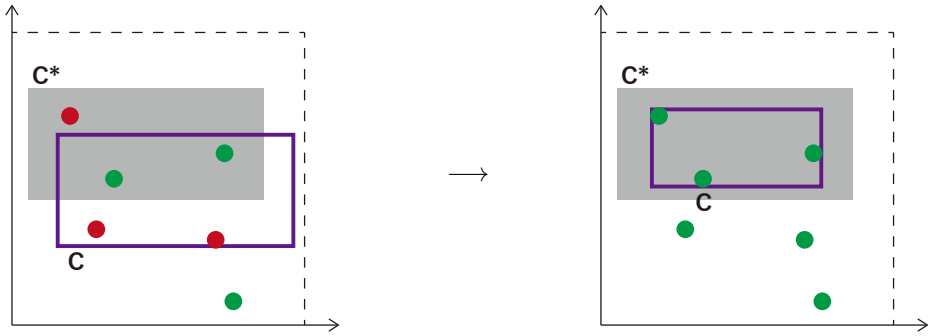
- $R(h) \leftrightarrow$ 下記の関数.

$$\begin{aligned} \mathbf{E}_{\mathsf{P}}\, L(C; Z) &= \mathbf{P}\left\{h_C(X) \neq Y\right\} \\ &= \mathbf{P}\left\{X \in (C \cup C^*) \setminus (C \cap C^*)\right\}. \end{aligned}$$

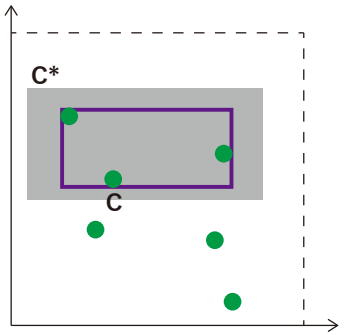
かくて長方形の学習は，PAC 学習的な性能保証つきのリスク最小化の具体例だとわかる.

期待損失最小化 (ERM)

候補の更新方法 (正例包絡) を思い出してみよう.



経験期待損失最小化 (ERM)



損失値の平均値を経験期待損失と呼ぶ.

$$\widehat{R}(C) \coloneqq \frac{1}{n}\sum_{i=1}^{n} L(C; Z_i) \approx R(C).$$

正例包絡アルゴリズムの定義上,

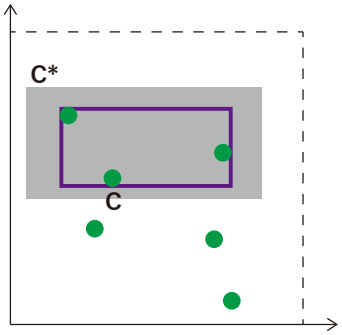
$$\widehat{R}(\widehat{C}_{\text{fit}}) = 0.$$

経験期待損失の最小化 (ERM) そのものである.

ERM は古典的な学習理論の大黒柱といえる.[4]

[4] Vapnik (1998); Mohri et al. (2012); Shalev-Shwartz and Ben-David (2014)

期待損失最小化 (ERM)

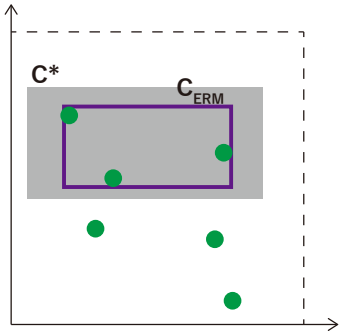


以下を満たすアルゴリズムを ERM 学習則と呼ぶ.

$$\widehat{h}_{\text{ERM}} \in \arg\min_{h \in \mathcal{H}} \widehat{R}(h)$$

実装方法が捨象されているため, 異なる ERM 学習則が異なる解に至ることは多々ある.

期待損失最小化 (ERM)

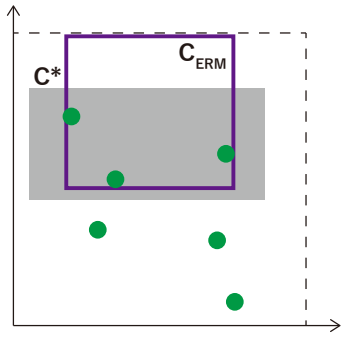


以下を満たすアルゴリズムを ERM 学習則と呼ぶ.

$$\widehat{h}_{\text{ERM}} \in \arg\min_{h \in \mathcal{H}} \widehat{R}(h)$$

実装方法が捨象されているため, 異なる ERM 学習則が異なる解に至ることは多々ある.

期待損失最小化 (ERM)

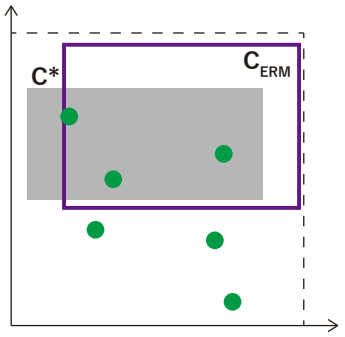


以下を満たすアルゴリズムを ERM 学習則と呼ぶ.

$$\widehat{h}_{\mathrm{ERM}} \in \arg\min_{h \in \mathcal{H}} \widehat{R}(h)$$

実装方法が捨象されているため, 異なる ERM 学習則が異なる解に至ることは多々ある.

期待損失最小化 (ERM)



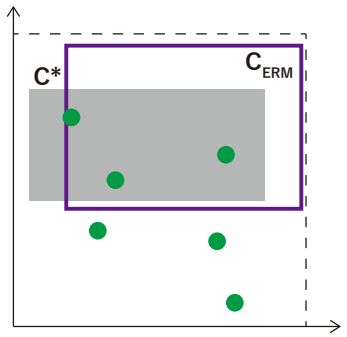
以下を満たすアルゴリズムを ERM 学習則と呼ぶ.

$$\widehat{h}_{\mathrm{ERM}} \in \arg\min_{h \in \mathcal{H}} \widehat{R}(h)$$

実装方法が捨象されているため, 異なる ERM 学習則が異なる解に至ることは多々ある.

「実データで学ぶ人工知能講座」機械学習の基礎　2020 iLDi 研究拠点 データビリティ人材育成教材　37

期待損失最小化 (ERM)



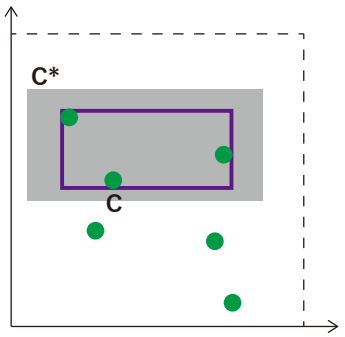
その結果, 汎化能力の意味では，ERM 学習則の間では, 優劣がつく.

例の正例包絡 $\widehat{C}_{\mathrm{fit}}$ は ERM 学習則で, 確率 $1-\delta$ で以下を満たす.

$$R(\widehat{C}_{\mathrm{fit}}) \leq \frac{4\log(4\delta^{-1})}{n}.$$

より一般的な ERM 学習則はどう？ERM でも悪い方法はたくさんありそう...

一般的な ERM 学習則の性能保証



ERM の基本戦略

1. $\widehat{R} \approx R$ を示す.
2. なるべく $\widehat{R}$ を小さくする.

より平易な言葉で

ステップ 1 :「本当の目的と計算可能な目的が最大どれほど離れているか？」

ステップ 2 :「計算可能な目的関数をいかにして最小にする？」

一般的な ERM 学習則の性能保証

ステップ 1 : $\widehat{R} \approx R$ を示す.

まず, 任意の確率変数 $X \geq 0$ と整数 $k > 0$ について, 以下が成り立つ.

$$\varepsilon^k I\{X > \varepsilon\} \leq X^k.$$

期待値をとって, ε^k で割れば Markov の不等式を得る.

$$\mathbf{P}\{X > \varepsilon\} \leq \frac{\mathbf{E}\, X^k}{\varepsilon^k}.$$

これにより, 有名な Chebyshev の不等式も自ずと出てくる.

$$\mathbf{P}\{|X - \mathbf{E}\, X| > \varepsilon\} \leq \frac{\mathbf{E}\,|X - \mathbf{E}\, X|^2}{\varepsilon^2} = \frac{\mathrm{var}(X)}{\varepsilon^2}. \tag{6}$$

一般的な ERM 学習則の性能保証

ステップ 1 : $\widehat{R} \approx R$ を示す.

次, 標本平均を下記のように表記し,

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

データが期待値 $\mathbf{E}\, X = \theta$ の Bernoulli 確率変数 $(X_1, \ldots, X_n \in \{0, 1\})$ であるとする. この場合, 以下が成り立つ.

$$\mathrm{var}(\overline{X}) = \frac{\mathrm{var}(X)}{n} = \frac{\theta(1-\theta)}{n}.$$

Chebyshev の不等式 (6) にある X を $\bar{x}$ に置き換えると,

$$\mathbf{P}\left\{|\overline{X} - \mathbf{E}\, X| > \varepsilon\right\} \leq \frac{\theta(1-\theta)}{\varepsilon^2 n}.$$

一般的な ERM 学習則の性能保証

ステップ 1 : $\widehat{R} \approx R$ を示す.

実は, より良いバウンドを Hoeffding (1963) の技法によって得られるケースが多い.

先ほどと同様にベルヌーイ確率変数の標本平均の誤差が ε を超える確率は高くても

$$\mathbf{P}\left\{|\overline{X} - \mathbf{E}\, X| > \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 n\right). \tag{7}$$

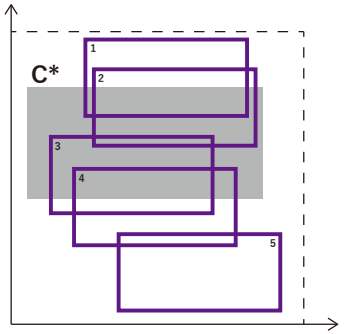
$\theta = 0.25, \varepsilon = 0.1$ のときの数値例 :

$n = 100$ case:
- **Chebyshev** ≈ 0.19.
- Hoeffding ≈ 0.27.

$n = 200$ case:
- Chebyshev ≈ 0.09.
- **Hoeffding** ≈ 0.04.

一般的な ERM 学習則の性能保証

ステップ 2 : なるべく $\widehat{R}$ を小さくする.



モデル $\mathcal{H}$ が有限 (左図の例では $|\mathcal{H}| = 5$) ならば, 話は簡単.

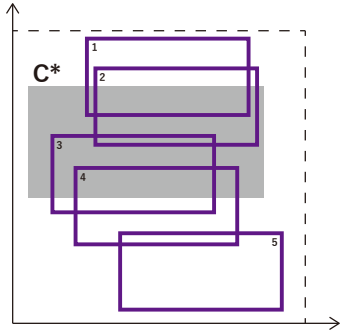
候補の数が少ない場合 :
愚直に全候補を見て, 成績トップを選ぶ.

候補の数が多い場合 :
ここが難しい...

候補の制約がない場合 :
正例包絡のアルゴリズムは当然使える.

一般的な ERM 学習則の性能保証

ステップ 2 :なるべく $\widehat{R}$ を小さくする.



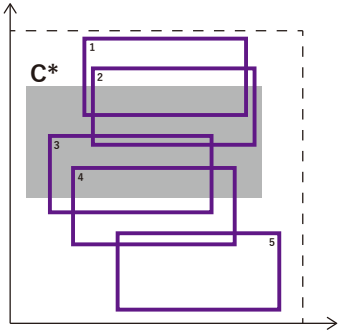
$\widehat{R}$ を小さくして, 何が言える？

任意の学習則の出力 $\widehat{h}$ について,

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

一般的な ERM 学習則の性能保証

ステップ 2 :なるべく $\widehat{R}$ を小さくする.

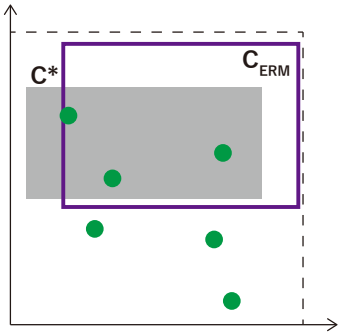


したがって，$1-\delta$ 以上の確率で,

$$R(\widehat{h}) \leq \widehat{R}(\widehat{h}) + \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{2n}} \qquad (8)$$

つまり，ERM は式 (8) の上界を最小化する.

ERM の優劣



アルゴリズムによって，$1-\delta$ の確率で約束できることを比較してみる.

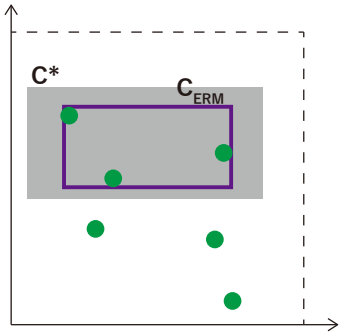
$$R(\widehat{C}_{\text{ERM}}) \leq \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{2n}}$$

vs.

$$R(\widehat{C}_{\text{fit}}) \leq \frac{4\log(4\delta^{-1})}{n}.$$

- 後者のほうが断然強い.
- 標本数 n の影響や $|\mathcal{H}|$ の有無.
- **Take-away:** より強い保証が欲しければ, 実装法を考慮すべし.

ERM の優劣



アルゴリズムによって，$1-\delta$ の確率で約束できることを比較してみる.

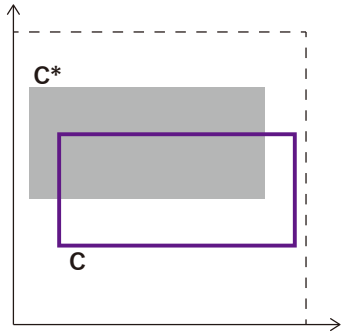
$$R(\widehat{C}_{\text{ERM}}) \leq \sqrt{\frac{\log(2|\mathcal{H}|\delta^{-1})}{2n}}$$

vs.

$$R(\widehat{C}_{\text{fit}}) \leq \frac{4\log(4\delta^{-1})}{n}.$$

- 後者のほうが断然強い.
- 標本数 n の影響や $|\mathcal{H}|$ の有無.
- **Take-away:** より強い保証が欲しければ, 実装法を考慮すべし.

モデル誤差

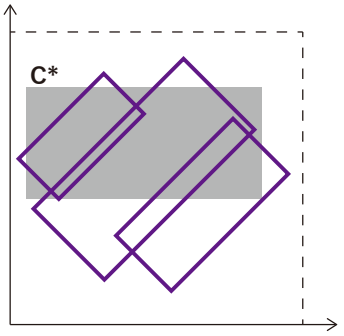


これまでの例では「長方形の学習であること」が既知．これが重要な前提である．

実世界では，普通はここまでの情報はない．

- 向きが違う場合？
- 長方形ではなく，円形だと思われる場合？
- さらに楕円形だと思われる場合？...

モデル誤差



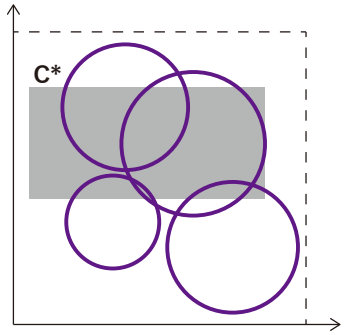
これまでの例では「長方形の学習であること」が既知．これが重要な前提である．

実世界では，普通はここまでの情報はない．

- 向きが違う場合？
- 長方形ではなく，円形だと思われる場合？
- さらに楕円形だと思われる場合？...

「実データで学ぶ人工知能講座」機械学習の基礎　2020 iLDi 研究拠点 データビリティ人材育成教材　47

モデル誤差



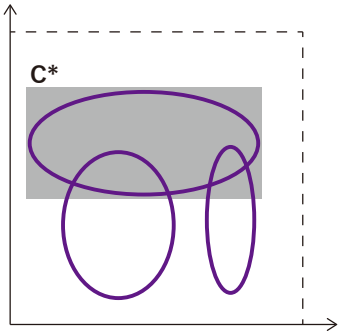
これまでの例では「長方形の学習であること」が既知．これが重要な前提である．

実世界では，普通はここまでの情報はない．

- 向きが違う場合？
- 長方形ではなく，円形だと思われる場合？
- さらに楕円形だと思われる場合？...

「実データで学ぶ人工知能講座」機械学習の基礎　2020 iLDi 研究拠点 データビリティ人材育成教材　47

モデル誤差



これまでの例では「長方形の学習であること」が既知．これが重要な前提である．

実世界では，普通はここまでの情報はない．

- 向きが違う場合？
- 長方形ではなく，円形だと思われる場合？
- さらに楕円形だと思われる場合？...

「実データで学ぶ人工知能講座」機械学習の基礎　2020 iLDi 研究拠点 データビリティ人材育成教材　47

モデル誤差

最良の性能を以下のように定義する．

$$R^* = \inf_h R(h), \quad \text{over all functions } h.$$

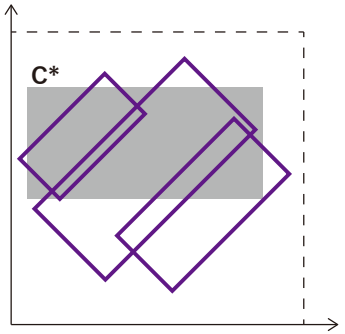
次は誤差を分解していくと，

$$\begin{aligned} R(\widehat{h}) - R^* &= R(\widehat{h}) - \inf_{h\in\mathcal{H}} R(h) + \inf_{h\in\mathcal{H}} R(h) - R^* \\ &\le |R(\widehat{h}) - \inf_{h\in\mathcal{H}} R(h)| + |\inf_{h\in\mathcal{H}} R(h) - R^*| \\ &:= \mathcal{E}(\widehat{h},\mathcal{H}) + \mathcal{A}(\mathcal{H}). \end{aligned}$$

- $\mathcal{E}(\widehat{h},\mathcal{H})$: 推定誤差 (estimation error)．
- $\mathcal{A}(\mathcal{H})$: モデル誤差 (approximation/model error)．

この枠組みを明快に繰り広げた Cucker and Smale (2002) の名著も必読．

モデル誤差



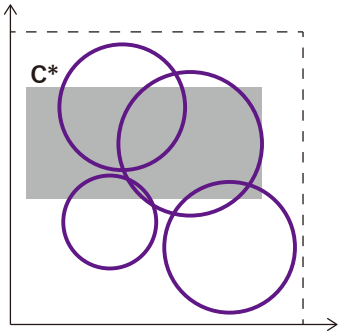
角度を事前に固定した場合：

- $\mathcal{E}(\widehat{h},\mathcal{H})$ = 小
- $\mathcal{A}(\mathcal{H})$ = 大

角度が自由ならば：

- $\mathcal{E}(\widehat{h},\mathcal{H})$ = 大
- $\mathcal{A}(\mathcal{H})$ = ゼロ

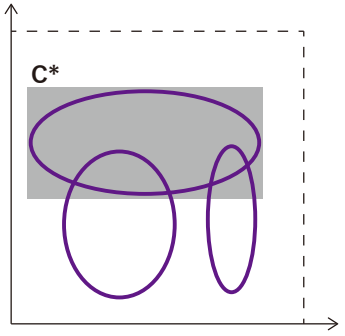
モデル誤差



円形を使う場合は，整合性の高い候補は決めやすいが，フィットがそもそも悪い．

- $\mathcal{E}(\widehat{h},\mathcal{H})$ = 小
- $\mathcal{A}(\mathcal{H})$ = 大

モデル誤差



楕円形なら整合性が高まるが，一番良い候補を選ぶことが少し難しくなる．

- $\mathcal{E}(\widehat{h},\mathcal{H})$ = 大
- $\mathcal{A}(\mathcal{H})$ = 小

モデルの表現力・複雑度が高いほど，より豊富な表現が可能になる．一方，そのなかの最良の候補をデータに基づいて選ぶことが難しくなる．

この原理は *bias-variance tradeoff* と呼ばれることが多い．

まとめ

長方形の学習例

- 平面上の長方形をラベルつきデータから推し量る問題.
- 「正例包絡」という具体的なアルゴリズムを提案した.
- これが ERM の一例で，またかなり良いほうでもあることを示した.
- 任意のデータの分布に対して成り立つ汎化能力の保証を明示した.

評価方法と **PAC** と **ERM**

- 学習則の性能評価と PAC 学習の枠組み.
- 損失とその期待値が汎化能力の指標となる.
- 真の分布を知らない場合でも，性能保証を導出.
- モデル誤差と推定誤差

※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. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data*.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1):1–49.

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

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

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

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Wald, A. (1949). Statistical decision functions. *Annals of Mathematical Statistics*, 20(2):165–205.