機械学習基礎

赤穂 昭太郎

akaho@ism.ac.jp

https://github.com/toddler2009/ml-tutorial

2025 年 4 月 18 日

目次

第 5 章　モデル選択と学習理論　5

5.1　汎化能力と過学習　. 5

5.1.1　モデルの複雑さと汎化誤差 6

5.2　AIC と BIC：情報量規準によるモデル選択 8

5.2.1　一般形と定義 8

5.2.2　AIC の導出：KL ダイバージェンスの近似最小化 8

5.2.3　BIC の導出：ベイズモデル選択の近似　. 9

5.2.4　AIC と BIC の比較と使い分け　. 9

5.3　MDL 原理と BIC との関係　. 10

5.3.1　MDL 原理の基本的な考え方 10

5.3.2　MDL と BIC の関係 10

5.3.3　理論的背景と直観的解釈 11

5.3.4　MDL と AIC との違い 11

5.4　特異モデルと WAIC によるモデル評価　. 12

5.4.1　特異モデルと情報量規準の問題点 12

5.4.2　WAIC の導入と定義 12

5.4.3　WAIC の特徴と利点 13

5.4.4　WAIC と他の規準の比較　. 13

5.5　PAC 学習と VC 次元による汎化誤差評価 14

5.5.1　PAC 学習の基本枠組み　. 14

5.5.2　有限仮説クラスに対する PAC 保証（基本不等式）. 14

5.5.3 無限仮説クラスと VC 次元による一般化 15
5.5.4 代表的なモデルの VC 次元の具体例 16
5.6 正則化と逆問題：一般的枠組みとパラメータ選択 20
5.6.1 不良設定問題と安定性の欠如 20
5.6.2 正則化による安定化 20
5.6.3 正則化パラメータのスケジューリング 21
5.6.4 応用例：確率密度関数の推定と経験分布 21
5.6.5 補足：正則化パラメータのスケジューリング条件と収束の証明スケッチ 22
5.7 機械学習における代表的な正則化手法 23
5.7.1 一般化された ℓ_p 正則化 23
5.7.2 Elastic Net 正則化 24
5.7.3 Fused Lasso（結合ラッソ） 24
5.7.4 Group Lasso（グループラッソ） 24
5.7.5 Total Variation（全変動）正則化 25
5.7.6 その他の正則化手法 25
5.7.7 まとめ 25
5.8 変数選択の方法：正則化ベースと逐次選択法の比較 26
5.8.1 正則化ベースの変数選択（Lasso など） 26
5.8.2 逐次的な変数選択法（Stepwise 法） 27
5.8.3 比較と選択の指針 27
5.9 モデル選択後の推論：Selective Inference の考え方 28
5.9.1 問題の背景と動機 28
5.9.2 Selective Inference の基本的な考え方 28
5.9.3 条件付き推論による補正法（例：LASSO 後の推論） 29
5.9.4 他の方法との比較と応用例 29
5.9.5 まとめ 29

参考文献 31

第 5 章

モデル選択と学習理論

これまでに扱ってきた機械学習モデルは，訓練データに対して最適化された関数近似・識別を行うものであったが，実際の応用においては **未知データに対する性能（汎化能力）** が重要となる．

この章では，機械学習におけるモデルの選択と評価に関する理論的枠組みを扱う．主な内容は以下のとおりである：

- 情報量規準（AIC, BIC, MDL）によるモデル比較
- 統計的学習理論に基づく PAC 学習の枠組み
- 正則化の理論的解釈とその設計原理

5.1　汎化能力と過学習

学習モデルが訓練データに対して高い性能を示していても，未知のデータに対しても同様に良好な性能を示すとは限らない．これが **過学習（overfitting）** の問題である．

一方で，モデルが単純すぎる場合には，データの構造を捉えきれずに **未学習（underfitting）** を引き起こす．

この両者のバランスを取ることが，**モデル選択（model selection）** の本質的な課題である．

5.1.1　モデルの複雑さと汎化誤差

図 5.1 に，学習理論における基本的な枠組みを模式的に示す．ここでは，真に最小化したい損失関数 $L(f)$ が存在し，関数 f を学習によって最適化することで，その最小値の達成を目指す．しかし，$L(f)$ は未知であるため，実際にはその代替として経験損失 $L_n(f)$ を最小化することになる．このとき，$L(f)$ と $L_n(f)$ のずれによって過学習が生じる可能性がある．

$L_n(f)$ を最小にして得られた関数を $\hat{f}_n$，$L(f)$ を最小にする関数を f^* とおくと，以下のような量を評価することにより，過学習の度合いや学習結果の信頼性を推定することができる：

- **汎化ギャップ**：$L(\hat{f}_n) - L(f^*)$
  学習によって得られた関数が，最適な関数と比べてどれほど真の損失を増やすか．
- **過学習バイアス**：$L(\hat{f}_n) - L_n(\hat{f}_n)$
  学習によって最小化したと考えた損失と，実際の損失との間にどれほど差があるか．
- **楽観評価ギャップ**：$L(f^*) - L_n(\hat{f}_n)$
  学習器によって得られた最小の経験損失が，真に最小であるはずの損失と比べてどれほど楽観的に評価されているか．

ただし，これらの指標のうち $L(f)$ や f^* は観測できない量であるため，実際にはなんらかの仮定のもとで，期待値，分散，または分位点の上限などにより評価を行う必要がある．たとえば，AIC や BIC はデータ数 n が十分大きいという漸近的な仮定のもとで期待損失を補正する指標であり，一方，PAC 学習では高い確率で成立する損失の上界（分位点）を評価することを目的とする．

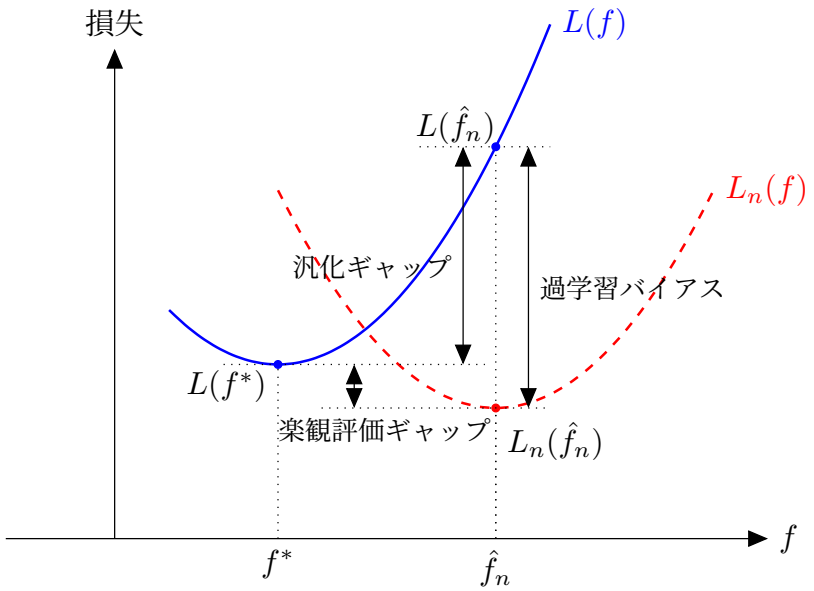

図 5.1　学習理論における損失関数の模式図：真の損失 $L(f)$，経験損失 $L_n(f)$，最適点 $f^*, \hat{f}_n$ と各評価ギャップ

本章で扱う内容の位置づけ

モデル選択と学習理論は，統計的最適性と計算効率の両立を目指す上で中心的な問題である．本章では以下のような多様な観点からこの問題を考察する：

- **情報量規準（AIC, BIC）**：尤度とモデル複雑性のトレードオフ
- **MDL 原理**：データの圧縮長によるモデル評価
- **PAC 学習理論**：確率的保証を伴う汎化能力の定量評価
- **正則化とバイアス・バリアンスのトレードオフ**
- **交差検証の漸近的一致性と分散性の評価**

また，リッジ回帰や LASSO に代表される正則化法に加え，Elastic Net などの代表的な手法も紹介し，それぞれの性質と理論的裏付けについても検討する．

5.2　AIC と BIC：情報量規準によるモデル選択

機械学習や統計的モデリングにおいて，複数の候補モデルから最も適切なモデルを選ぶ際に，尤度とモデルの複雑さをバランスよく評価するための基準が重要となる．ここでは代表的な情報量規準である AIC（Akaike Information Criterion）と BIC（Bayesian Information Criterion）を紹介する．

5.2.1　一般形と定義

モデル $\mathcal{M}_k$（k はモデルの自由度またはパラメータ数）における観測データ $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ に対して，最尤推定値 $\hat{\boldsymbol{\theta}}$ を用いた対数尤度を $\ell(\hat{\boldsymbol{\theta}})$ とすると，以下のように定義される：

- **AIC**（赤池情報量規準）：

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2k$$

- **BIC**（ベイズ情報量規準）：

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + k \log n$$

ここで：

- $\ell(\hat{\boldsymbol{\theta}})$：最尤推定値による対数尤度
- k：モデルのパラメータ数
- n：サンプルサイズ

5.2.2　AIC の導出：KL ダイバージェンスの近似最小化

AIC は，真の分布 $g(x)$ に対してモデル $f(x \mid \theta)$ を近似させるとき，Kullback – Leibler（KL）ダイバージェンス：

$$D_{\mathrm{KL}}(g \parallel f_\theta) = \int g(x) \log \frac{g(x)}{f(x \mid \theta)}\, dx$$

を最小化することを目指した近似推定量である．

データから得られる最尤推定値 $\hat{\theta}$ に対し，平均化された尤度の期待値を補正することで，以下の近似が得られる（赤池, 1974）：

$$\mathbb{E}[-2\ell(\hat{\theta})] \approx -2\ell(\hat{\theta}) + 2k$$

この補正項 $2k$ がモデルの複雑さを評価するペナルティとなる．

5.2.3　BIC の導出：ベイズモデル選択の近似

BIC はベイズ的なモデル比較の枠組みに基づいており，モデル $\mathcal{M}_k$ における周辺尤度（モデル証拠）：

$$p(\boldsymbol{x} \mid \mathcal{M}_k) = \int p(\boldsymbol{x} \mid \theta_k, \mathcal{M}_k)\, p(\theta_k \mid \mathcal{M}_k)\, d\theta_k$$

の対数をラプラス近似により評価した結果として得られる：

$$\log p(\boldsymbol{x} \mid \mathcal{M}_k) \approx \ell(\hat{\boldsymbol{\theta}}) - \frac{k}{2} \log n + \text{const.}$$

これを -2 倍して，AIC と同様の形式にそろえると BIC が得られる．

5.2.4　AIC と BIC の比較と使い分け

- **AIC** は汎化誤差の期待値最小化に基づいており，**予測精度重視のモデル選択**に適している．
- **BIC** はベイズ的なモデルの周辺尤度に基づいており，**真のモデルを識別する能力（consistent model selection）**に優れる．
- AIC はパラメータ数に比例したペナルティ，BIC は $\log n$ に比例したペナルティを持つため，**サンプルサイズが大きくなるほど BIC の方が複雑なモデルに厳しくなる**．

■使い分けの指針

- **予測精度重視：** AIC（汎化性能の近似）

- **モデル同定重視：** BIC（真のモデルへの収束性）
- **サンプルサイズが小さい場合：** AIC（柔軟）
- **サンプルサイズが大きい場合：** BIC（過剰適合の抑制）

5.3　MDL 原理と BIC との関係

MDL（Minimum Description Length）原理は，機械学習や統計モデリングにおいて，データを最も効率的に記述するモデルを選ぶという情報理論的な立場に基づくモデル選択の方法である（Rissanen, 1978）.

この原理では，モデルの選択を **データの圧縮**という観点から考える.

5.3.1　MDL 原理の基本的な考え方

観測データ $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ を与えられたモデルクラス $\mathcal{M} = \{\mathcal{M}_k\}$ により記述する際,

$$\text{総記述長} = \underbrace{\text{モデルの複雑さ}}_{\text{モデルの符号長}} + \underbrace{\text{データの誤差}}_{\text{データの符号長}}$$

を最小化するモデルを選ぶというのが MDL の基本原理である.

より具体的には，パラメータ θ をもつ確率モデル $p(\boldsymbol{x} \mid \theta)$ に対し，次のような「符号長」の近似を用いる：

$$L(\boldsymbol{x}, \hat{\theta}) = -\log p(\boldsymbol{x} \mid \hat{\theta}) + \frac{k}{2} \log n + \text{定数}$$

この表現は，第一項がデータの対数尤度（誤差項），第二項がモデルの複雑さ（パラメータ空間のボリュームに対するペナルティ）を表す.

5.3.2　MDL と BIC の関係

この表式は，**BIC（Bayesian Information Criterion）** の導出と同じ構造を持っている．すなわち，ラプラス近似を用いてベイズモデルの周辺尤度：

$$\log p(\boldsymbol{x} \mid \mathcal{M}_k) \approx \ell(\hat{\theta}) - \frac{k}{2}\log n + \text{定数}$$

を最大化することは，以下の BIC を最小化することと等価である：

$$\mathrm{BIC} = -2\ell(\hat{\theta}) + k\log n$$

よって，**MDL 原理と BIC は数学的に等価な基準**であり，両者は「より短く符号化できるモデルがより良いモデルである」という思想のもとで一致する．

5.3.3 理論的背景と直観的解釈

MDL の理論的背景には，シャノンの情報理論における**符号長と確率の関係**：

$$L(x) \approx -\log p(x)$$

がある．すなわち，**確率が高いほど短く符号化できる**という原理をベースに，データとモデルの情報量をトータルで最小化する．

このように，MDL は統計的推論を情報理論的な観点から再解釈したものであり，モデル選択とデータ圧縮の間に深い関係を与える．

5.3.4 MDL と AIC との違い

- AIC は **予測性能（汎化誤差）** の近似最小化に基づく．
- BIC および MDL は **モデルの同定（一貫性）** に基づく．
- サンプルサイズが大きくなると，MDL/BIC はより強く複雑なモデルにペナルティを与える傾向がある．

■まとめ

- MDL 原理は，モデル選択を「データとモデルの最小符号化長」に帰着させる情報理論的アプローチ．
- 数学的には BIC と一致し，ラプラス近似によって導出される．
- モデルの選択を「圧縮効率」という観点から再解釈する枠組みを与える．

5.4　特異モデルと WAIC によるモデル評価

5.4.1　特異モデルと情報量規準の問題点

AIC や BIC は，正則な統計モデル（regular model）を前提として導出されている．すなわち，以下の仮定が必要である：

- 真の分布がモデルに含まれている（well-specified）.
- パラメータ空間において最尤推定量が一意に存在し，**Fisher 情報行列が正則（非退化）**.
- 推定量が漸近的に正規分布に従う（漸近正規性）.

しかし，実際の機械学習モデルでは以下のような **特異モデル（singular model）** が多数存在する：

- 隠れ変数を持つ混合モデル（例：混合ガウス分布）
- ニューラルネットワーク
- 階層ベイズモデル
- 深層生成モデル（例：VAE）

これらのモデルでは，**Fisher 情報行列が退化（rank 欠落）**し，最尤推定量の漸近正規性が成り立たず，AIC や BIC の理論的正当性が崩れる．

5.4.2　WAIC の導入と定義

この問題を解決するために，**Watanabe（2010）** により提案されたのが WAIC（Widely Applicable Information Criterion）である．

WAIC は，ベイズ推論の枠組みにおいて，**周辺化された予測分布の汎化誤差**を推定する情報量規準であり，特異モデルにも適用可能である．

■WAIC の定義　観測データ $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ に対して，事後分布 $p(\theta \mid \boldsymbol{x})$ が得られているとき，WAIC は以下のように定義される：

$$\mathrm{WAIC} = -2\left(\sum_{i=1}^{n}\log \mathbb{E}_\theta[p(x_i \mid \theta)] - \sum_{i=1}^{n}\mathrm{Var}_\theta[\log p(x_i \mid \theta)]\right)$$

ここで：

- $\mathbb{E}_\theta[\cdot]$ は事後分布に対する期待値（ベイズ平均）
- $\mathrm{Var}_\theta[\cdot]$ は事後分布に対する分散
- 第一項はベイズ予測誤差，第二項は**複雑さの補正項（effective number of parameters）**として解釈される

5.4.3 WAIC の特徴と利点

- 特異モデル（非正則モデル）においても有効な汎化誤差の推定量である．
- クロスバリデーションの近似量と一致することが知られている（Watanabe, 2010）．
- モデルの予測分布を基準とするため，AIC/BIC よりも柔軟性が高い．
- ベイズ推論（MCMC 等）によって事後分布が得られていれば，WAIC は数値的に容易に計算できる．

5.4.4 WAIC と他の規準の比較

- AIC や BIC はモデルが正則であることを仮定して導出される．
- WAIC は **正則性を仮定せず，任意のベイズモデルに適用可能**．
- 特異モデルにおいては，AIC/BIC は過度に楽観的な汎化誤差を与える可能性があるのに対し，WAIC はより堅牢な評価を与える．

■まとめ

- 特異モデルでは，最尤推定の性質が壊れるため，AIC/BIC は信頼できない．
- WAIC はベイズ的な枠組みにより，汎化誤差の期待値を予測分布ベースで評価する．

- 汎化性能を適切に評価したい場合，特に **複雑な生成モデルや深層モデル**においては WAIC が有力な選択肢となる．

5.5　PAC 学習と VC 次元による汎化誤差評価

5.5.1　PAC 学習の基本枠組み

PAC（Probably Approximately Correct）学習は，Valiant（1984）によって提案された，**誤差と確率の二重の保証**をもつ学習の枠組みである．

学習アルゴリズムが誤差 ε 以下の予測を，確率 $1-\delta$ 以上で達成できるならば，そのアルゴリズムは (ε, δ)-PAC であるという．

■設定

- 入力空間 $\mathcal{X}$ とラベル空間 $\mathcal{Y} = \{0, 1\}$
- 未知の分布 $\mathcal{D}$ に従ってデータ $(x_i, y_i) \sim \mathcal{D}$ を取得
- 仮説集合（モデルクラス）$\mathcal{H}$ 上で学習を行い，$h \in \mathcal{H}$ を選ぶ

■目的　十分なサンプル数 n があれば，任意の $h \in \mathcal{H}$ に対して次が成り立つ：

$$\mathbb{P}_{\mathcal{D}^n}\left[|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathrm{emp}}(h)| > \varepsilon\right] \leq \delta$$

ここで：

- $\mathrm{err}_{\mathcal{D}}(h)$ は真の汎化誤差（test error）
- $\mathrm{err}_{\mathrm{emp}}(h)$ は経験誤差（training error）

5.5.2　有限仮説クラスに対する PAC 保証（基本不等式）

仮説集合 $\mathcal{H}$ が有限集合である場合，$\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$ のとき，1 つの h に対して Hoeffding の不等式を用いると：

$$\mathbb{P}\left[|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathrm{emp}}(h)| > \varepsilon\right] \leq 2\exp(-2n\varepsilon^2)$$

仮説空間全体に対して合成すると（合成確率の加法）：

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathrm{emp}}(h)| > \varepsilon\right] \leq 2K \exp(-2n\varepsilon^2)$$

これが δ 以下になるためには：

$$n \geq \frac{1}{2\varepsilon^2}\left(\log(2K) + \log\frac{1}{\delta}\right)$$

すなわち，**PAC 的に学習可能であるためには，仮説空間のサイズ K に対数的に依存する数のサンプルが必要**である．

5.5.3 無限仮説クラスと VC 次元による一般化

$\mathcal{H}$ が無限集合（例：線形分類器など）である場合には，単純に仮説数を数えることができない．その代わりに，**仮説クラスの複雑さ**を測るために VC 次元（Vapnik – Chervonenkis dimension）が導入される．

■定義（VC 次元） 仮説クラス $\mathcal{H}$ の VC 次元 d_{VC} とは，$\mathcal{X}$ 上の任意の d_{VC} 個の点を任意のラベルで完全に分離（shatter）できる最大の d のこと．

$$d_{\mathrm{VC}} := \max\left\{d \,\middle|\, \exists x_1, \ldots, x_d \in \mathcal{X}, \text{すべての } \{0,1\}^d \text{ に対応する } h \in \mathcal{H} \text{ が存在}\right\}$$

■Sauer の補題と帰着 サンプルサイズ n に対して，$\mathcal{H}$ が生み出すラベルのパターンの数は高々：

$$|\mathcal{H}|_{\mathrm{restricted}} \leq \sum_{i=0}^{d_{\mathrm{VC}}} \binom{n}{i} \leq \left(\frac{en}{d_{\mathrm{VC}}}\right)^{d_{\mathrm{VC}}}$$

したがって，有限仮説集合における PAC の議論と同様に，仮説空間を「効果的なサイズ（ラベルの区別可能な数）」に制限して議論できる．

結果として，次のような汎化誤差境界が得られる：

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathrm{emp}}(h)| > \varepsilon\right] \leq c\left(\frac{n}{d_{\mathrm{VC}}}\right)^{d_{\mathrm{VC}}} \exp(-n\varepsilon^2)$$

ここで $c > 0$ は定数．この不等式から，必要なサンプルサイズ n は d_{VC} に依存して増加する．

まとめ

- PAC 学習は「たいてい，ある程度正しく予測できる」ことを確率的に保証する学習の枠組み．
- 有限仮説クラスでは，サイズ K に対して $O(\log K)$ のサンプル数で汎化保証が可能．
- 無限仮説クラスでは，**VC 次元**によって「効果的なサイズ」に帰着し，PAC 保証を導出できる．

5.5.4　代表的なモデルの VC 次元の具体例

仮説クラスの複雑さを測る指標として **VC 次元（Vapnik – Chervonenkis 次元）** は，PAC 学習における汎化能力の理論的評価に重要である．ここでは，代表的な学習モデルの VC 次元を具体的に紹介する．

線形識別器（線形分類器）

D 次元の入力空間 $\mathbb{R}^D$ における線形識別器とは，以下のような仮説クラス：

$$\mathcal{H}_{\mathrm{linear}} = \left\{ x \mapsto \mathrm{sign}(\boldsymbol{w}^\top x + b) \mid \boldsymbol{w} \in \mathbb{R}^D, b \in \mathbb{R} \right\}$$

このクラスの VC 次元は次のように与えられる：

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{linear}}) = D + 1$$

■理由（直感的な説明）　$D + 1$ 個の点を適切に配置すれば，全てのラベルパターンに対応する線形識別超平面が存在するが，$D + 2$ 個では必ずしも分離可能でない（Radon's 定理などが根拠）．

SVM（ハードマージン）

SVM は線形識別器と同様の仮説クラスを持つが，最大マージンを学習目的とする点が異なる．VC 次元は基本的に線形識別器と同様：

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{SVM}}) = D + 1$$

ただし，**マージンを陽に制約**した場合（例えばマージンが γ 以上であることを仮定）には，VC 次元はより小さく評価できる．

■マージン付き分類器の VC 次元（Bartlett, 1998）

$$\mathrm{VCdim} \leq \min\left(\frac{R^2}{\gamma^2}, D\right) + 1$$

ここで：

- R はデータの半径（$\|x_i\| \leq R$）
- γ はマージン幅

このように，SVM においては **マージンが大きいほど汎化性能が良くなる** という理論的背景がある．

多項式識別器（次数 k の多項式）

次数 k の多項式識別器は，以下のような形の関数を考える：

$$\mathcal{H}_{\mathrm{poly},k} = \left\{ x \mapsto \mathrm{sign}(P_k(x)) \mid P_k(x) \text{ は次数 } \leq k \text{ の多項式} \right\}$$

このとき，仮説空間の次元は $\binom{D+k}{k}$（モノミアルの数）となり，VC 次元もおおよそこれに比例する：

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{poly},k}) = \Theta\left(\binom{D+k}{k}\right)$$

すなわち，**多項式次数が増えると急激に複雑さが増大する**．

k-NN 分類器（固定 k）

k-近傍法（k-NN）は，距離に基づいてラベルを決定するノンパラメトリック法である．このような識別器の VC 次元は，通常は無限である：

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{kNN}}) = \infty$$

ただし，データ数 n を固定した場合は有限となる．これは，k-NN がすべてのラベルに柔軟に対応できてしまうため．

決定木（最大深さ h の木）

最大深さ h の決定木モデルでは，仮説空間の表現能力が制限されるため，VC 次元も深さに依存する：

$$\mathrm{VCdim} \leq O(2^h)$$

これは，分割によって実現できる領域の数が指数的に増えることに由来する．すなわち，**木の深さはモデルの複雑さ（過学習のしやすさ）と密接に関係**する．

まとめ

モデル	VC 次元
線形識別器（$\mathbb{R}^D$）	$D+1$
SVM（ハードマージン）	$D+1$（または $\frac{R^2}{\gamma^2}+1$）
多項式識別器（次数 k）	$\Theta\left(\binom{D+k}{k}\right)$
k-NN 分類器	∞（柔軟すぎる）
決定木（深さ h）	$O(2^h)$

例：$\mathrm{sign}(\sin(ax))$ の VC 次元は無限大

関数クラス $\mathcal{H} = \{x \mapsto \mathrm{sign}(\sin(ax)) \mid a \in \mathbb{R}\}$ に対して，その VC 次元が無限大であることを示す．

■主張　$\mathcal{H}$ は任意個の点集合をシャッターできる．すなわち，

$$\mathrm{VCdim}(\mathcal{H}) = \infty$$

■証明のアイディア 任意の個数 n の点 $x_1, \ldots, x_n$ をとって，それらの各点に対する任意のラベル割り当てを実現する $a \in \mathbb{R}$ を構成する．

例えば，点を次のように選ぶ[*1]：

$$x_i = 2^i, \quad i = 1, 2, \ldots, n$$

任意のラベルパターン $(y_1, \ldots, y_n) \in \{0, 1\}^n$ に対し，次のような a を選ぶ：

$$a = -\pi * (0.y_1 y_2 \ldots y_n 1)_2$$

なおカッコ内は 2 進数表現を表す．

このように選ぶと，a を i bit 左シフトする (x_i をかける) ことで，y_i が 0 のとき

$$\pi < a x_i \pmod{2\pi} < 2\pi$$

y_i が 1 のとき

$$0 < a x_i \pmod{2\pi} < \pi$$

となるので，$\mathrm{sign}(\sin(a x_i)) = y_i$ となる．

従って，任意のラベルパターンに対して対応する a を構成できるので，どんな n に対しても n 点をシャッター可能：

$$\Rightarrow \quad \mathrm{VCdim}(\mathcal{H}) = \infty$$

■補足：有限 VC 次元モデルとの対比 この例は，**非常に単純な関数形式でも，周期性とパラメータの自由度によってモデルが任意の複雑さを持つ**ことを示している．これは，機械学習におけるモデル制御の重要性を物語るものである．

[*1] x_i が互いに異なる任意の値の時も a をうまく選べばシャッターできると予想するが著者の知る限り証明は得られていない．

5.6　正則化と逆問題：一般的枠組みとパラメータ選択

機械学習における多くの学習問題は，関数のパラメータを観測データから復元するという **逆問題（inverse problem）**の形式で定式化できる．

特に，作用素方程式

$$Af = F$$

の形で，未知の関数 f を，既知の線形（または非線形）作用素 A と観測された右辺 F から推定するという問題は，多くの実用的な推論に共通する．

5.6.1　不良設定問題と安定性の欠如

右辺 F が理想的な値ではなく，何らかのノイズを含む近似値 F^δ のみが得られるとする：

$$\|F^\delta - F\| \leq \delta$$

このとき，単に $A^{-1}F^\delta$ を計算すると，ノイズによって大きな誤差が増幅され，f の推定が不安定になる．このような問題は **不良設定問題（ill-posed problem）**と呼ばれ，以下の性質が欠如している：

- 解の存在性
- 解の一意性
- データに対する解の連続依存性（安定性）

5.6.2　正則化による安定化

このような不安定性を回避するために，**正則化（regularization）**という手法が用いられる．これは，目的関数に制約やペナルティを導入することで，解の安定性を高めるものである．

代表的な方法は，Tikhonov 型正則化であり，次のような最適化問題として定式化される：

$$\min_f \left\{ \|Af - F^\delta\|^2 + \lambda \mathcal{R}(f) \right\}$$

ここで：

- $\mathcal{R}(f)$ は正則化項（例：ノルム，変動，スパース性など）
- $\lambda > 0$ は正則化パラメータで，安定性と精度のトレードオフを制御する

5.6.3　正則化パラメータのスケジューリング

正則化パラメータ λ の選び方は，F^δ がどれだけ真の F に近いか（すなわちノイズレベル δ）によって決定されるべきである．

理想的には，$\delta \to 0$ に対して $\lambda = \lambda(\delta)$ も適切に収束させることで，以下の性質を満たす：

$$\lambda(\delta) \to 0 \quad \text{かつ} \quad \frac{\delta^2}{\lambda(\delta)} \to 0 \quad (\delta \to 0)$$

この条件の下で，正則化解 f^λ は真の解 f に収束することが知られている．

5.6.4　応用例：確率密度関数の推定と経験分布

一つの典型例として，確率密度関数 f を推定したいとする．ここで：

- A：密度 f から分布関数 F への積分作用素：

$$(Af)(x) = \int_{-\infty}^{x} f(t)\, dt = F(x)$$

- F^δ：観測に基づく近似，すなわち経験分布関数 F_n

このとき，F_n は F に収束するが，有限標本ではノイズを含む近似にすぎない．したがって，f を単に F_n を微分して求めると不安定である．

このようなときに，例えば **カーネル密度推定**[*2]や **正則化項付き推定（例：変動最小化，ノルム制約など）**を用いることで，より安定な推定が可能になる．

■**結論**　このように，逆問題においてはノイズの存在によって推定が不安定になるため，正則化によって解の構造を制約することが重要である．正則化パラメータはノイズの大きさと精度のバランスを取るように調整される必要がある．

5.6.5　補足：正則化パラメータのスケジューリング条件と収束の証明スケッチ

線形作用素 $A : \mathcal{X} \to \mathcal{Y}$（ヒルベルト空間間の有界線形作用素）と，観測値 $F^\delta \in \mathcal{Y}$ がノイズ付きであるとする：

$$\|F^\delta - F\| \leq \delta$$

ここで，未知関数 $f^* \in \mathcal{X}$ は $Af^* = F$ を満たすと仮定する．これに対して，次のような Tikhonov 正則化問題を考える：

$$f^\lambda := \arg\min_{f \in \mathcal{X}} \left\{ \|Af - F^\delta\|^2 + \lambda \|f\|^2 \right\}$$

このとき，f^λ が f^* に近づくための条件として以下のようなスケジューリングが知られている：

$$\lambda = \lambda(\delta) \to 0, \quad \text{かつ} \quad \frac{\delta^2}{\lambda(\delta)} \to 0 \quad (\delta \to 0)$$

■**証明スケッチ**　正則化解 f^λ は，正則化項により安定化されており，次の評価が得られる（定常性条件と変分不等式に基づく）：

$$\|f^\lambda - f^*\| \leq C_1 \cdot \frac{\delta}{\sqrt{\lambda}} + C_2 \cdot \sqrt{\lambda}$$

ここで：

[*2] カーネル密度推定も正則化問題として定式化できる [17]

- 第一項はデータのノイズが作用素の逆作用によって増幅される誤差項
- 第二項は正則化によって発生するバイアス項（過剰なスムージング）

この上界がゼロに収束するには：

$$\frac{\delta}{\sqrt{\lambda}} \to 0, \quad \sqrt{\lambda} \to 0 \quad \Rightarrow \quad \lambda \to 0, \quad \frac{\delta^2}{\lambda} \to 0$$

よって，これが正則化解の一致性（consistency）を保証するスケジューリング条件である．

■結論

$$\boxed{\lambda(\delta) \to 0, \quad \frac{\delta^2}{\lambda(\delta)} \to 0 \quad \text{ならば} \quad \|f^{\lambda} - f^*\| \to 0}$$

このようなパラメータ選択は，正則化における**バイアスと分散のトレードオフ**に基づいて導かれる自然な戦略である．

5.7　機械学習における代表的な正則化手法

正則化は，学習において解の一意性や安定性，または構造的な性質（スパース性，平滑性など）を誘導するために導入される．以下では，代表的な正則化手法を定義とともに簡潔にまとめる．

5.7.1　一般化された ℓ_p 正則化

パラメータベクトル $\boldsymbol{w} \in \mathbb{R}^d$ に対して，以下のようなノルムを導入する：

$$\mathcal{R}_p(\boldsymbol{w}) = \|\boldsymbol{w}\|_p^p = \sum_{j=1}^{d} |w_j|^p$$

- $p = 2$：リッジ正則化（滑らかな収束）
- $p = 1$：Lasso 正則化（スパース性を誘導）
- $p < 1$：より強いスパース性（非凸最適化，解の計算が困難）
- $p > 2$：ノルムが大きな成分を強く罰する（通常あまり用いられない）

■性質の分岐

- $p \geq 1$：凸最適化となり，計算が安定
- $p < 1$：非凸となるが，スパース性の促進がより強力

5.7.2 Elastic Net 正則化

Lasso とリッジの両方の性質を兼ね備えた正則化．以下のように定義される：

$$\mathcal{R}_{\mathrm{EN}}(\boldsymbol{w}) = \alpha\|\boldsymbol{w}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{w}\|_2^2, \quad \alpha \in [0,1]$$

- Lasso のスパース性と，リッジの安定性を両立
- 多重共線性のあるデータや，高次元データに有効

5.7.3 Fused Lasso（結合ラッソ）

通常のラッソに加え，隣接成分間の差を罰することで，**分段定数的な構造**を促進する：

$$\mathcal{R}_{\mathrm{Fused}}(\boldsymbol{w}) = \lambda_1 \sum_{j=1}^{d} |w_j| + \lambda_2 \sum_{j=1}^{d-1} |w_{j+1} - w_j|$$

- 時系列や空間データなど，順序構造のある変数に対して有効
- 変数のスパース性と滑らかさの両方を表現可能

5.7.4 Group Lasso（グループラッソ）

パラメータベクトルをあらかじめ定められたグループ $\{G_1, \ldots, G_m\}$ に分け，それぞれのグループ単位でスパース性を誘導：

$$\mathcal{R}_{\mathrm{Group}}(\boldsymbol{w}) = \sum_{g=1}^{m} \lambda_g \|\boldsymbol{w}_{G_g}\|_2$$

- グループ単位で全体を選択・除外（変数選択）
- 各グループ内ではスパースでなくてもよい

5.7.5 Total Variation（全変動）正則化

連続的な関数 f に対して，変動（勾配）の合計を罰する正則化：

$$\mathcal{R}_{\mathrm{TV}}(f) = \int |f'(x)| dx$$

離散化すると：

$$\mathcal{R}_{\mathrm{TV}}(\boldsymbol{w}) = \sum_{j=1}^{d-1} |w_{j+1} - w_j|$$

- 信号や画像処理においてエッジを保ったままノイズを除去する
- 平滑性よりも「分段定数性」を強調する

5.7.6 その他の正則化手法

- **核ノルム（trace norm）**：行列の低ランク化を誘導．推薦システムなどで利用

$$\mathcal{R}(\mathbf{W}) = \sum_i \sigma_i(\mathbf{W}) \quad \text{（特異値の和）}$$

- **エントロピー正則化**：確率分布において滑らかさを促進するために用いられる

$$\mathcal{R}(p) = \sum_i p_i \log p_i$$

- **非凸正則化（SCAD, MCP など）**：スパース性と推定バイアスのバランスを取る目的で設計された非凸関数を使用する

5.7.7 まとめ

以下のように，正則化にはさまざまな形式があり，目的に応じて適切な選択が重要である：

正則化手法	主な特徴	適用例
ℓ_1（Lasso）	スパース性	変数選択
ℓ_2（Ridge）	平滑性，安定性	多重共線性
Elastic Net	スパース＋安定性	高次元回帰
Fused Lasso	スパース＋構造的連続性	時系列，画像
Group Lasso	グループ選択	分類器の選択
Total Variation	分段定数性	信号処理，画像復元
Trace Norm	低ランク誘導	行列補完，推薦
エントロピー	分布の滑らかさ	確率モデル
非凸（SCAD 等）	小さいバイアス，強いスパース性	高精度推定

5.8　変数選択の方法：正則化ベースと逐次選択法の比較

回帰や分類などのモデルにおいて，説明変数の数が多くなると，**どの変数をモデルに含めるべきか**という変数選択の問題が生じる．

この問題に対しては，主に以下のような 2 つのアプローチが存在する：

1. **正則化ベースの方法**（例：Lasso）
2. **逐次的な変数選択法（stepwise 法）**

5.8.1　正則化ベースの変数選択（Lasso など）

正則化項に ℓ_1 ノルムなどを加えることで，回帰係数をスパースにし，自然に変数選択を行う手法．Lasso は以下のように定式化される：

$$\min_{\boldsymbol{w}} \left\{ \|X\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{w}\|_1 \right\}$$

■特徴

- 変数選択と係数推定を同時に行う．
- 高次元データ（$p > n$）にも対応可能．

- 連続的かつ安定的な変数選択経路を持つ（正則化パス）.
- アルゴリズム的に効率的な最適化が可能（例：座標降下法）.

5.8.2　逐次的な変数選択法（Stepwise 法）

伝統的な統計モデリングでよく用いられる方法で，変数を一つずつ加えたり削除したりしてモデルを構築する．代表的には以下の 2 つがある：

- **前進選択法（forward selection）**：変数を一つずつ追加し，モデルの評価指標（AIC, BIC など）を改善する方向に進む.
- **後退除去法（backward elimination）**：全変数から出発して，寄与の少ない変数を一つずつ削除する.

■特徴

- 計算は比較的単純で，モデルの構造が直感的に理解しやすい.
- モデル選択の判断には AIC や BIC，p 値などが用いられる.
- 変数間の相関に対して脆弱で，多重共線性があると不安定.
- 高次元データ（$p > n$）には適用が困難.

5.8.3　比較と選択の指針

特徴	正則化ベース（Lasso など）	Stepwise 法
変数選択の仕組み	正則化による自動選択	手続き的に加除
計算方法	凸最適化問題	繰り返しのモデル比較
安定性	一定の連続性あり	小さな変更でモデル激変の可能性
多重共線性への対応	一部選択（ただしバイアスあり）	不安定になりやすい
高次元データ	対応可（$p > n$）	非対応
可視化・解釈性	正則化パスなど可視化容易	モデル構造の解釈は比較的容易

■結論　近年では，高次元データや計算効率の観点から，**正則化ベースの方法（特に Lasso やその拡張）**が主流となっている．しかし，Stepwise 法はモデル構造の可視性や，統計的仮説検定と結びつけやすいという点で，**小規模で説明重視のモデル構築**において今も有用である．

目的やデータ構造に応じて，両者を適切に使い分けることが重要である．

5.9　モデル選択後の推論：Selective Inference の考え方

モデル選択に基づいて得られた推定値や統計量に対して，標準的な方法で信頼区間や p 値を計算することは，**選択バイアス（selection bias）**により過大評価となる可能性がある．

この問題に対処する枠組みとして **Selective Inference（選択的推論）** が注目されている．

5.9.1　問題の背景と動機

モデル選択（変数選択，クラスタ数決定，スパース回帰など）を行ったあと，その選択結果に基づいて通常の方法で推定や検定を行うと，**モデル選択の不確かさを考慮していないため，過剰な信頼度を持つことになる**．

- 特に，変数選択後の係数推定において，選択された変数の効果を通常の t 検定などで評価すると，p 値が過小評価されやすい．

5.9.2　Selective Inference の基本的な考え方

Selective Inference では，「**選択されたという条件のもとでの分布**」に基づいて推論を行う．

- 観測された統計量 T の分布を，モデル選択イベント A のもとで条件付き分布として評価する：

$$\mathbb{P}(T \in C \mid \text{モデル選択結果} \in A)$$

- この条件付き分布を用いて，正しい信頼区間や p 値を構成する．

5.9.3　条件付き推論による補正法（例：LASSO 後の推論）

Tibshirani ら（2016）による枠組みでは，LASSO によって選択された変数集合に対して，以下の手順で推論が行われる：

- LASSO の選択イベントが **線形不等式の集合（多面体領域）**で表現できる
- 選択された変数の線形統計量（例：$\hat{\beta}_j$）の分布を，この多面体領域に条件づけて評価する
- この条件付き分布は truncated normal（切断正規分布）になることが多く，これに基づいた信頼区間や p 値が計算できる

■例：条件付き p 値の計算　ある選択された変数の統計量 T が標準正規分布に従い，選択イベントが $T \in [a, b]$ に対応する場合：

$$p = \mathbb{P}_{H_0}(|T| \geq |t_{\text{obs}}| \mid T \in [a, b]) = \frac{\mathbb{P}_{H_0}(|T| \geq |t_{\text{obs}}|, T \in [a, b])}{\mathbb{P}_{H_0}(T \in [a, b])}$$

このような形で，選択イベントに条件づけた p 値や信頼区間が構成される．

5.9.4　他の方法との比較と応用例

- **Post-selection inference**：LASSO やステップワイズ法などによる変数選択後の推論
- **データ分割法（data splitting）**：データを選択用と推定用に分割することで選択バイアスを回避（ただし分割のばらつきの問題あり）
- **Selective CLT**：選択された統計量に対して漸近的な正規性を仮定して補正

5.9.5　まとめ

- モデル選択後の推論には，選択によるバイアスを補正する必要がある．
- Selective Inference は，**選択イベントに条件づけた分布に基づいて推論**を行う

ことで，正当な p 値や信頼区間を提供する．

- LASSO など一部の選択手法では，選択イベントが線形不等式で記述可能であるため，理論的・実用的な枠組みが整備されている．

参考文献

[1] James, G., Witten, D., Hastie, T., Tibshirani, R. (2023) An Introduction to Statistical Learning, Second Edition, Springer. https://www.statlearning.com

[2] Bishop, C.M. (2026) Pattern recognition and machine learning, Springer (ビショップ：パターン認識と機械学習（上下），丸善)

[3] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23 (pp. 35-50). Springer Berlin Heidelberg.

[4] 萩原克幸 (2022)，入門 統計的回帰とモデル選択，共立出版.

[5] 赤穂昭太郎 (2008), カーネル多変量解析，岩波書店.

[6] Wahba, G. (1990). Spline models for observational data. Society for industrial and applied mathematics.

[7] 金森敬文, 竹之内高志, 村田昇 (2009). パターン認識 (R で学ぶデータサイエンス 5) 共立出版.

[8] Cover, T., Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[9] 青嶋誠，矢田和善 (2019)，高次元の統計学，共立出版

[10] 小西貞則. (2010). 多変量解析入門: 線形から非線形へ. 岩波書店.

[11] Hyvärinen, A. (2013). Independent component analysis: recent advances. Philosophical Transactions of the Royal Society A: Mathematical, Physical

and Engineering Sciences, 371(1984), 20110534.

[12] 赤穂昭太郎 (2018). ガウス過程回帰の基礎. システム/制御/情報, 62(10), 390-395.

[13] Domingos, P., Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29, 103-130.

[14] Yedidia, J. S., Freeman, W. T., Weiss, Y. (2001). Bethe free energy, Kikuchi approximations, and belief propagation algorithms. Advances in neural information processing systems, 13(24).

[15] Srinivas, N., Krause, A., Kakade, S. M., Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995.

[16] Watanabe, S., Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of machine learning research, 11(12).

[17] Vapnik, V.N. (1998). Statistical Learning Theory, Wiley.