カーネル多変量解析
第2章
カーネル多変量解析の仕組み（2）

サ ミンソン

２．３確率モデルからの導入

- 多変量解析の目的・・・与えられたデータに基づいて、背後にある構造を推論

- ただし・・・データにノイズがのる、関数のうちの有限個の点でしか関数値が与えられない・・・などの問題

↓

データの生成過程を確率分布を用いてモデル化する

(a)線形モデルのベイズ推論

・まず、$y=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$(式(1.1)) による関数近似の確率モデルを考える。

出力 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ に、ランダムノイズ ε がのった、

$y=f(\boldsymbol{x})+\varepsilon$ (2.28) が出力として与えられる。

ε が独立な正規分布であるとすると、x、fが与えられたもとで、分散を σ^2 として、

・yの条件付き確率精度

$$p(y \mid \boldsymbol{x}; f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-f(\boldsymbol{x}))^2}{2\sigma^2}\right) \qquad (2.29)$$

続き

- 線形モデル $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$

ーパラメータwを決めることとfを決めることは等価

・wの各成分が独立に平均0、

分散が $1/\lambda$ である正規分布から生成されたとする。

wの分布を

$$p(\boldsymbol{w}) = \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-\frac{\lambda}{2}\|\boldsymbol{w}\|^2\right) \qquad (2.30)$$

とする。

サンプルの生成過程

- まず、p(w)に従ってパラメータwがランダムに決められる。
- wで決まる関数f(x)をn個の点で計算し、(2.29)の分布に従ってサンプル出力yが観測される。

- パラメータとデータの同時確率分布

$$p(y^{(1)}, \ldots, y^{(n)}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{j=1}^{n} p(y^{(j)} \mid \boldsymbol{x}^{(j)}; f) \qquad (2.31)$$

- データやパラメータの生成過程を確率分布で表す

→生成モデル

続き

- 生成モデル―関数のパラメータが先に決まる
  サンプルはその後に決まる

サンプルから関数を推定・・・ベイズの公式を利用

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \qquad (2.32)$$

Bを観測した事後におけるAの分布・・・事後分布
最も大きくするAの値を見つける・・・MAP推定
―右辺の分子を最大

続き

- パラメータwがA,出力yがBに相当する。
- ベイズの公式の右辺の分子は式(2.31)になり、対数を取ると、

$$\sum_{i=1}^{n} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; f) + \log p(\boldsymbol{w})$$
$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - f(\boldsymbol{x}^{(i)}))^2 - \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \text{定数} \qquad (2.33)$$

- MAP推定は、線形モデルで正規化付きの二乗誤差を最小にすること

(b)正規過程からカーネルへ

- 入力値 $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$ に対して、$f(\boldsymbol{x}^{(1)}), f(\boldsymbol{x}^{(2)}), \ldots, f(\boldsymbol{x}^{(n)})$ が平均0、分散共分散行列 $V=(V_{ij})_{i,j=1,\cdots,n}$ をもつ多次元正規分布に従うとする。
- $f(\boldsymbol{x}^{(i)})$ は正規確率変数、$f(\boldsymbol{x}^{(i)})$ と $f(\boldsymbol{x}^{(j)})$ の共分散が

$$V_{ij} = \mathrm{E}_{f(\boldsymbol{x}^{(i)}), f(\boldsymbol{x}^{(j)})}[f(\boldsymbol{x}^{(i)}) f(\boldsymbol{x}^{(j)})] \qquad (2.34)$$

となる。(Eはxに関する期待値)

- この確率過程をx上の正規過程と呼ぶ

つづき

- 任意の入力点 $\boldsymbol{x}^{\mathrm{new}}$を固定し、その点での関数値の値の分布を調べる。サンプルに対するグラム行列を $K=(K_{ij})_{i,j=1,\cdots,n}$ とおくと、サンプル $\boldsymbol{x}^{(i)}$ と $\boldsymbol{x}^{\mathrm{new}}$に関する成分は $k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{\mathrm{new}})$ で与えられる。式2.29と式2.31に相当する関数値とデータは

$$
\begin{aligned}
&p(y^{(1)}, \ldots, y^{(n)}, f(\boldsymbol{x}^{(1)}), \ldots, f(\boldsymbol{x}^{(n)}), f(\boldsymbol{x}^{\mathrm{new}})) \\
&\quad = p(f(\boldsymbol{x}^{(1)}), f(\boldsymbol{x}^{(2)}), \ldots, f(\boldsymbol{x}^{(n)}), f(\boldsymbol{x}^{\mathrm{new}})) \prod_{i=1}^{n} p(y^{(i)} \mid \boldsymbol{x}^{(i)}; f) \quad (2.36)
\end{aligned}
$$

と書ける。

つづき

- 分布は全て正規分布→サンプルが与えられたもとでの関数値fの事後分布も正規分布。関数値fの平均は

$$\mathrm{E}_{f(\boldsymbol{x}^{\mathrm{new}})}[f(\boldsymbol{x}^{\mathrm{new}}) \mid \mathcal{D}] = \sum_{i=1}^{n} \{(K+\sigma^2 I_n)^{-1}\boldsymbol{y}\}_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{\mathrm{new}}) \qquad (2.37)$$

となる。（Dは、与えられたサンプル。）

- 関数近似の場合の $\boldsymbol{\alpha} = (K+\lambda I_n)^{-1}\boldsymbol{y}$ (1.17)

のαを $f(\boldsymbol{x}) = \sum \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x})$ に代入したものになっている。

ー＞カーネル関数の重みつき和のモデルを正規化付きで求めた関数近似の結果＝正規過程を事前分布として用いた場合のMAP推定の結果

２．４汎化能力の評価とモデル選択

- データ解析においては、汎化能力が重要

ーサンプルにフィットするだけではなく、背後にある構造を正しく抽出することが求められる。

・モデル選択・・・汎化能力を高めるためにモデルの複雑度を調整すること

・汎化能力の理論的な話・・・7章

汎化能力を評価する方法を説明

(a)クロスバリデーション

- 汎化能力は学習に使ったサンプル以外のデータに対する性能

①学習用のサンプルで学習

②残しておいたテスト用のサンプルで性能評価を行う

- 問題点ー・テストデータを残しすぎ→学習データが少ない
  ・テストデータが少なすぎ→評価結果が不安定

- 学習用サンプルとテスト用サンプルの分け方をいろいろ変えて得たテスト誤差を平均する・・・クロスバリデーション

k−foldクロスバリデーション法 (k−fold CV)

- 「1」サンプルをk個のデータに分ける
- 「2」i=1,・・・,kに対し以下を繰り返す
  (1)i番目のグループを除いたデータで学習を行う
  (2)i番目のグループでテスト誤差を評価し r_i とおく
- 「3」$\sum_{i=1}^{k} r_i/k$ をテスト誤差の推定値とする

(b)線形モデルのlevave-one-out クロスバリデーション

・k-fold CV法は、一つだけのサンプルをテストデータとして除いておく方法・・・leave-one-outクロスバリデーションと呼ぶ

・y＝f(x)を学習したとする。

i番目のサンプルの入力 x^i → $\tilde{y}^{(i)}=f(\boldsymbol{x}^{(i)})$

＝ サンプルの出力 $y^{(i)}$ のノイズ成分を除去した推定値

・カーネル回帰の場合、$\tilde{y}^{(j)}=\sum_{i=1}^{n}\alpha_i k(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)})$ と

$$\boldsymbol{\alpha}=(K+\lambda I_n)^{-1}\boldsymbol{y} \qquad (1.17)$$ より、

$$\tilde{\boldsymbol{y}}=K\boldsymbol{\alpha}=(K+\lambda I_n)^{-1}K\boldsymbol{y} \qquad (2.38)$$ が得られる

つづき

- 一般に $\tilde{\boldsymbol{y}}=H\boldsymbol{y}$ という線形関係があるとする。

カーネル回帰の場合は $H=(K+\lambda I_n)^{-1}K$ である。

・leave-one-out CV誤差は、 学習サンプルとテストサンプルを分ける手続きなしで、$y^{(i)}$ と $\tilde{y}^{(i)}$ の重みつきの誤差平均

$$\mathrm{CV}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y^{(i)}-\tilde{y}^{(i)}}{1-H_{ii}}\right)^2 \qquad (2.39)$$

・ただし、H_{ii} はHの第i対角成分である。

(c)具体例

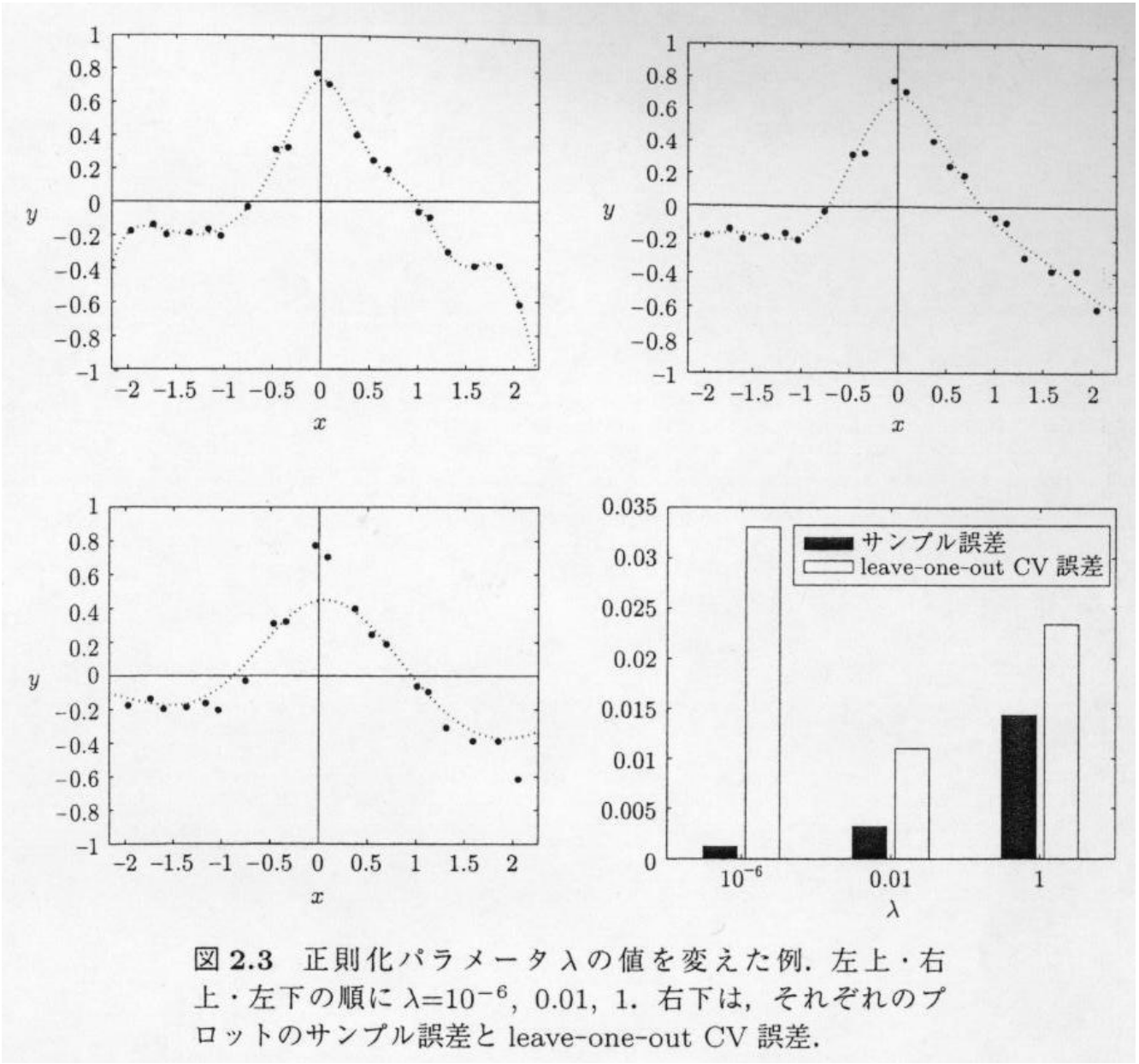

図 2.3　正則化パラメータ λ の値を変えた例．左上・右上・左下の順に $\lambda=10^{-6}$, 0.01, 1．右下は，それぞれのプロットのサンプル誤差と leave-one-out CV 誤差．

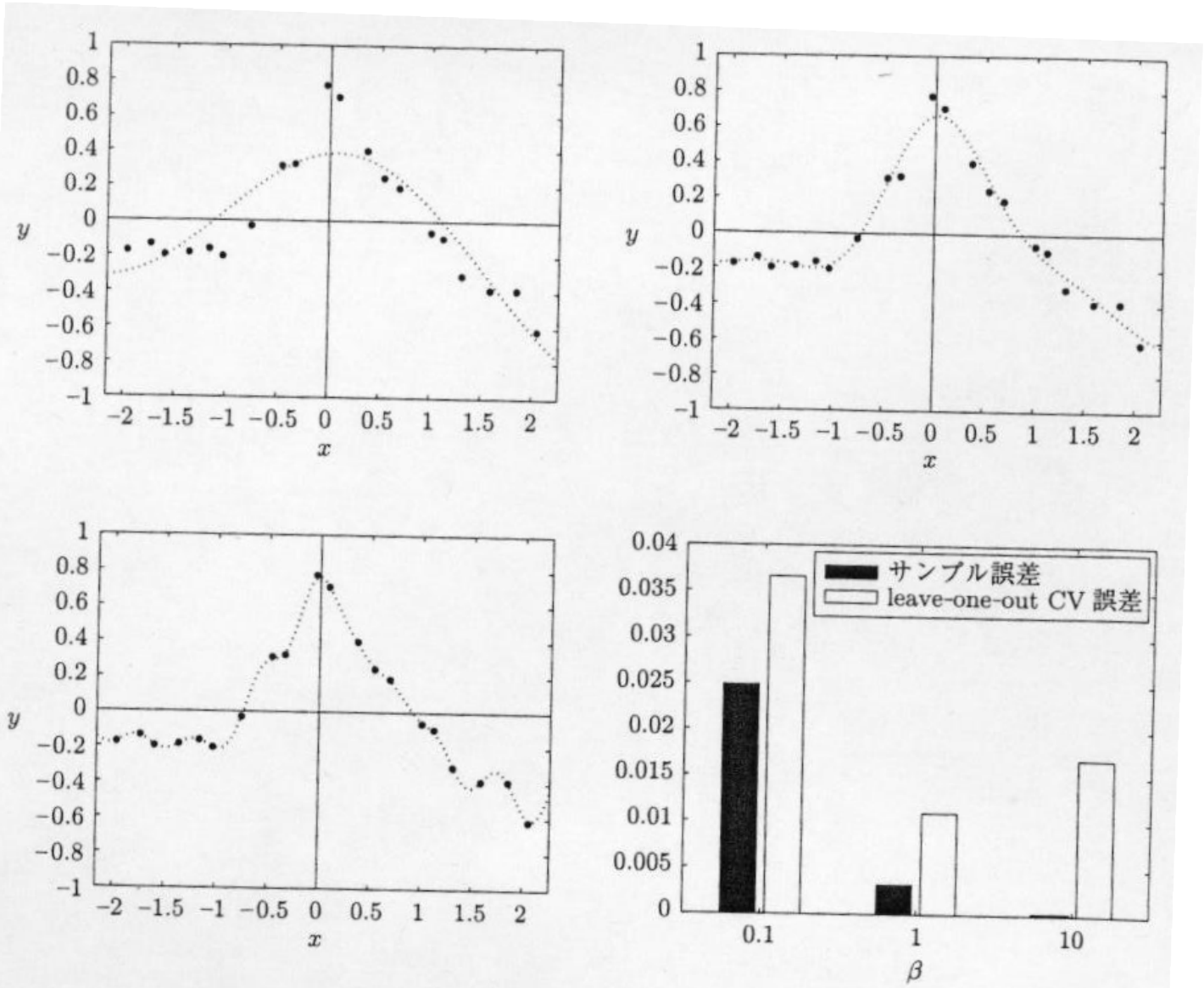

図 2.4 ガウスカーネルのパラメータ β の値を変えた例. 左上・右上・左下の順に β=0.1, 1, 10. 右下は, それぞれのプロットのサンプル誤差と leave-one-out CV 誤差.

つづき

・leave−one−out　CV誤差は汎化能力を評価する一つの尺度と考えることができるので、モデル選択を行うことができる。

・いくつかの β と λ に対してleave−one−out CV誤差を計算し、最も小さな値となる β と λ を選ぶ