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

２．３ 確率モデルからの導入
２．４ 汎化能力の評価とモデル選択

茨城大学工学部情報工学科
國井慎也

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すること

  -ただし、データには、ノイズがのっていたり、関数のうち有限個の点のデータしか関数値が与えられていなかったり、不確定な要素が多い

- このときの解決策として有効なのは、データの生成過程を確率分布を用いてモデル化すること

線形モデルによる関数近似の確率モデル

- ここで考える確率モデルは、出力yは関数値 $f(x)=w^t x$ そのものではなく、ランダムノイズがのった

$$y=f(x)+\varepsilon$$

- x,fが与えられたもとでのyの条件付き確率

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

ここで、ε は独立な正規分布
この正規分布の分散を σ^2 とする

関数fの確率モデル

- 線形モデル$y=w^{T}x$ではパラメータwを決めることとfを決めることは等価である。

- Wの各成分が独立に平均0で分散が $\frac{1}{\lambda}$であるような正規分布から生成されたとすると、wの分布は

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

生成モデル（１）

- サンプルは以下のような過程で生成されるものと仮定する。

  - p(w)に従ってパラメータwがランダムに決められる。

  - そのwで決まる関数f(x)をn個の点$x^{(1)},\ldots,x^{(n)}$で計算し、p(y|x;f)の分布に従ってノイズがのってサンプル出力$y^{(1)},y^{(2)},\ldots,y^{(n)}$が観測される。

生成モデル（2）

- パラメータとデータの同時確率分布は

$$p\left(y^{(1)},\ldots,y^{(n)},w\right)=p(w)\prod_{j=1}^{n}p\left(y^{(j)}|x^{(j)};f\right)$$

- このように、データやパラメータの生成過程を確率分布で表したものを生成モデルという

ベイズの定理(1)

- 生成モデルでは関数のパラメータが予め決まって、それからサンプルが決められた。

  -しかし、サンプルから関数を推定したいときは、それを逆向きに行う必要がある。それをするのがベイズの定理である。

  確率変数A,Bがあるとき

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

  これをベイズの定理という

ベイズの定理(2)

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

- ベイズの定理は、AからBが生成される生成モデルがあるときに、逆にBを観測した上で、Aを推論する方法である。

- この確率はBを観測した事後におけるAの分布なので事後分布という

MAP推定

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

- p(A|B)を最大にするようなAの値を見つけるにはMAP推定という手法を使う。

  -右辺の分母はAの値に依存しないため、右辺の分子を最大にすることだけを考えればよい。

  -線形モデルでは、パラメータwがA、出力値$y^{(1)},y^{(2)},..,y^{(n)}$Bに相当する。

線形モデルのMAP推定(1)

- 線形モデルのベイズの公式の右辺の分子は式(2.31)と同じになる

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

- 式(2.31)の対数をとると、

$$\sum_{i=1}^{n}\log p\left(y^{(i)}|x^{(i)};f\right)+\log p(w)$$

$$= \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)}-f\left(x^{(i)}\right)\right)^2-\frac{\lambda}{2}\|w\|^2+\text{定} \quad \square$$

線形モデルのMAP推定(2)

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)}-f\left(x^{(i)}\right)\right)^2-\frac{\lambda}{2}\|w\|^2+\text{定}\quad\square$$

- 第一項は二乗誤差にマイナスを付けたもので、第二項は正則化項に対応している。

- つまり、MAP推定は線形モデルでの正則化付きの二乗誤差を最小にすることと等価になっている。

正規過程

- パラメータwではなく、関数f(x)が正規分布に従うと考える。
- f(x)が正規分布に従うとは、
  任意の個数n個の入力値$x^{(1)},x^{(2)},\ldots,x^{(n)}$に対して、
  $f(x^{(1)}),\ldots,f(x^{(n)})$平均0、分散共分散行列$V=(V_{ij})_{i,j=0,\ldots,n}$をもつ
  多次元正規分布N[0,V]に従う。
- $f(x^{(i)})$は正規確率変数で、$f(x^{(i)})$と$f(x^{(j)})$の共分散が

$$V_{ij} = E_{f(x^{(i)}),\, f(x^{(j)})}\left[f\left(x^{(i)}\right) f\left(x^{(j)}\right)\right]$$

ここでE[ ]は確率変数の期待値を表す。この確率過程をX上の正規過程(GP=Gaussian Process)と呼ぶ

正規過程からカーネルへ

- V_{ij}は$x^{(i)}$と$x^{(j)}$で決まるので、関数$k(x^{(i)},x^{(j)})$とみなすことができ,関数kが正定値であって、カーネル関数とみなすことができる。

- カーネル関数kは、そのグラム行列を分散共分散行列とみなすことによって、正規過程と等価になる。

正規過程のMAP推定

- サンプルが与えられたときの関数の推定をMAP推定で考える。

  -パラメータwの分布を求めるのではなく、関数

  f(x)の分布を求める。

  -サンプル点のfにf(x^{new})を加えたn+1個の関数値の事前分布は正規過程でモデル化できる。

  サンプルに対するグラム行列を$\mathbf{K}=(\mathbf{K}_{ij})_{i,j=0,\ldots,n}$とおくとf同時確率分布は

$$p\left(f\left(x^{(1)}\right), f\left(x^{(2)}\right), \ldots, f\left(x^{(n)}\right), f\left(x^{new}\right)\right)$$

正規過程のMAP推定

$$p\left(f\left(x^{(1)}\right), f\left(x^{(2)}\right), \ldots, f\left(x^{(n)}\right), f\left(x^{new}\right)\right)$$

- 上の式は平均0の正規分布
- 分散共分散行列は、サンプル$x^{(i)}$と$x^{(j)}$に関する成分がK_{ij}で、サンプル$x^{(i)}$とx^{new}に関する成分は$k(x^{(i)},x^{new})$で与えられる。
- f(x)を固定したもとでのyの分布は線形モデルと同じく式(2.29)となる

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

- 関数値とデータの同時分布は

$$p\left(f\left(y^{(1)}\right),\ldots,f\left(y^{(n)}\right),f\left(x^{(1)}\right),f\left(x^{(2)}\right),\ldots,f\left(x^{(n)}\right),f\left(x^{new}\right)\right)$$

$$=f\left(\left(x^{(1)}\right),f\left(x^{(2)}\right),\ldots,f\left(x^{(n)}\right),f\left(x^{new}\right)\right)\prod_{i=1}^{n}p\left(y^{(i)}|x^{(i)};f\right)$$

- ここに出てくる分布はすべて正規分布

  サンプルが与えられたもとでの$f(x^{(new)})$の事後確率も正規分布になる。

f(x^{new})の事後確率の平均

$$E_{f(x^{new})}\left[f\left(x^{new}\right)\middle|D\right] = \sum_{i=1}^{n}\left\{\left(K+\sigma^2 I_n\right)^{-1} y\right\}_i k\left(x^{(i)}, x^{new}\right)$$

- ただし、与えられたサンプルをDと書いた

  $\{\}_i$はかっこの中のベクトルの第i成分を表す
- 事後分布が正規分布なら、MAP推定はその平均に一致する。

正規過程とカーネルの等価性

$$\sum_{i=1}^{n} \left\{ \left(K + \sigma^2 I_n \right)^{-1} y \right\}_i k\left(x^{(i)}, x^{new} \right)$$

- 上の式は、関数近似の場合の式(1.17)のαを

$$f(x) = \sum_i \alpha_i k\left(x^{(i)}, x \right)$$

に代入したもである。（ノイズの分散σ2が正則化パラメータλに対応）

- つまり、カーネル関数の重みつき和のモデルを正則化付きで求めた関数近似の結果が、そのカーネル関数と等価な正規過程を事前分布として用いた場合のMAP推定と一致する

汎化能力の評価とモデル選択

- データ解析においては、汎化能力が高いあてはめを行うことが重要

  -カーネル法では、サンプル数と同じだけの自由度があるため、過学習を引き起こす。

  -モデルの形があまり複雑になりすぎないようにすることが必要

  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をモデル選択という

クロスバリデーション

- 汎化能力とは、学習に使ったサンプル以外のデータに対する性能

  -サンプルを学習用とテスト用に分け、初めに学習用のサンプルで学習した後に、テスト用のサンプルで性能評価を行う。

  -問題点として、テストデータを残しすぎると学習データが少なくて学習に十分な性能が出せず、一方、テストデータが少なすぎると、評価結果が不十分になる。

k-foldクロスバリデーション

- K-foldクロスバリデーション

  [1]まず、サンプルをk個のグループに分ける。

  [2]i=1,...,kに対して以下を繰り返す。

  (i)i番目のグループを除いたデータで学習を行う

  (ii)i番目のグループでのテスト誤差をr_iとする

  [3]$\sum_{i=1}^{k} \frac{r_i}{k}$ をテスト誤差の推定値(クロスバリデーション誤差)とする

線形モデルの leave-one-outクロスバリデーション（1）

- サンプルがn個あるとき、n-foldクロスバリデーションは1つだけのサンプルデータとしてのぞいておく方法。特にleave-one-outクロスバリデーションと呼ばれる。

- 関数近似でy=f(x)を学習したとする。

  このとき、入力$x^{(i)}$を学習した関数に入れると
  $\tilde{y}^{(i)} = f\left(x^{(i)}\right)$ という結果が得られる。

  これは、サンプル出力y(i)のノイズを除去した推定値としてみなすことができる

線形モデルの leave-one-outクロスバリデーション（2）

- 線形回帰やカーネル回帰では、$\tilde{\boldsymbol{y}}=(\tilde{y}^{(1)},\ldots,\tilde{y}^{(1)})^T$ が $\boldsymbol{y}=(y^{(1)},\ldots,y^{(1)})^T$ の線形変換でかける。
- 実際、カーネル回帰の場合、$\tilde{y}^{(i)}=\sum_{i=1}^{n}\alpha_i k(x^{(i)},x^{(j)})$
- と $\boldsymbol{\alpha}=(K+\lambda I_n)^{-1}y$ より、

$$\tilde{y}=K\boldsymbol{\alpha}=(K+\lambda I_n)^{-1}K\,\boldsymbol{y}$$

- 一般に$\tilde{\boldsymbol{y}}=H\,\boldsymbol{y}$という線形関係があるときを考える。カーネル回帰の場合は、H＝$(K+\lambda I_n)^{-1}K$ である。

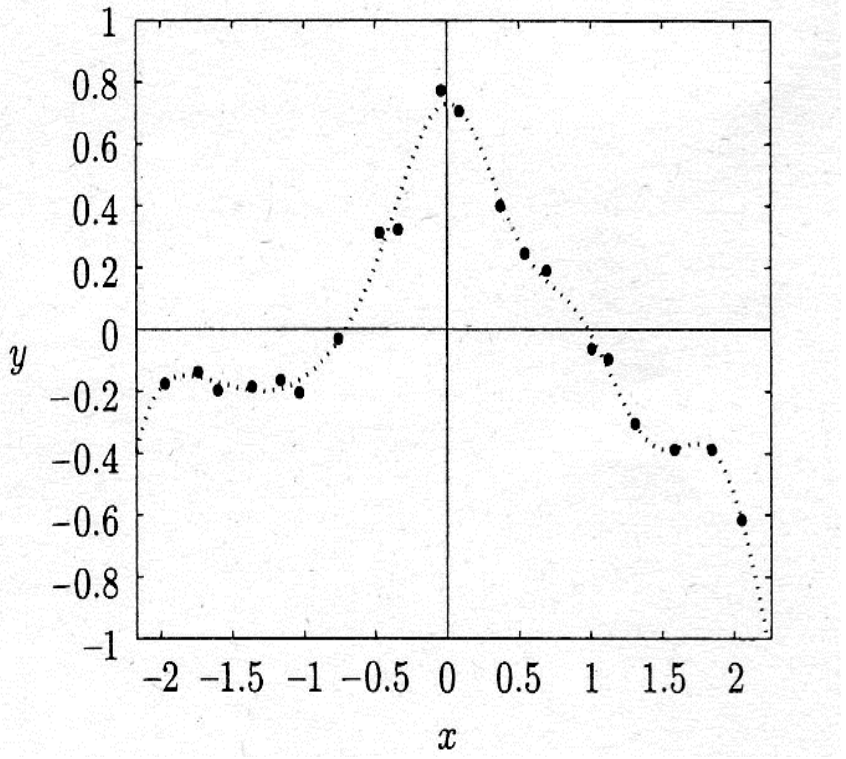
線形モデルの leave-one-outクロスバリデーション（3）

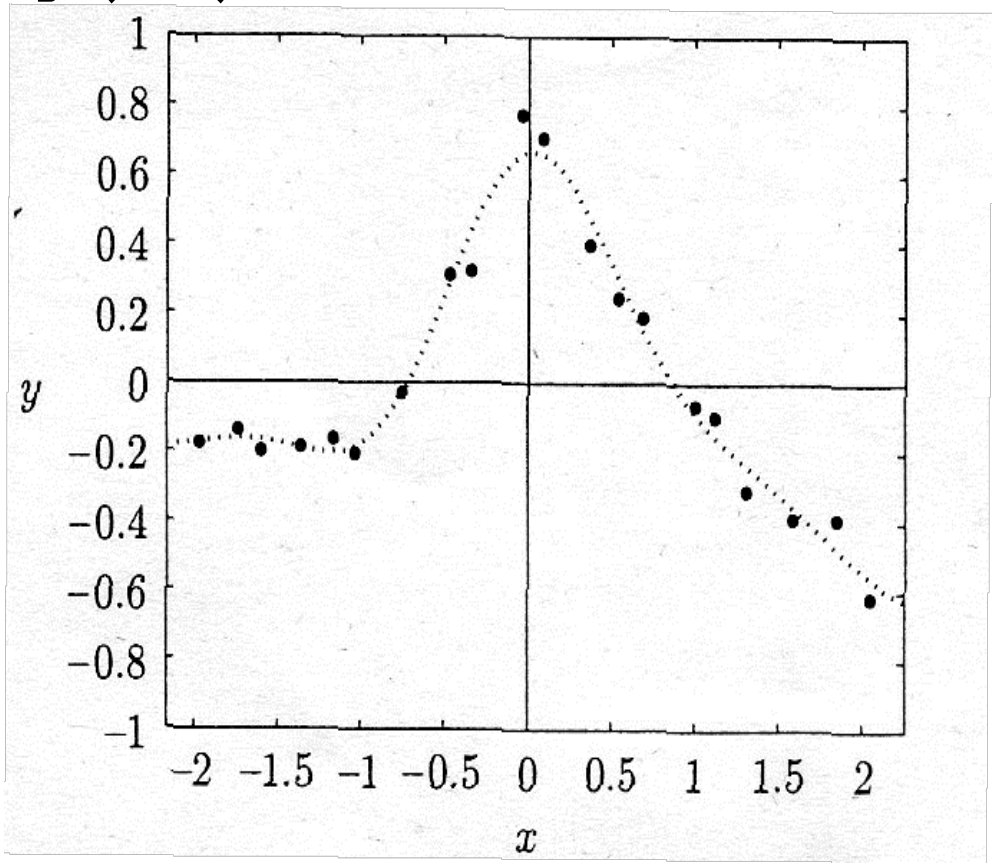
- このときのleave-one-outクロスバリデーション誤差（CV誤差）は、学習サンプルとテストサンプルを分ける手続きなしで、サンプル出力$y^{(i)}$とノイズ成分を除去した推定値 $\tilde{y}^{(i)}$ の重みつきの平均誤差として、

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

- ただし、H_{ii}はHの第i対角成分である。

具体例(1)



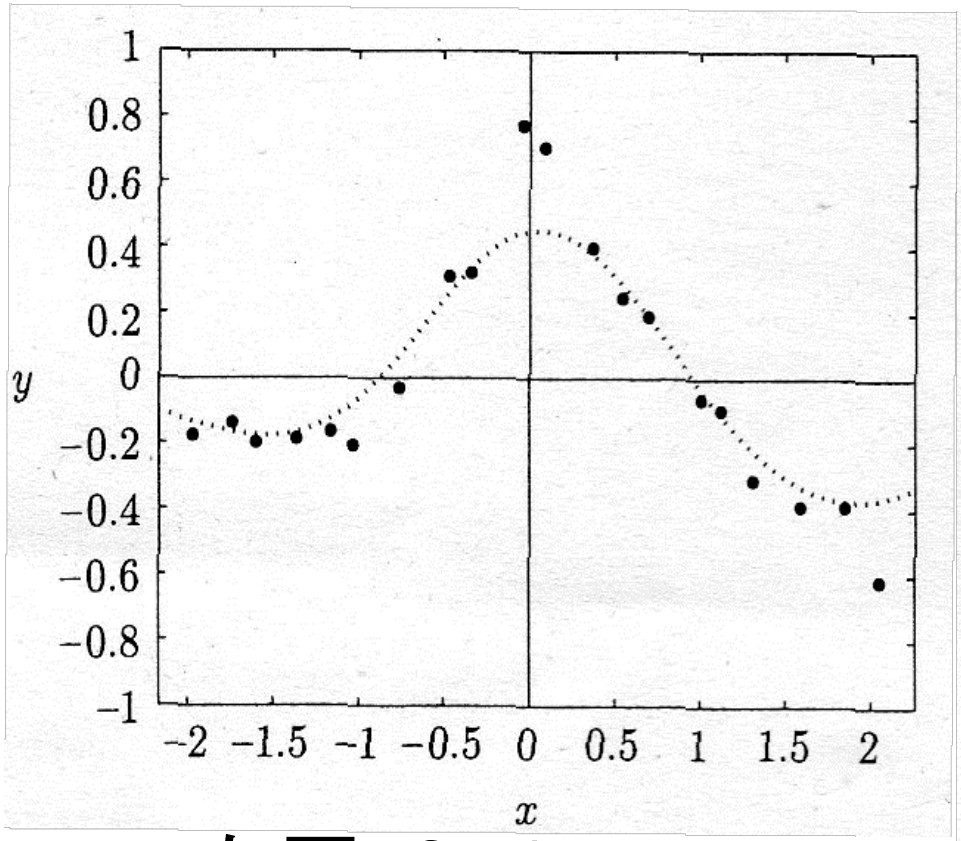


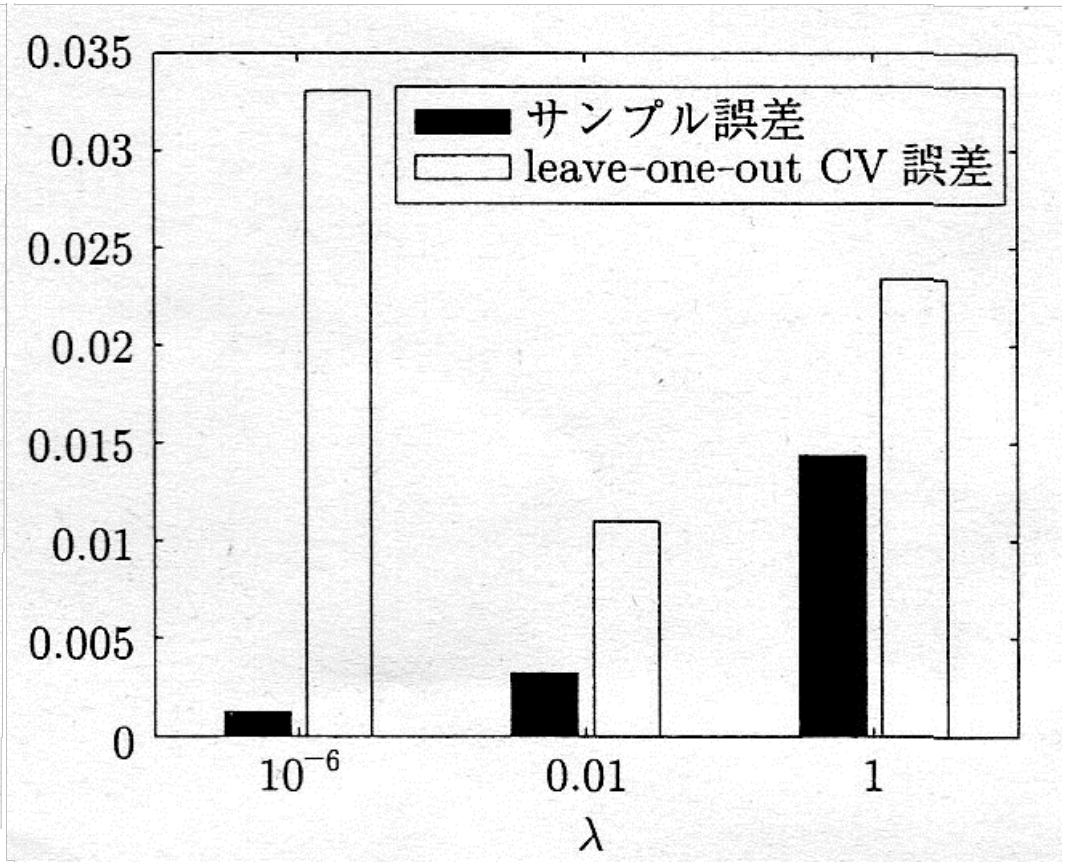
- β＝1を固定

  左図:λ＝10^{-6}

  右図:λ＝0.01

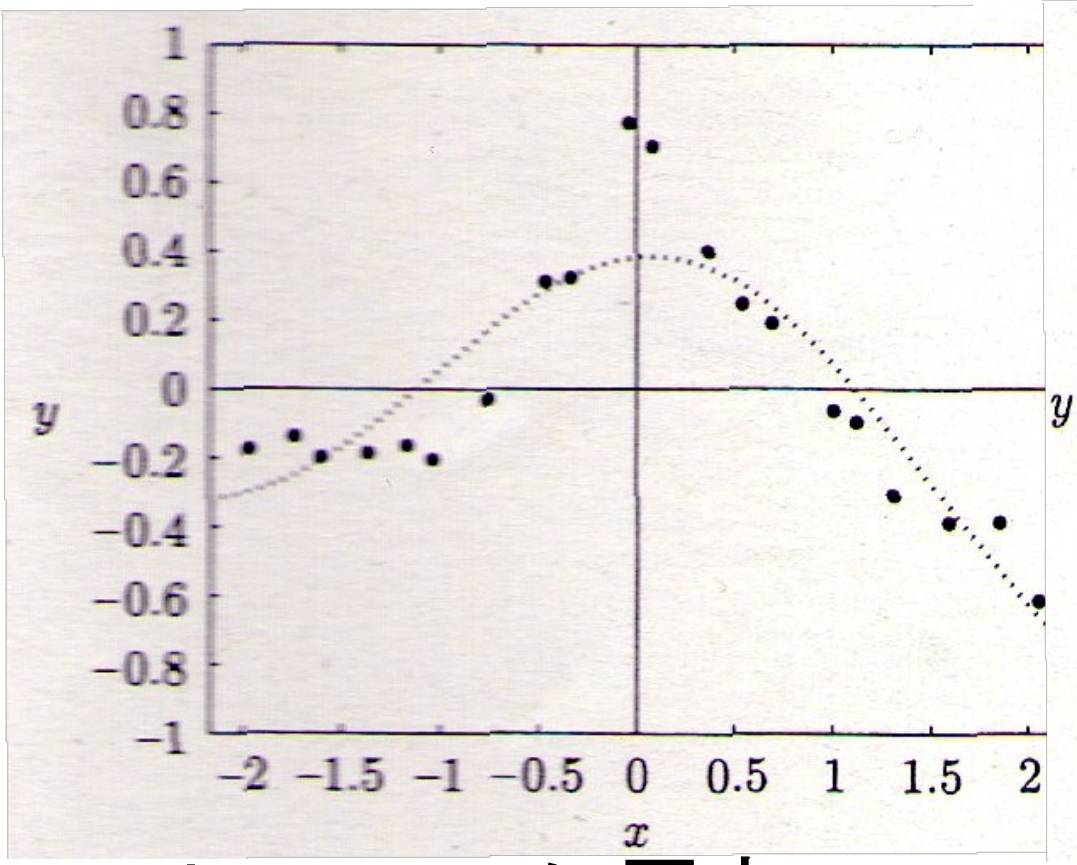
具体例(2)

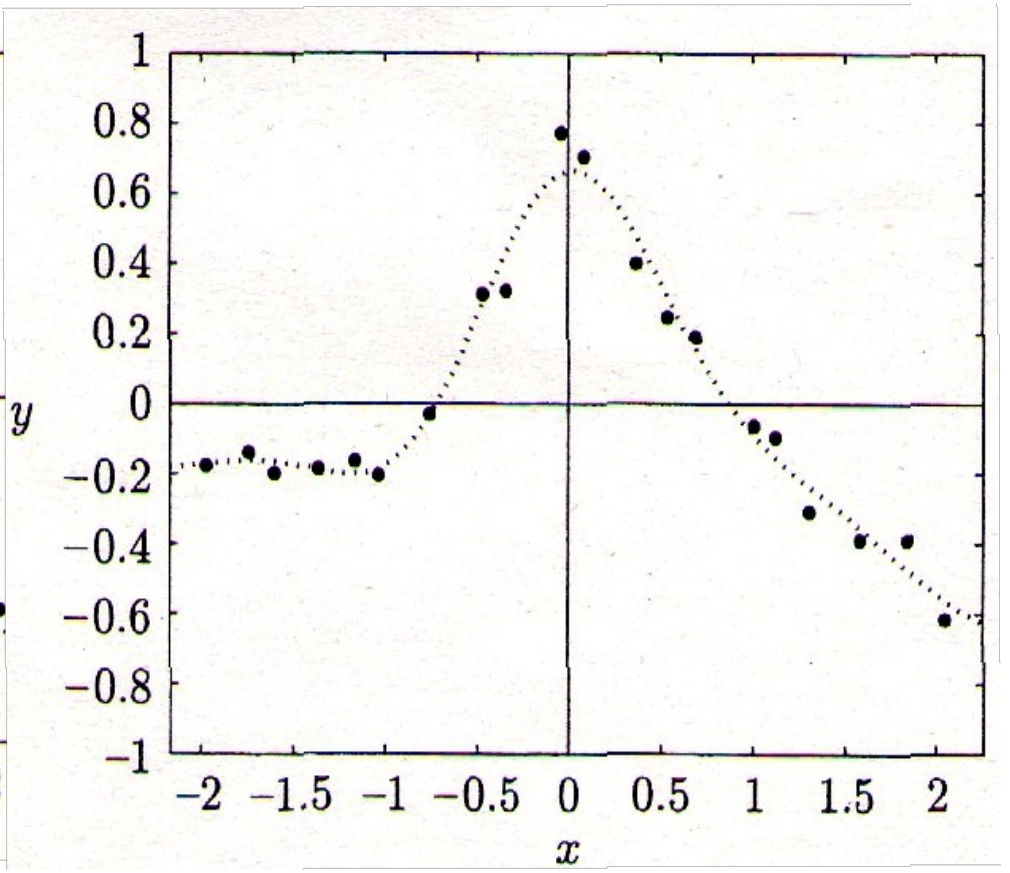




- 左図:β=1,λ＝1
- 右図:サンプル誤差とCV誤差の比較

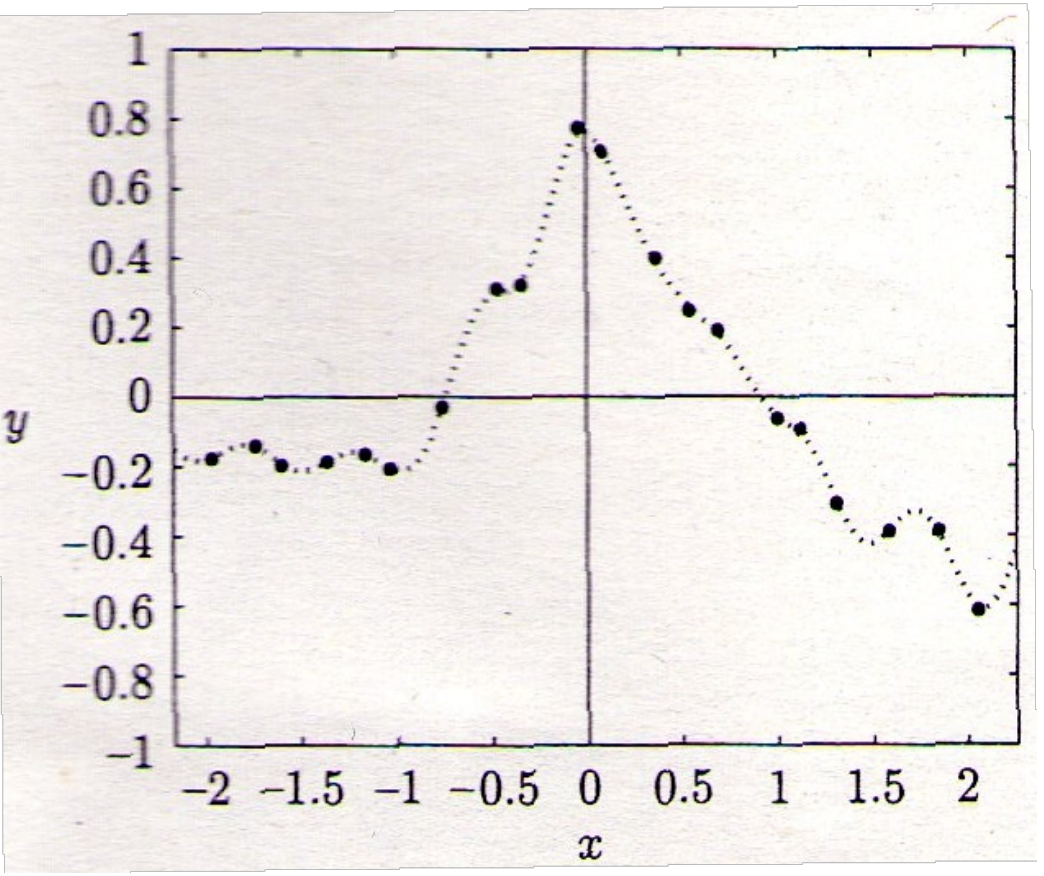
具体例(3)

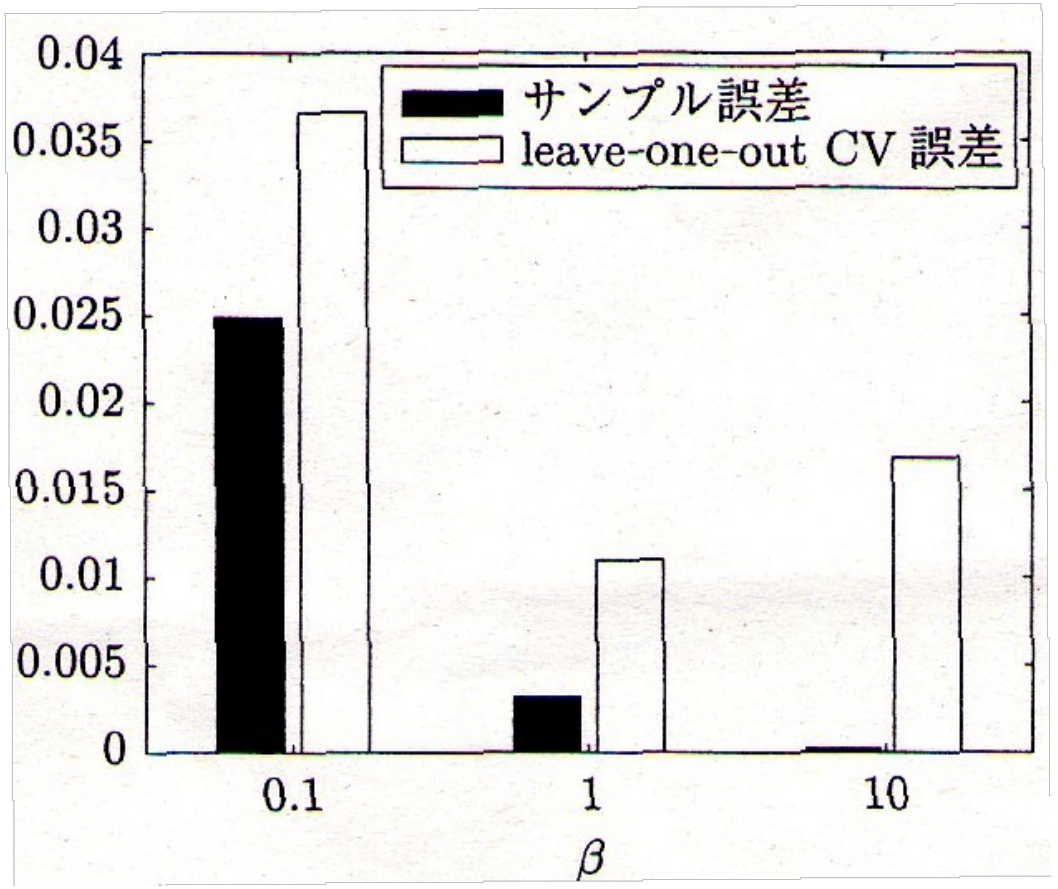




- λ＝0．01に固定、
- 左図:β＝0.1
- 右図:β＝1

具体例(4)





- 左図:λ＝0.01, β=10
- 右図:サンプル誤差とCV誤差の比較

βとλの選択

- βを増やすことと、λを減らすことが同じような振る舞いを示す。

- 結果的に、いくつかのβとλに対してleave-one-out CV誤差を計算し、そのうち最も小さな値をとるようなβとλを選べばよい。