横浜国立大学集中講義「現代物理学」
第9回　2021年11月29日

カーネル法とガウス過程

岡山大学 異分野基礎科学研究所

大槻純也

ベイズ線形回帰の課題

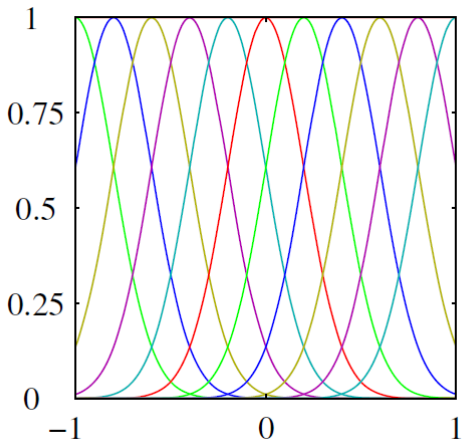
計算コスト $O(M^3)$

M: 基底関数 $\{\phi_j(\mathbf{x})\}$ の数 ＝ 重み $\{w_j\}$ の数

$$S = (\beta\Phi^{\mathrm{T}}\Phi + \alpha\mathbf{I})^{-1}$$

変数 $\mathbf{x} = \{x_1, x_2, \cdots, x_D\}$ の次元が増えると
M が指数関数的に増大

$$M \sim L^D \qquad L: \text{各変数 } x_i \text{ の分割数}$$



対処法１：
モンテカルロ法を使う
(Markov Chain Monte Carlo法)

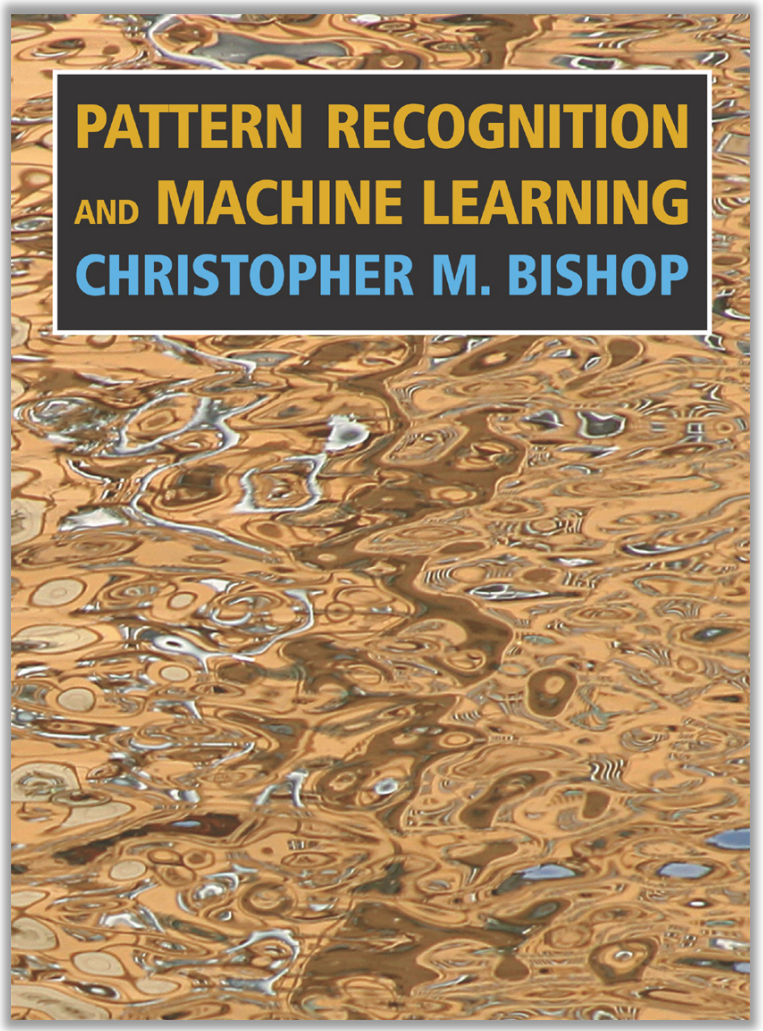
尤度関数

$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left[-\beta E(\boldsymbol{w})\right]$$

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left[y(x_n, \boldsymbol{w}) - t_n\right]^2$$

対処法２：
重みを消去 → カーネル法 (今日の内容)

参考文献 (ガウス過程)



←
初回に紹介した
PRMLの第6章
“Kernel Methods”

→
こちらの本もおすすめです。
表記がPRMLと同じにしてあるので、PRMLと並行して読んでも混乱しません。

カーネル

ベイズ線形回帰における予測分布 $p(t|x, \mathbf{x}, \mathbf{t})$ の平均

$$
\begin{aligned}
m(x) &= \beta \boldsymbol{\phi}^{\mathrm{T}}(x) S \Phi^{\mathrm{T}} \mathbf{t} \\
&= \sum_{n=1}^{N} \beta \boldsymbol{\phi}^{\mathrm{T}}(x) S \boldsymbol{\phi}(x_n) t_n \\
&\equiv \sum_{n=1}^{N} k(x, x_n) t_n
\end{aligned}
$$

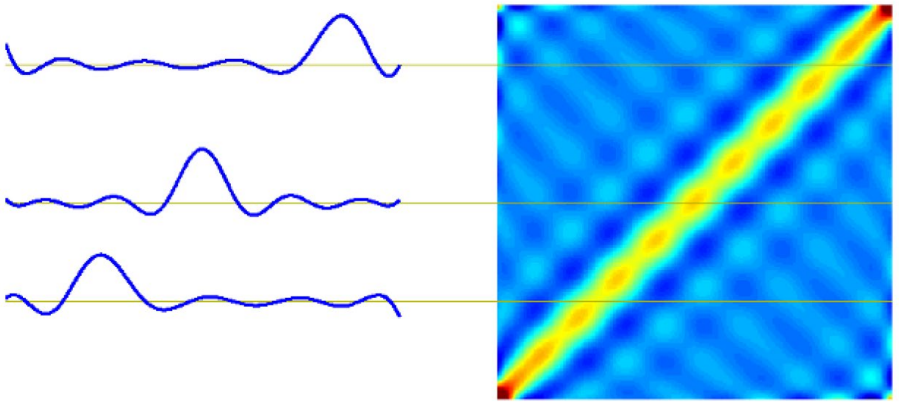
$\{t_n\}$ を含まない部分

等価カーネル (equivalent kernel)

$$k(x, x') = \beta \boldsymbol{\phi}^{\mathrm{T}}(x) S \boldsymbol{\phi}(x)$$

基底関数 $\{\phi_j(x)\}$ とデータ点 $\{x_n\}$ に依存

$k(x, x')$ の例：ガウス関数基底の場合



PRML Fig. 3.10

$x = x'$ にピークを持つ (局在している)

→ t の予測において、近くの点が強く影響する

カーネルトリック (kernel trick)

基底関数を用意して $k(x, x')$ を計算する代わりに、$k(x, x')$ の関数形を仮定してしまう

ガウス過程

恒例の線形モデル

$$y(x, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(x)$$

今までは $\boldsymbol{w}$ の分布に注目していた
つまり、$\boldsymbol{w}$ の事後確率分布 → $y(x, \boldsymbol{w})$ の事後確率分布
今度は関数 $y(x, \boldsymbol{w})$ の分布を直接考える

パラメータ $\boldsymbol{w}$ の事前分布を仮定

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

データ点がN個ある

$$\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$$

このとき $y_n \equiv y(x, \boldsymbol{w})$ の分布は？

$$y_n \equiv y(x_n, \boldsymbol{w}) \quad \mathbf{y} = \{y_1, y_2, \cdots, y_N\}$$

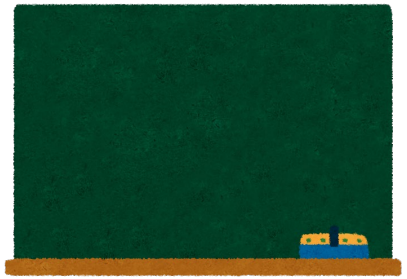
結果： $\boldsymbol{y}$ の分布はガウス分布になる

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, K)$$

導出

$$\mathbb{E}[\mathbf{y}] = \mathbf{0}$$

$$\mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \frac{1}{\alpha}\Phi\Phi^{\mathrm{T}}$$



より共分散 K は

$$K = \frac{1}{\alpha}\Phi\Phi^{\mathrm{T}}$$

kernel matrix
カーネル行列

$A^T A$, AA^T の形を
一般にグラム行列
(Gram matrix)と呼ぶ

$$K_{nm} = \frac{1}{\alpha}\boldsymbol{\phi}^{\mathrm{T}}(x_n)\boldsymbol{\phi}(x_m) \equiv k(x_n, x_m)$$

kernel function
カーネル関数

基底ベクトルの内積
x_n と x_m がどれだけ似ているか

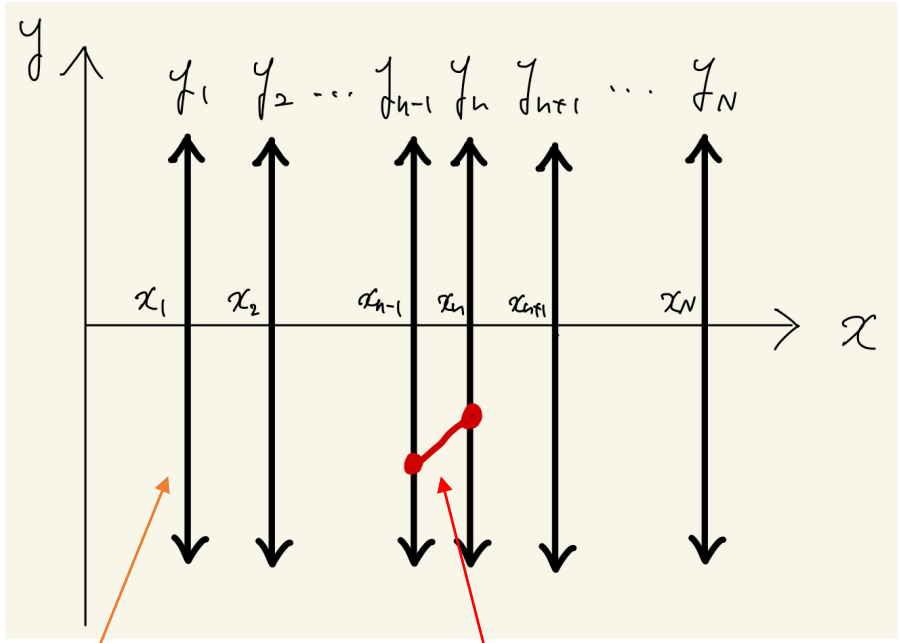
ガウス過程のイメージ

測定前

$\{y_1, \cdots, y_N\}$ は決まっていない
ただし $\{x_1, \cdots, x_N\}$ は与えられている

関数 $y(x)$ の事前確率

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, K)$$



分散

$$K_{nn} = k(x_n, x_n)$$

関数 $y(x)$ は揺らいでいる

相関している

$$K_{nm} = k(x_n, x_m)$$

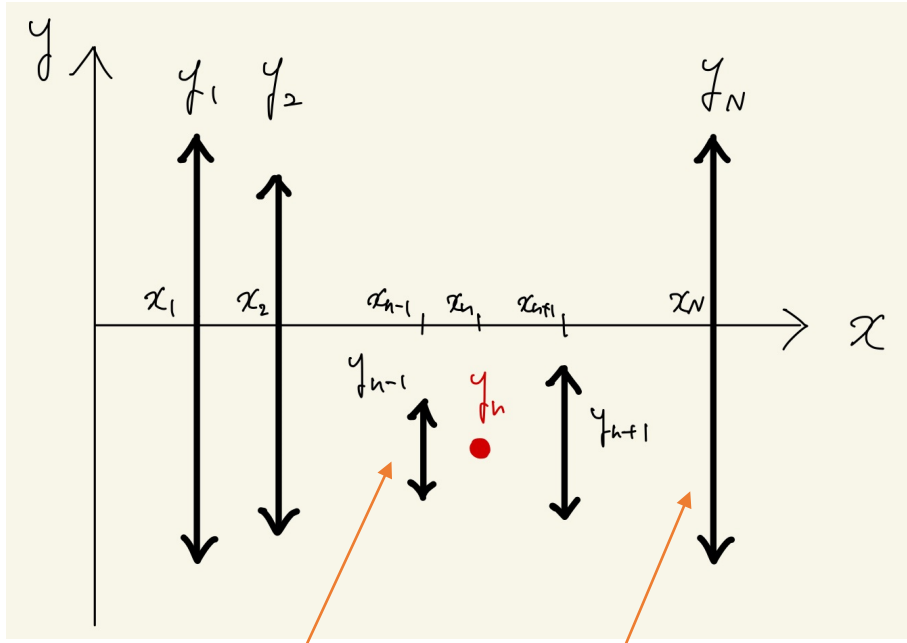
独立に揺らいでいるわけではない

測定後

y_n の値が決まった

関数 $y(x)$ の条件付き確率

$$p(\mathbf{y}|y_n)$$



測定点の近くは
分散が小さくなる

測定点から遠い点は
分散が大きいまま

カーネル関数

カーネル関数の定義　　$\mathbf{x} = \{x_1, \cdots, x_D\}$

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}') = \sum_{j=1}^{M} \phi_j(\mathbf{x})\phi_j(\mathbf{x}')$$

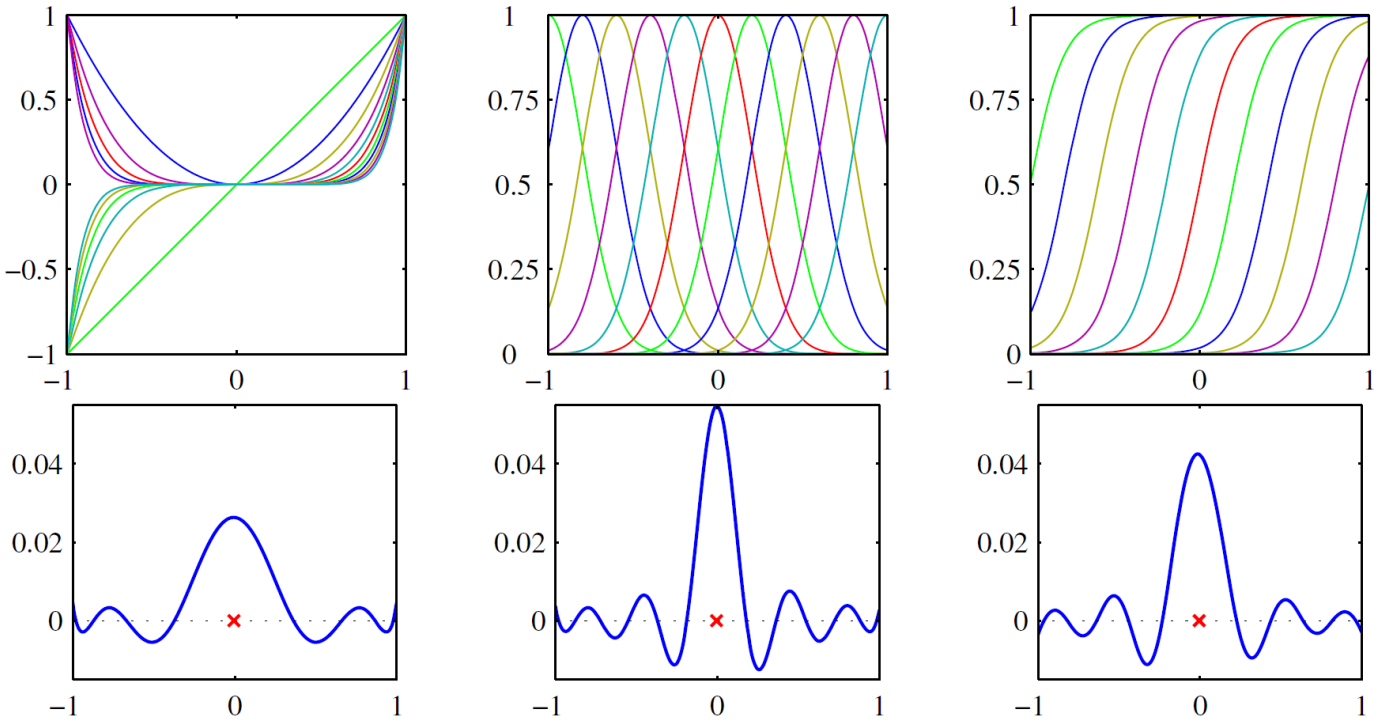
基底関数 $\{\phi_j(\mathbf{x})\}$ を仮定すれば計算できる

共通する性質

- $\mathbf{x} = \mathbf{x}'$ でピーク
- 速く減衰

このような性質を持つ $k(\mathbf{x}, \mathbf{x}')$ の
関数形を直接与えてもよい

基底関数から作ったカーネルの例 (D =1)



PRML Fig. 6.1

カーネル関数

基底関数を仮定して計算する代わりに
$k(\mathbf{x}, \mathbf{x}')$ の関数形を直接与える

$k(\mathbf{x}, \mathbf{x}')$ が満たすべき条件

1. $\mathbf{x}, \mathbf{x}'$ について対称
2. $\mathbf{x} = \mathbf{x}'$ でピーク。減衰。
3. 基底関数 $\phi_j(\mathbf{x})$ への分解が可能
   (対応する基底が存在)

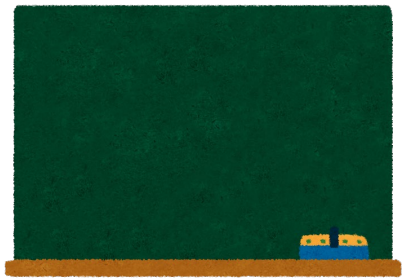
条件3の必要十分条件
任意の $\{\mathbf{x}_n\}$ から作ったカーネル行列 K が
positive semidefinite (半正定値性)

Gaussian kernel (ガウスカーネル)

$$k(\mathbf{x}, \mathbf{x}') = \theta_1 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\theta_2}\right)$$

対応する基底関数は無限個のガウス関数基底

証明



ノイズ

観測データへのノイズの影響を考える

$$t_n = y_n + \epsilon_n$$

観測値　予測値

ノイズ = 確率変数

$$p(\epsilon) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

ノイズ ϵ_n が確率分布するので、
観測値 t_n も確率分布する

$$p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1})$$

N 個の観測データ $\mathbf{t} = (t_1, \cdots, t_N)$ の確率分布

$$p(\mathbf{t}|\mathbf{y}) = \prod_{n=1}^{N} \mathcal{N}(t_n|y_n, \beta^{-1})$$

$$= \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}_N)$$

多変量ガウス分布

観測データ t の確率分布
＝ marginal distribution ($\boldsymbol{y}$ について周辺化)

sum rule

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y}$$

ガウス過程

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, K)$$

$$= \int \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}_N)\mathcal{N}(\mathbf{y}|\mathbf{0}, K)d\mathbf{y}$$

ガウス分布の畳み込み積分はガウス分布
公式PRML (2.115)

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, K + \beta^{-1}\mathbf{I}_N)$$

観測ノイズがある場合、
ノイズの分散 β^{-1} をカーネル行列 K の対角項に足す

予測分布 (Predictive distribution)

N個の測定値が与えられているとして
$(N+1)$個めの測定値を予測したい

計算するべき量は conditional distribution

$$p(t_{N+1}|\mathbf{t}_N)$$

これを計算するために joint distribution

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, C_{N+1})$$

$$C_N = K_N + \beta^{-1}\mathbf{I}_N$$

から出発し、次の式を使う

$$p(\mathbf{t}_N, t_{N+1}) = p(t_{N+1}|\mathbf{t}_N)p(\mathbf{t}_N)$$

product rule

確率変数 t_{N+1} と共分散 C_{N+1} を分割

$$\mathbf{t}_{N+1} = \begin{pmatrix} \mathbf{t}_N \\ \hline t_{N+1} \end{pmatrix} \qquad C_{N+1} = \left(\begin{array}{c|c} C_N & \boldsymbol{k} \\ \hline \boldsymbol{k}^{\mathrm{T}} & c \end{array}\right)$$

$\mathbf{x}_n$ と $\mathbf{x}_{N+1}$ の相関

ガウス分布の条件付き確率はガウス分布
公式PRML (2.81), (2.82)

$$p(t_{N+1}|\mathbf{t}_N) = \mathcal{N}(t_{N+1}|m(\mathbf{x}_{N+1}), \sigma^2(\mathbf{x}_{N+1}))$$

$$m(\mathbf{x}_{N+1}) = \boldsymbol{k}^{\mathrm{T}} C_N^{-1} \mathbf{t}_N$$

$$\sigma^2(\mathbf{x}_{N+1}) = c - \boldsymbol{k}^{\mathrm{T}} C_N^{-1} \boldsymbol{k}$$

次回の予告

OKAYAMA UNIVERSITY

【実習】ガウス過程を使って実際に予測分布を計算してみる