6.4. ガウス過程

Shunsuke Shigemitsu

M2 Bilab@UT

28, September, 2012

おさらい

カーネル法：$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ のような非線形の写像 (基底関数) $\boldsymbol{\phi}$ を用いた線形結合を考えるときに、カーネル関数

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}') \tag{6.1}$$

を計算することで、$\boldsymbol{\phi}$ による写像を直接計算せずにすむという方法。
(実際の関数は、$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2)$ みたいな感じ)

今回は・・・
これを確率的識別モデルに適用してみる。

確率過程って？

確率過程 (stochastic process) とは、任意の有限な値集合 $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$ に対して、矛盾のない同時分布を与えるもの。

→ たとえば、時間とともに変化していく確率変数。
確率変数 ($y(\mathbf{x})$ とか) がたくさんあって、それを全部まとめたものが確率過程。
その中から適当に N 個を取り出したときの確率分布を考える。
実生活で目にする (?) ものでいうと、ブラウン運動などがある。

ガウス過程 (Gaussian process) は、確率過程の中でもたくさんの $y(\mathbf{x})$ の同時分布 $p(y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N))$ がガウス分布に従うようなもののこと。

6.4.1 線形回帰モデル再び

3 章でやった線形回帰モデルを用いる。

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) \quad \ldots(6.49)\ \text{モデルはこんな感じ}$$

$\mathbf{w}$ ... パラメータベクトル (M 次元)

$\boldsymbol{\phi}$ ... 非線形な写像、基底関数。M 個ある

$\mathbf{x}$ ... 入力値

流れ：

- $\mathbf{w}$ の事前分布 ($p(\mathbf{w})$) を決める
- 関数 $y(\mathbf{x}, \mathbf{w})$ に対する事前分布も決まる。
- 訓練データの集合 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ が与えられると、今度は $\mathbf{w}$ の事後分布 $p(\mathbf{w}|\mathbf{x}_1, \ldots, \mathbf{x}_N)$ が求まる。
- そこから最終的な目的である新しい入力ベクトル $\mathbf{x}$ に対する予測分布 $p(t|\mathbf{x})$ が求まる。(→ 6.4.2 節)

線形回帰ふたたび (2)

まずは、パラメータ $\mathbf{w}$ の事前分布を決める (どんな分布なのかよくわからないので、適当に決める)。

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}). \tag{6.50}$$

このとき、α は超パラメータで、分布の精度を表す (大きければ大きいほどばらつきが少ない)。
特定の $\mathbf{w}$ が決まると、$\mathbf{x}$ についての特定の関数 $y(\mathbf{x})$ が決まる。ここで、回帰を行うには、訓練データ点の集合 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ における関数 (の集合) を評価したい。その関数の値の集合を、要素

$$y_n = y(\mathbf{x}_n) \quad (n = 1, \ldots, N)$$

を持つようなベクトル $\mathbf{y}$ として表す。(6.49) 式を $\mathbf{y}$ を用いてまとめると、

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} \tag{6.51}$$

となる。ただし、$\mathbf{\Phi}$ は要素 $\phi_{nk} = \phi_k(\mathbf{x}_n)$ をもつような計画行列。

線形回帰再び (3)

$\mathbf{y}$ はガウス分布に従う $\mathbf{w}$ の線形結合なので、$\mathbf{y}$ もガウス分布に従う。ということは、平均と共分散がわかれば、$\mathbf{y}$ がどんなものなのかがわかる。(6.50) 式 $(p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}))$ より、

$$\mathbb{E}[\mathbf{y}] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}] = \mathbf{0} \tag{6.52}$$

$$\mathrm{cov}[\mathbf{y}] = \mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} = \frac{1}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{K} \tag{6.53}$$

となる。なお、$\mathbf{K}$ は、

$$K_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m) \tag{6.54}$$

を要素に持つグラム行列である。

なにがすごいの？

N 個の変数の同時分布が、平均と共分散だけで表せてしまうというところがすごい。
ちなみに、$y(\mathbf{x})$ の平均は、事前知識がない場合対称性から 0 とすることが多い。
これは、($y(\mathbf{x})$ を決める) パラメータの事前分布 $p(\mathbf{w}|\alpha)$ の平均を 0 にするのと同じことである。

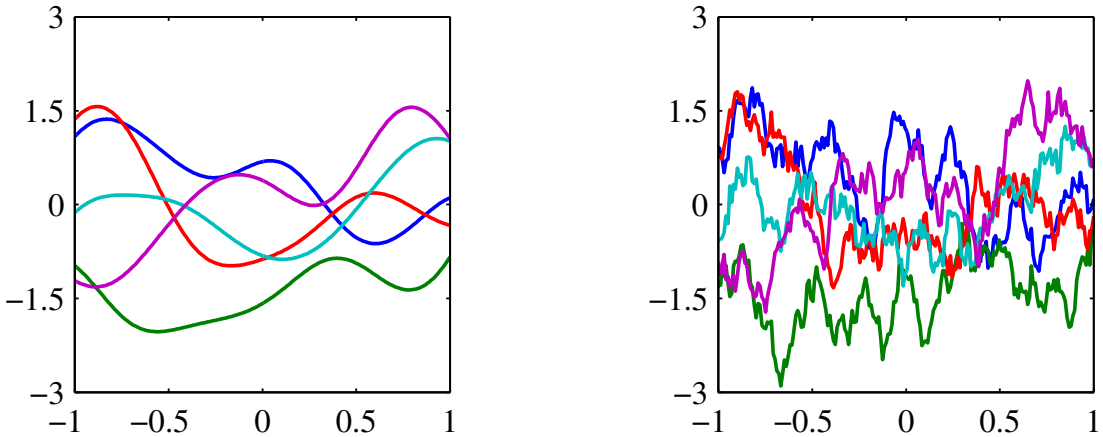
このとき、ガウス過程は、カーネル関数

$$\mathbb{E}[y(\mathbf{x}_n)y(\mathbf{x}_m)] = k(\mathbf{x}_n, \mathbf{x}_m) \tag{6.55}$$

で与えられる共分散によって定まる。
(ガウス過程は $y(\mathbf{x})$ の寄せ集めで、実際の $\mathbf{w}$ はいろんなものがあり得る)

図解

カーネル関数 k は、基底関数 ϕ を決めることで求めることもできるが、実際は直接定義することのほうが多い (たぶん)。下図は、異なる 2 つのカーネル関数で定まるガウス過程からサンプルされた関数を示す。



左：$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2)$ (6.23)
右：$k(x, x') = \exp(-\theta|x - x'|)$ (6.56)

(6.56) は、もともとは**オルンシュタイン-ウーレンベック過程** (Ornstein-Uhlenbeck process) というブラウン運動を記述するために作られたモデル。

より一般的には...

ガウス過程とは、$y(\mathbf{x})$ がガウス分布に従うような確率過程のことである (つまり、関数の形は必ずしも $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ でなくてもよい)。

入力ベクトルが 2 次元のときは、特別に**ガウス確率場** (Gaussian random field) と呼ばれる。

6.4.1 まとめ

- ガウス過程とは、関数の値の集合 $\mathbf{y}$ の同時確率が、ガウス分布に従うような確率過程 (確率変数の寄せ集め) のこと。
- 線形回帰も、パラメータ $\mathbf{w}$ がガウス分布に従うとすると、各データ点 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ が与えられた時の関数の値 $\mathbf{y}$ もガウス分布に従うので、ガウス過程である。
- $y(\mathbf{x})$ を $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ とすると、

$$\mathbb{E}[\mathbf{y}] = \boldsymbol{\Phi}\mathbb{E}[\mathbf{w}] = \mathbf{0} \tag{6.52}$$

$$\mathrm{cov}[\mathbf{y}] = \mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \frac{1}{\alpha}\boldsymbol{\Phi}\Phi^{\mathrm{T}} = \mathbf{K} \tag{6.53}$$

  で表されるような確率過程となる。
- カーネル関数をどんなものにするかによって、サンプリングされる関数の形も変わってくる。

6.4.2 ガウス過程による回帰

出てくる変数：

- t_n :観測変数。$y_n = y(\mathbf{x}_n)$ として、

$$t_n = y_n + \epsilon_n \tag{6.57}$$

  で定義される。まとめて $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$.
- ϵ_n :ノイズ。それぞれの観測値に対して独立に決まる。β をノイズの精度を表す超パラメータとして、

$$p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1}) \tag{6.58}$$

  となる。

N 個の訓練集合をまとめて書く

ノイズは各データ点 ($y(\mathbf{x}) =$ 真の値) に対してランダムに決まるので、(6.58) 式をまとめて書くと、

$$p(\mathbf{t}|\mathbf{y}) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}_N). \tag{6.59}$$

となる。ただし、$\mathbf{I}_N$ は $N \times N$ の単位行列。
ガウス過程の定義 ($\mathbb{E}[\mathbf{y}] = \mathbf{0}$, $\mathrm{cov}[\mathbf{y}] = \mathbf{K}$) より、周辺分布 $p(\mathbf{y})$ は、

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}). \tag{6.60}$$

である。
$\mathbf{K}$ を決めるカーネル関数は、二つの点 $\mathbf{x}_n$ と $\mathbf{x}_m$ が似ているほど $y(\mathbf{x}_n)$ と $y(\mathbf{x}_m)$ の相関が高いようなものが選ばれる (たとえばガウスカーネル)。

データ点のサンプリング

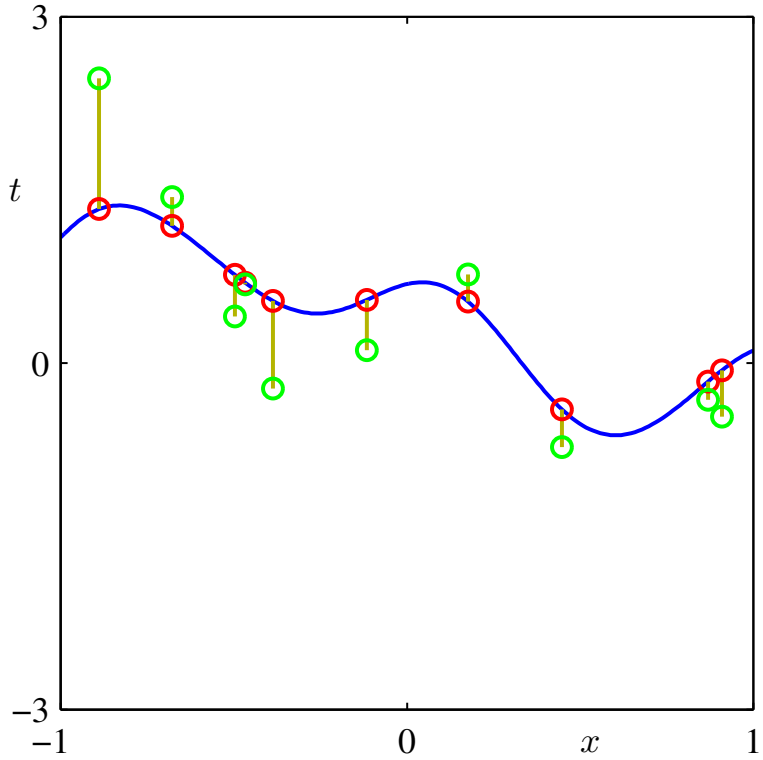


Figure : 実線はガウス過程からサンプリングした関数、赤い点は入力集合 x_n に対応する値 y_n を表す。緑の丸は、y_n にそれぞれ独立にガウスノイズを加えた点 t_n を示している。

観測値の周辺分布

$p(\mathbf{t}|\mathbf{y})$ と $p(\mathbf{y})$ がわかったので、次は周辺分布 $p(\mathbf{t})$ を求めたい。すなわち、(2.115) 式 (x) を用いて、

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\mathrm{y} = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}). \tag{6.61}$$

と求まる。共分散行列 $\mathbf{C}$ の各要素は、

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\gamma_{nm} \tag{6.62}$$

である。$y(\mathbf{x})$ と ϵ が互いに独立であるため、共分散もこの二つを足し合わせるだけで良い。

新しい入力を予測する

$p(\mathbf{t}|\mathbf{y})$ と $p(\mathbf{y})$、$p(\mathbf{y})$ が求まったので、新しい入力ベクトル $\mathbf{x}_{N+1}$ に対する目標変数 t_{N+1} を予測したい。そのためには、予測分布 $p(t_{N+1}|\mathbf{t}_N)$ を求めなければならない ($\mathbf{t}_N = (t_1, \ldots, t_N)^{\mathrm{T}}$)。
このとき、$p(t_{N+1}|\mathbf{t}_N)$ は $\mathbf{t}_N$ だけでなく $\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}$ にも依存しているが、簡単のためにそこは省略。

(6.61) より、$t_1, \ldots, t_{N+1}$ の同時分布は、

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1}). \tag{6.64}$$

このとき、$\mathbf{C}_{N+1}$ は、$(N+1) \times (N+1)$ の共分散行列で、各要素は前頁の (6.62) 式で与えられる。この式から予測分布を求めることができる。

回帰で求める平均と分散

(6.64) 式の $\mathbf{C}_{N+1}$ を分割して、

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^{\mathrm{T}} & c \end{pmatrix} \tag{6.65}$$

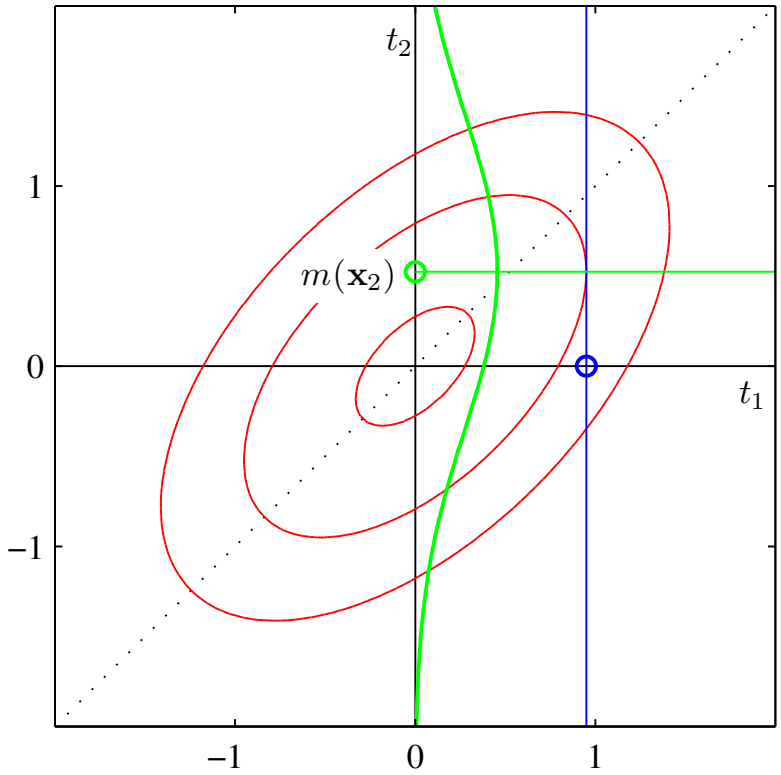
とする。$\mathbf{k}$ は、要素 $k(\mathbf{x}_n, \mathbf{x}_{N+1}(n = 1, \ldots, N)$ を持つベクトルであり、$c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}$ とする。

いろいろ計算すると、$p(t_{N+1}|\mathbf{t})$ は、以下のような平均と共分散を持つガウス分布であることがわかる。

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t} \tag{6.66}$$

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k} \tag{6.67}$$

ガウス過程による回帰：例



Figure： 訓練データとテストデータが 1 つずつの場合のガウス過程による回帰のしくみ。赤色の楕円が、同時分布 $p(t_1, t_2)$ の 等高線を示している。t_1 は訓練データ (青い点)。緑色の線は $p(t_2|t_1)$。楕円を青いところで切るとこんな感じになる。

ガウス過程の適用例

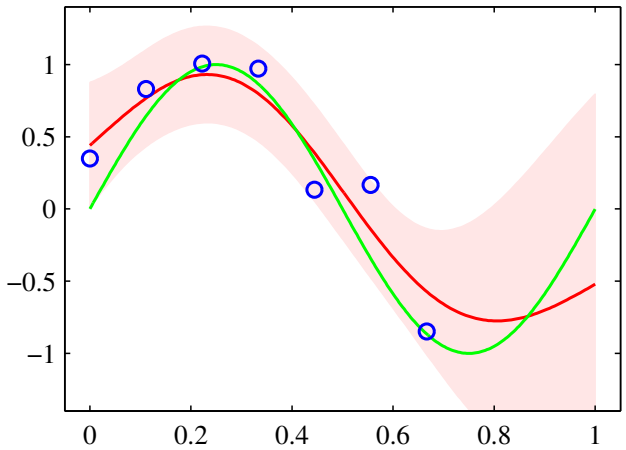
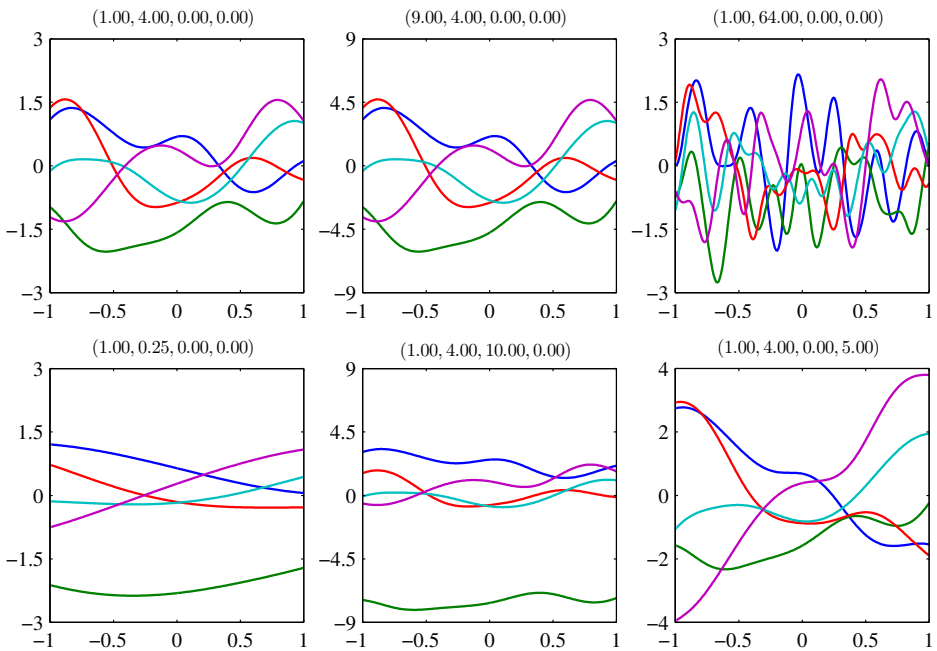


Figure : 赤い線は正弦関数を表している。青い点はそこからガウス分布に従うノイズを加えてサンプリングされたデータ点。緑色の実線がガウス過程による平均、ピンク色の領域がガウス過程による分散 (標準偏差の 2 倍) を表している。データが疎な部分 (右端付近) では不確かさ (分散) が大きくなっているのがわかる。

よく使われるカーネル

ガウス過程回帰に使われるカーネル関数として、以下のようなものがある。
4 つの超パラメータ $(\theta_0, \theta_1, \theta_2, \theta_3)$ をもつ。

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m. \tag{6.63}$$



制約など

カーネル関数は何でも良いが、(6.62) で与えられる共分散行列が正定値でなければならない。

また、予測分布の平均 (6.66) は、$\mathbf{x}_{N+1}$ の関数として

$$m(\mathbf{x}_{N+1}) = \sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x}_{N+1}) \tag{6.68}$$

と表すことができる。ただし、a_n は $\mathbf{C}_N^{-1}\mathbf{t}$ の n 番目の要素。この式を用いると、$\mathbf{K}(N \times N$ 次元) の代わりに M 個の基底関数で表すことができ、$M \times M$ の行列の逆行列を求めればすむ。
このため、データ数 N に比べてモデル数 M が少ないような場合は、計算量を減らすことができる。ただし、カーネル関数の中には無限個の基底関数でしか表せないものがあり、そうしたものにはこの方法は使えない。

6.4.2 まとめ

- ノイズ ϵ_n を加えた観測値 $\mathbf{t}$ から、新たなデータに対する予測分布 $p(t_{N+1}|\mathbf{t})$ を求めることができた。
- 予測分布は、

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t} \tag{6.66}$$

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k} \tag{6.67}$$

  という平均と共分散を持つガウス分布になる。
- $N \times N$ の共分散行列の逆行列 $\mathbf{C}_N^{-1}$ を求めなければならないので、計算時間は $O(N^3)$ となってしまう。基底関数を用いて $\mathbf{S}_N$ の逆行列を求めれば (カーネルトリックは使えないが) 計算量は $O(M^3)$ になる。

6.4.3 超パラメータの学習

ガウス過程による予測は、共分散を決める際の超パラメータに依存している。実際の応用では、あらかじめ超パラメータを決めておくよりも、これを学習させることが多い。
超パラメータの学習は、尤度関数 $p(\mathbf{t}|\boldsymbol{\theta})$ に基づいて行われることが多い。最も単純なアプローチは、対数尤度関数 $\ln p(\mathbf{t}|\boldsymbol{\theta})$ を最大化するような $\boldsymbol{\theta}$ を求めるという方法である。

ガウス過程における対数尤度関数は、

$$\ln p(\mathbf{t}|\boldsymbol{\theta}) = -\frac{1}{2}\ln|\mathbf{C}_N| - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t} - \frac{N}{2}\ln(2\pi) \tag{6.69}$$

で与えられる。

対数尤度関数の最大化

先ほどの続きで、パラメータベクトル $\boldsymbol{\theta}$ の勾配も求める。

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} \tag{C.21}$$

$$\frac{\partial}{\partial x}\ln|\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right) \tag{C.22}$$

を用いて (付録 C 参照)、

$$\frac{\partial}{\partial \theta_i}\ln p(\mathbf{t}|\boldsymbol{\theta}) = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}_N^{-1}\frac{\partial \mathbf{C}_N}{\partial \theta_i}\right) + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}_N^{-1}\frac{\partial \mathbf{C}_N}{\partial \theta_i}\mathbf{C}_N^{-1}\mathbf{t} \tag{6.70}$$

である。ただし、$\ln p(\mathbf{t}|\boldsymbol{\theta})$ は非凸関数なので極大点は複数あり得る。

6.4.3 まとめ

- ガウス過程に現れる超パラメータの学習は、対数尤度の最大化という一般的な手法が用いられることが多い。
- ただし対数尤度関数はカーネル関数のせいで非線形なことが多いので、そうした場合は超パラメータの勾配も求める必要がある。
- ベイズ的な手法を用いて予測することもあるが、その際は近似を用いなければならない ($p(\boldsymbol{\theta})$ や $p(\mathbf{t}|\boldsymbol{\theta})$ を厳密に周辺化することができない)。