おさらい

カーネル法の章では、基底関数の代わりにカーネル関数を用いてモデル化する方法を学んでいる。6.4.1 項までで、カーネル関数の構成法や Nadaraya-Watson モデル、ガウス過程の導入を扱った。この章の残りとしては、ガウス過程の回帰、関連度自動決定、ガウス過程の分類を扱う予定である。

ガウス過程は以下の定義であった。

**「ガウス過程は関数 $y(x)$ 上の確率分布で、$x_1, x_2, \cdots, x_N$ に対する $y(x)$ の値の同時分布がガウス分布となる。」**

教科書の導入では、わかりやすさのために基底関数を使った次のモデルを使ってガウス過程の形を導いた。

モデル : $y(x) = w^T \phi(x)$

事前分布 : $p(w) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$

$x_1, \cdots, x_N$ に対する予測値 $y$ の分布は、次のようなガウス分布になった。

$$y \sim \mathcal{N}(0, K)$$

ここで、$K_{nm} = k(x_n, x_m)$

6.4.2 以降で扱う内容では、ガウス過程を $y \sim \mathcal{N}(0, K)$ を起点として議論を進める。よって、$y(x) = w^T \phi(x)$ や $p(w) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$ が背景に存在しないことに注意したい。すなわち、グラム行列 $K$ の構成方法はこちらで柔軟に設定できる。

### 6.4.2 ガウス過程による回帰

この項で学ぶことは次の2点である。

■ ガウス過程の事前分布におけるモデルのイメージを掴む

■ データが得られた後の予測分布がガウス分布となることを示す

ガウス過程の優れている点は、パラメトリックなモデルに比較して柔軟な表現力がある点である。モデルの制約としてはガウス過程の定義とカーネル関数の設定だけであるのに対し、図6.5で示すような様々な形の回帰関数が表現できる。また、事後分布も図6.8のように当てはまりのよい予測分布を構成できる。

まずは今回の問題設定を確認する。観測される変数にはガウス分布のノイズを仮定する。

$$t_n = y_n + \varepsilon_n$$

（$t_n$：観測、$y_n$：真の値、$\varepsilon_n$：ノイズ）

$$p(t \mid y) = \mathcal{N}(t \mid y, \beta^{-1} I_N) \qquad (6.59)$$

ガウス過程の定義より入力 $x$ に対して、 $y$ の予測分布は次のようになる。

$$p(y) = p(y \mid x) = \mathcal{N}(y \mid 0, K) \qquad (6.60)$$

これを用いて $t$ の分布を求めると、これもガウス分布となる。

$$p(t) = \int p(t, y)\, dy$$
$$= \int p(t \mid y)\, p(y)\, dy$$

（$p(t \mid y)$：ガウス分布(6.59)、$p(y)$：ガウス分布(6.60)）

ここで、上巻(2.115)を使う。

$$\begin{cases} p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1}) \\ p(y \mid x) = \mathcal{N}(y \mid Ax + b, L^{-1}) \end{cases} \quad \text{のとき、}$$

$$p(y) = \mathcal{N}(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

今回、$\mu = 0$, $\Lambda^{-1} = \beta^{-1} I_N$, $A = I_N$, $b = 0$, $L^{-1} = K$ を代入する

$$= \mathcal{N}(t \mid 0, \underbrace{K + \beta^{-1} I_N}_{= C \text{ と置く}}) \qquad (6.61)$$

ここで、 $\mathbb{K}$ を構成するカーネル関数には次のものがよく使われる。

$$k(x_n, x_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|x_n - x_m\|^2\right\} + \theta_2 + \theta_3 x_n^T x_m$$

$\theta_0$ $\theta_1$ $\theta_2$ $\theta_3$ は ハイパーパラメータ.

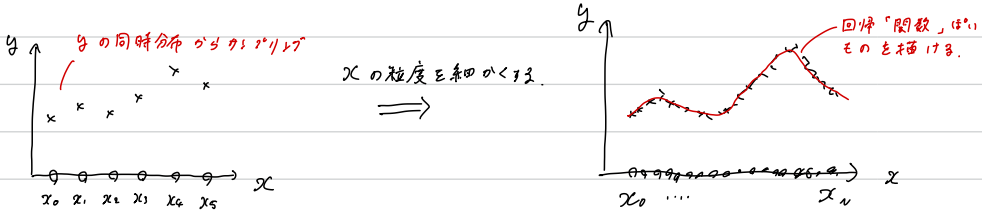
ハイパーパラメータをいくつか変更した場合に、事前分布から**サンプリングされるモデル**を図6.5（次ページ）に示す。

（コメント）

ここで、モデルがサンプリングされるイメージが非常に難解だった。基底関数モデル $y = w^T \phi(x)$ であれば、 $w$ を事前分布からサンプリングすることで、関数 $y(x)$ をサンプリングすることができ、話は簡単である。

一方で、ガウス過程では事前分布から得られるモデルパラメータが存在しない。この場合のモデルのサンプリングとは、入力ベクトル $x$ に対して、 $y \sim N(0, \mathbb{K})$ を直接サンプリングすることに相当するらしい。

＜イメージ＞

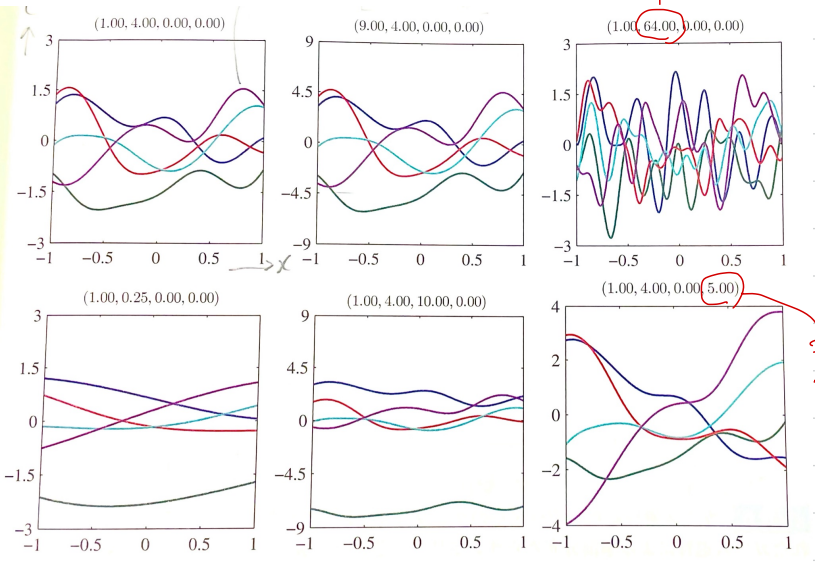



ガウス過程の事前分布からサンプリングされたモデルは、与えられたデータに対する予測値しか返すことができない。なので、図6.5の回帰線は間の点を補完して得られた線と考えらえる。パラメトリックモデルと比較すると、事前分布から得られるモデルが大きく異なることがわかる。

図 6.5

$$k(x_n, x_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|x_n - x_m\|^2\right\} + \theta_2 + \theta_3 x_n^T x_m$$

RBFの部分の精度パラメータが大きい




ここまで、事前分布から得られるモデルを見てきたが、実際の応用では訓練データが得られた元での予測分布が重要である。訓練集合として次のデータが得られたとする。

$$x_1, x_2, \cdots x_N \quad , \qquad t_N = (t_1, \cdots, t_N)^T$$

新たにデータ $x_{N+1}$ が得られたときの $t_{N+1}$ の分布を求めるのが目的である。

$$p(t_{N+1} \mid x_1, \cdots, x_N, x_{N+1}, t_N)$$

これを求めるために、同時分布 $p(t_{N+1}, t_N \mid x_1, \cdots, x_{N+1})$ を求める。これが求まると、目的の条件付き分布もガウス分布の性質から求まるからである。

6.61 式より、

$$p\left(\begin{pmatrix} t_{N+1} \\ t_N \end{pmatrix} \mid x_1, \cdots, x_{N+1}\right) = \mathcal{N}\left(\begin{pmatrix} t_{N+1} \\ t_N \end{pmatrix} \middle| 0, \begin{pmatrix} c & k^T \\ k & C_N \end{pmatrix}\right) \qquad (6.64 \text{ 改})$$

ここで、

$$\begin{cases} k^T = (k(x_{N+1}, x_1), \cdots, k(x_{N+1}, x_N)) \\ c = k(x_{N+1}, x_{N+1}) + \beta^{-1} \end{cases}$$

PRML上巻の2.81、2.82式を用いると、目的の条件付き分布が求まる。2.81、2.82式をおさらいする。

$$p\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} \middle| \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right)$$

のとき、

$$p(x_a \mid x_b) = \mathcal{N}(x_a \mid \mu_{a|b}, \Sigma_{a|b})$$

$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$ (2.81)

$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$ (2.82)

(6.64改) 式にこれを適用すると、目的の条件付き分布が導かれる。

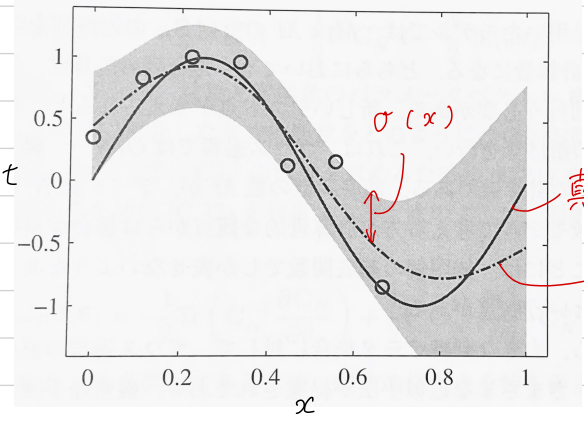
$$p(t_{N+1} \mid x_1, \cdots, x_{N+1}, \mathbf{t}_N) = \mathcal{N}(t_{N+1} \mid m(x_{N+1}), \sigma^2(x_{N+1}))$$

$t_{N+1}$ が $x_a$、$\mathbf{t}_N$ が $x_b$ に相当

$m(x_{N+1}) = 0 + \mathbf{k}^T \mathbf{C}_N^{-1}(\mathbf{t} - 0) = \mathbf{k}^T \mathbf{C}_N^{-1}\mathbf{t}$ (6.66)

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

平均 $m(x_{N+1})$ と分散 $\sigma^2(x_{N+1})$ は $x_{N+1}$ の関数であり、これをプロットすると図6.8の予測分布が得られる。



最後に、ガウス過程を利用するための制約について確認する。

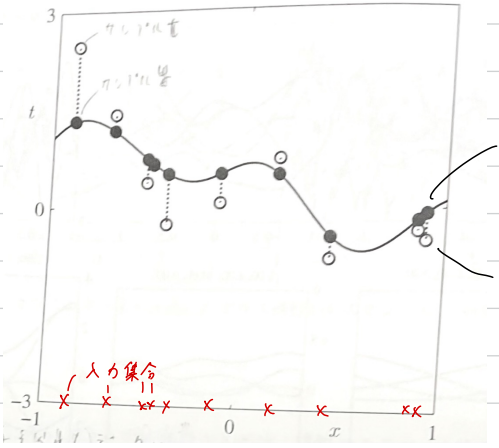
まず、ガウス過程により得られた $t$ の確率分布 $N(0, C)$ の共分散行列 $C$ が正定値行列であるかを確かめる。これはグラム行列 $K$ が半正定値であることから確かめられる。

$K$ の固有値を $\lambda_i$ とすると、$C$ の固有値は $\lambda_i + \beta^{-1}$ と表される。ここで、カーネル関数の必要十分条件が、任意の $\{x_n\}$ に対するグラム行列が半正定値行列であるという事実を使う（p.5 下段）と、$\lambda_i \geq 0$ すなわち $\lambda_i + \beta^{-1} > 0$ となる。これより、$C$ は正定値行列である。

次に、$m(x_{N+1})$ と $\sigma^2(x_{N+1})$ の計算に必要な、$C_N$ の逆行列の計算量を考える。一般的な逆行列の計算では、$O(N^3)$ の計算量となるため、データが多量な場合にこの計算は現実的ではなくなる。その代わりに近似的な手法がいくつか提案されている (Gibbs, 1997; Trespass, 2001; …)。

文章の流れとして飛ばした図6.6 と図6.7 についても触れておく。

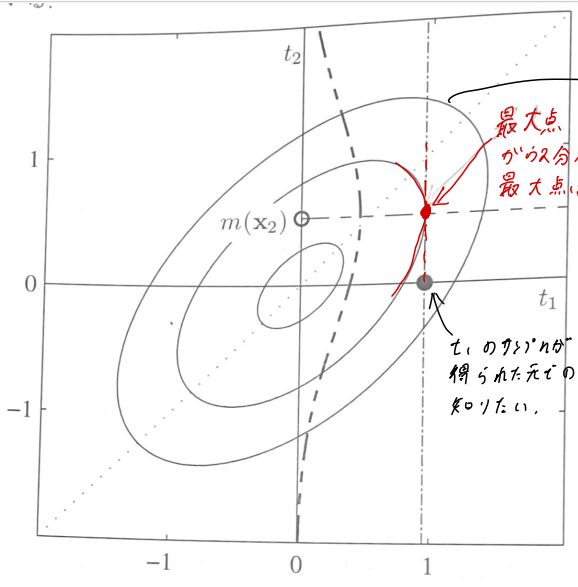
図6.6 事前分布からサンプリングされる $y$ と $t$ の例




黒点：ガウス過程 $N(y|0, K)$ から得られたサンプル $y$

白点：$p(t|y) = N(t|y, \beta^{-1}I_N)$ から得られたサンプル $t$

図6.7　$p(t_1, t_2)$ から　$p(t_2 \mid t_1)$ が得られる過程を描いた図




楕円：$p(t_1, t_2)$ の等高線
（ガウス分布とする）

最大点
ガウス分布における
最大点はつまり期待値 $m(x_2)$

$t_1$ のサンプルが
得られた元での $p(t_2 \mid t_1)$ を
知りたい．

### 6.4.3 超パラメータの学習

ガウス過程の予測は、共分散関数（行列） $C$ に依存しており、いくつかの超パラメータ $\theta_0, \theta_1, \theta_2, \theta_3, \beta$ に依存している。この節では超パラメータを学習する方法を学ぶ。

本来であれば、パラメータ集合 $\theta$ の事前分布を仮定して、$p(t) = \int p(t|\theta)p(\theta)d\theta$ を予測分布として扱うべきである。しかしこれを解析的に扱うのは難しい。そこで、対数尤度 $p(t|\theta)$ を最大化するパラメータ $\hat{\theta}$ を用いて、$p(t) = p(t|\hat{\theta})$ を予測分布とする方法がよく用いられる。これは第二種の最尤推定（エビデンス近似）と呼ばれる（上巻p.164）。

$p(t|\theta)$ はエビデンス関数と呼ばれる。これを最大化する超パラメータの推定は、3.5.1、3.5.2項で扱っているので、読み合わせとしたい。

ガウス過程におけるエビデンス関数は（6.61）式からガウス分布の式で与えられる。

$$\ln p(t|\theta) = -\frac{1}{2}\ln|C_N| - \frac{1}{2}t^T C_N^{-1} t - \frac{N}{2}\ln(2\pi)$$

教科書の付録「行列の微分」の（C.22）（C.21）式を用いると勾配は次のように表される。

$$\frac{\partial}{\partial\theta_i}\ln p(t|\theta) = -\frac{1}{2}\mathrm{Tr}\left(C_N^{-1}\frac{\partial C_N}{\partial\theta_i}\right) + \frac{1}{2}t^T C_N^{-1}\frac{\partial C_N}{\partial\theta_i}C_N^{-1}t$$

これを用いて、なんらかの繰り返し手法によって $\theta$ の最尤解を取得する。