# 1　ガウス過程

前回のおさらいから。セクション 6.4.2 ではガウス過程による回帰を考えて予測分布を構成することを考えた。そしてその予測分布の期待値や共分散行列がカーネル関数を用いた形で表せることを導いた（式 6.65-7）。

今回はガウス過程による分類を考えて予測分布を構成するための準備をしていく。そして分類の場合にも予測分布の期待値や共分散行列がカーネル関数を用いた形で表せることを確かめる[*1]。

## 1.1　関連度自動決定

冒頭の説明と反してしまうが、まだこのサブセクションでは分類問題を考えない。位置付けとしては 6.4.3 の続きとなるが、自分の理解では「ARD」というものの存在を紹介しているサブセクションと認識している。

先のサブセクションではスケールパラメタを求めるために最尤推定を用いた。本サブセクションも最尤推定の枠組みの中で、入力間の相対的な重要度[*2]についてガウス過程に対する ARD（Automatic Relevance Determination；関連度自動決定）という手法を例にとり学んでいくことにする。

まずは本手法が何をするものなのか具体的な式を導入しながら理解したい[*3]。今 2 次元のガウス過程を考える。カーネル関数を次の形で与える。

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_0 \exp\left[-\frac{1}{2}\sum_{i=1}^{2}\eta_i(x_i - x_i')^2\right] \tag{1}$$

ここで $\eta_i$ は分散に相当するパラメタである。このパラメタの組み合わせについて 2 種類考えて $y(\boldsymbol{x})$ をサンプリングしたものをプロットすると図 6.9 のようになる。左肩下がりの軸が $x_1$、右肩上がりの軸が $x_2$ に対応している。パラメタの取り方としては、左図：$(\eta_1, \eta_2) = (1, 1)$、右図：$(\eta_1, \eta_2) = (1, 0.01)$ と約束している。

この図から小さい分散パラメタに対応する入力変数、今回で言えば $\eta_2 = 0.01$ に対応する $x_2$ は予測分布への寄与が小さくなり、結局不要な入力変数になるということが言える（らしい）[*4]。このように最尤推定の枠組みで、不要な入力変数を取り除く手法のことを ARD と呼んでいる（らしい）。

なお、図 6.10 は読み合わせとする[*5]。

## 1.2　ガウス過程による分類

ここからが冒頭で述べた本編となる。本サブセクションではガウス過程を分類問題に適用できるようなフレームワークを紹介する。訓練データ $\boldsymbol{t}_N$ が与えられたもとでの予測分布 $p(t_{N+1} = 1 \mid \boldsymbol{t})$ を決定することが目的となる[*6]。予測分布を構成するための準備においては、シグモイド関数による非線形変換を用いることがポイントとなる。

---

[*1] 予測分布の最終形に辿り着くのは次回の範囲となる。本文にもある通り、本レジュメではその導入部分までを取り扱う。

[*2] 教科書にも書かれているこのぼんやりしたキーワード。初見では意味が掴めなかったが、後の内容を読むとなんとなくわかる。端的に言ってしまえば、この後 2 つの入力変数のうち予測に使うもの、使わないものを取捨選択することになるのだが、まさにそのことを表現したキーワードになっている。もっとわかりやすく具体的に説明した例については https://www.slideshare.net/KeisukeSugawara/slide0629 の頁 50 を参照されたい。

[*3] 実はこれを理解することは、本手法に「自動」という名前が組み込まれている背景を理解することに他ならないことがわかる。

[*4] 正直よくわからんが、なんとなくスパース性とのアナロジーと認識している

[*5] なんとも煮えきれない説明となってしまった。ただし、7 章で ARD についてもう少し詳しく取り扱うらしいのでそこで頑張ろう。

[*6] 入力変数も当然与えられる。ここでは $\boldsymbol{a}_{N+1} = (a(\boldsymbol{x}_1), \ldots, a(\boldsymbol{x}_{N+1}))$ として与えられているが、表記を簡便にするため条件付き部分には書かないものとする。

まず導入として、先にも述べたシグモイド関数による非線形変換を用いるモチベーションについて説明する。そもそも、ガウス過程のモデルの予測は実数値全体を取るため分類には適さない[*7]。分類を行うためにはクラスに所属する確率が $(0,1)$ に収まる必要がある。そこでガウス過程で得られる $a(\boldsymbol{x})$[*8]をロジスティックシグモイド関数 $y=\sigma(a)$ として非線形変換することにより、確率の値を $(0,1)$ に収めることを考える。これで分類問題を取り扱えることができるようになる。

この変換によってガウス過程が非ガウス過程に移り変わる様子を図 6.11 で確かめてほしい。変換後（右図）では確かに縦軸のスケールが $(0,1)$ になっていることがわかる。

ガウス過程で定義される $a(\boldsymbol{x})$ をシグモイド変換して得られる $y$ に関する確率密度関数は、2 クラス分類の問題の場合、次のようなベルヌーイ分布で書ける。

$$p(t \mid a) = \sigma(a)^t (1-\sigma(a))^{1-t} \tag{2}$$

ここでベルヌーイ分布を書き下したことと後に予測分布を導出することとは次のように結びついていると認識している。

・式 2 は予測分布の表式（6.76）における $p(t_{N+1}=1 \mid a_{N+1})$ を求めるために使う。

・今ベルヌーイ分布の形で書いたことでクラス 1 とクラス 0 の両方を表現した。

・式（6.76）の $p(t_{N+1}=1 \mid a_{N+1})$ はクラス 1 となる確率の値、つまり $t_{N+1}=1$ を満たすシグモイド関数の値となる。

・最終的にはクラス 1 となる確率の値をとってくればよいが、ここで先に両クラスとも表現しておきたかった。

以上、4 点踏まえたところで、この段階においてベルヌーイ分布の表式を導入したと理解している。ちなみに、クラス 0 の確率は余事象的な考え方によって自明に求められることも教科書で言及されている。

非線形関数による変換を用いるという導入は終わったので、ここからは本サブセクションの目的であった予測分布を構成していくことを考えよう。予測分布を構成するためのフレームワークは、これまでに何度も学んでいるが簡単にポイントだけ整理すると

・予測したい目標変数についての条件付き分布を書き下す。

・新しい入力変数の事後分布とのたたみ込みによって予測分布の表式を周辺化積分の形で表す。

の 2 つであった。

今、新しい入力変数の事後分布を求めるために事前分布 $p(\boldsymbol{a}_{N+1})$ を導入しておく。

$$p(\boldsymbol{a}_{N+1}) = N(\boldsymbol{a}_{N+1} \mid \boldsymbol{0}, \boldsymbol{C}_{N+1}) \tag{3}$$

と書ける。ここで共分散行列 $\boldsymbol{C}_{N+1}$ の要素を $C(\boldsymbol{x}_n, \boldsymbol{x}_m) = k(\boldsymbol{x}_n, \boldsymbol{x}_m) + \nu\delta_{nm}$ と約束する。共分散行列の正定値性を保証するためにノイズライクな項 $\nu\delta_{nm}$ を加えている。この事前分布の表式 3 は後で入力の事後分布 $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ を求めるために用いるので、ここで準備したものである。なお、後に確かめるが $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ が求まると新しい入力変数の事後分布が求まることがわかる。

所望の予測分布が $p(t_{N+1}=1 \mid \boldsymbol{t}_N)$ は次のような形式で書ける。

$$p(t_{N+1}=1 \mid \boldsymbol{t}_N) = \int p(t_{N+1}=1 \mid a_{N+1}) p(a_{N+1} \mid \boldsymbol{t}_N)\, da_{N+1} \tag{4}$$

これは新しい入力変数 $a_{N+1}$ で周辺化積分を行うことで求められる。

---

[*7] 全実数値をとるガウス過程に従う確率分布からサンプリングしてくるのだから当然のことである。

[*8] $a$ であることに注意。$y$ ではない。

しかしながらこの積分計算は解析的に実行できない[*9]。ではどうするか？それを次のサブセクションで考えたい。

## 1.3 ラプラス近似

先のセクションで導いた予測分布の積分表式 4 に対して近似を施すことで解析的な形を得る。ここでは近似手法としてラプラス近似を用いることにする。

今回ラプラス近似においては、入力の事後分布に相当する $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ の部分をガウス分布で近似してやることが求められる。そのためには事後ガウス分布のモードを求めなければならない。

モードを求めるために勾配計算が必要となるが、本レジュメにおいてはその勾配計算の直前である事後分布の対数表示を求めるところまでを担当する。その後の流れの概略については本レジュメの最後に記載しておく。

まず、準備として新しい入力に対する事後分布を次のように変形しておく。変形自体は初等的であるため省略する。

$$p(a_{N+1} \mid \boldsymbol{t}_N) = \cdots = \int p(a_{N+1} \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)\ d\boldsymbol{a}_N \tag{5}$$

被積分関数について $p(a_{N+1} \mid \boldsymbol{a}_N)$ と $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ の二つの部分に分けて考えることにする。

・$p(a_{N+1} \mid \boldsymbol{a}_N)$ について

これは頁 19 下部で得た条件付き分布 $p(t_{N+1} \mid \boldsymbol{t})$ についての結果（6.66-67）において、$t$ を $a$ に置き換えるだけで簡単に求められる。詳細は割愛。

・$p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ について

事後分布 $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ を求めるにはベイズの定理を用いる。そのために準備として事前分布 $p(\boldsymbol{a}_N)$ と条件付き分布 $p(\boldsymbol{t}_N \mid \boldsymbol{a}_N)$ を導入する。

まず事前分布 $p(\boldsymbol{a}_N)$ を考えよう。これは式 3 から自明である。ただし $N+1$ を $N$ に置き換えよ。

つぎに条件付き分布 $p(\boldsymbol{t}_N \mid \boldsymbol{a}_N)$ について考える。各データ点が独立であるという仮定のもとで、式 2 の積を取ることで

$$p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) = \Pi_{n=1}^{N} \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n} \tag{6}$$

という形で表現できる。ただし、各 $a$ や $t$ に下付き添字を記しておくことに注意せよ。これをさらに変形して

[*9] 一般的にシグモイド関数の積分を実行することは難しいと認識している。もちろん必ずそうであるわけではないだろうが。

いく。

$$
\begin{aligned}
p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) &= \Pi_{n=1}^{N} \sigma(a_n)^{t_n} (1-\sigma(a_n))^{1-t_n} \\
&= \Pi_{n=1}^{N} \left( \frac{1}{1+\exp(-a_n)} \right)^{t_n} \left( 1 - \frac{1}{1+\exp(-a_n)} \right)^{1-t_n} \quad (\because (4.59)) \\
&= \Pi_{n=1}^{N} \left( \frac{1}{1+\exp(-a_n)} \right)^{t_n} \left( \frac{1+\exp(-a_n)-1}{1+\exp(-a_n)} \right)^{1-t_n} \\
&= \Pi_{n=1}^{N} \left( \frac{1}{1+\exp(-a_n)} \right)^{t_n} \left( \frac{\exp(-a_n)}{1+\exp(-a_n)} \right)^{1-t_n} \\
&= \Pi_{n=1}^{N} \frac{(\exp(-a_n))^{1-t_n}}{(1+\exp(-a_n))^{t_n+1-t_n}} \\
&= \Pi_{n=1}^{N} \frac{(\exp(-a_n))^{1-t_n}}{1+\exp(-a_n)} \\
&= \Pi_{n=1}^{N} \frac{(\exp(-a_n))^{-t_n}}{1+\exp(a_n)} \\
&= \Pi_{n=1}^{N} \exp(a_n t_n) \frac{1}{1+\exp(a_n)} \\
&= \Pi_{n=1}^{N} \exp(a_n t_n) \sigma(-a_n) \ (\because (4.59)) \qquad (7)
\end{aligned}
$$

いささか冗長な式変形となってしまったが、これで条件付き分布も陽な形で書くことができた。

以上により、事後分布 $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ はベイズの定理を適用することで求められるが、計算の都合のため、これに対数をかました形で表すことにしよう。この対数をとった形式について正規化項を無視したものを $\Psi(\boldsymbol{a}_N)$ とする。

$$
\begin{aligned}
\Psi(\boldsymbol{a}_N) &\propto \ln p(\boldsymbol{a}_N) + \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) \\
&= \ln N(\boldsymbol{a}_N \mid \boldsymbol{0}, \boldsymbol{C}_N) + \ln \Pi_{n=1}^{N} \exp(a_n t_n) \sigma(-a_n) \\
&= \ln\{ \left[ \frac{1}{(2\pi)^{N/2} \mid \boldsymbol{C}_N \mid} \right] \exp\left[ -\frac{\boldsymbol{a}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{a}_n}{2} \right] \} + \sum_{n=1}^{N} t_n a_n - \sum_{n=1}^{N} (1+\exp(a_n)) \\
&= -\frac{1}{2} \boldsymbol{a}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{a}_n - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln \mid \boldsymbol{C}_N \mid + \boldsymbol{t}_N^T \boldsymbol{a}_N - \sum_{n=1}^{N} (1+\exp(a_n)) \qquad (8)
\end{aligned}
$$

これが求めるべき事後分布である。ただしベイズの定理を適用する際の分母部分は無視したとなっていることに注意せよ。もちろん、この後の計算には影響がないためこのようにして構わない。

ラプラス近似を適用するためには、この $\Psi(\boldsymbol{a}_N)$ について勾配をとりモードを求める必要がある。その計算については次回の輪講にお任せしたいと思う。