### 6.4.6 ラプラス近似（続）

前回はガウス過程の分類問題を扱った。その中で、予測分布を積分の形式で求め、積分を使わない表現を得るためにラプラス近似を導入するところまで行った。今回はラプラス近似を引き続き行い、予測分布の最終形を導く。その後、超パラメータの最適化と、ニューラルネットワークとの関係を考察する。

分類問題のガウス過程の予測分布は次のような形になった。

$$p(t_{N+1}|\mathbf{t}_N) = \int p(t_{N+1}=1|a_{N+1})\, p(a_{N+1}|\mathbf{t}_N)\, da_{N+1} \quad (6.76)$$

（$p(t_{N+1}=1|a_{N+1})$：ベルヌーイ分布なので $\sigma(a_{N+1})$　／　$p(a_{N+1}|\mathbf{t}_N)$：ガウス分布で近似する。）

重畳積分の近似公式（4.153）を使う。

$$= \int p(a_{N+1}|a_N)\, p(a_N|\mathbf{t}_N)\, da_N \quad (6.77)$$

（$p(a_{N+1}|a_N)$：ガウス分布（∵ 回帰のガウス過程）(6.66)(6.67)　／　$p(a_N|\mathbf{t}_N)$：ガウス分布（∵ ラプラス近似））

ガウス分布の周辺化の公式を使う（2.115）

$\log p(a_N|\mathbf{t}_N)$ をモード付近で2次近似する。

$$= \log p(a_N) + \log p(\mathbf{t}_N|a_N) + \text{const.}$$

$$= -\frac{1}{2} a_N^T C_N^{-1} a_N - \frac{1}{2}\ln|C_N| + \mathbf{t}_N^T a_N - \sum_{n=1}^{N} \ln(1+e^{a_n}) + \text{const.}$$

$\Psi(a_N)$ とおく。

つまり、 $\Psi(a_N)$ のモードの値とモード付近の2次微分の値が求まると、芋づる式に予測分布の形が得られるのである。まずモードから求める。

$$\nabla\Psi(a_N) = -C_N^{-1} a_N + \mathbf{t}_N - \left( \sum_{n=1}^{N} \frac{e^{a_n}}{1+e^{a_n}} \delta_{1n} \quad \sum_{n=1}^{N} \frac{e^{a_n}}{1+e^{a_n}} \delta_{2n} \quad \cdots \right)^T$$

$$= -C_N^{-1} a_N + \mathbf{t}_N - \left( \frac{e^{a_1}}{1+e^{a_1}} \quad \frac{e^{a_2}}{1+e^{a_2}} \quad \cdots \quad \frac{e^{a_N}}{1+e^{a_N}} \right)^T$$

$$= -C_N^{-1} a_N + \mathbf{t}_N - \left( \frac{1}{1+e^{-a_1}} \quad \frac{1}{1+e^{-a_2}} \quad \cdots \quad \frac{1}{1+e^{-a_N}} \right)^T$$

$$= -C_N^{-1} a_N + \mathbf{t}_N - \left( \sigma(a_1) \quad \sigma(a_2) \quad \cdots \quad \sigma(a_N) \right)^T$$

$$= -C_N^{-1} a_N + \mathbf{t}_N - \sigma(a_N)$$

新しく定義する。

$\nabla \Psi(a_N) = 0$ を満たす $a_N$ を求めたいが、$\sigma(a_N)$ が $a_N$ に依存するため陽に求めることができない。そこで**反復再重み付け最小二乗法**を用いる。

> ～おさらい：反復再重み付け最小二乗法～
>
> ニュートン・ラフソン法（a.k.a. ニュートン法）において、ヘッセ行列が更新のたびに変化する場合の更新手順のことを言う。ニュートン法の更新式は次の形である。
>
> $$w^{(new)} = w^{(old)} - H^{-1} \nabla E(w^{old})$$
>
> ここで、ヘッセ行列 $H$ が $w$ に依存する場合には、更新のたびに $H$ も更新する必要がある。$H$ を「重み」と考えると、都度重みを更新していることに相当するため、**反復再重み付け**最小二乗法と呼ばれる。(ref. Chapter.4/210228輪講.pdf)

ニュートン法に必要な2次微分 $\nabla\nabla\Psi(a_N)$ を求める。これは既に記載したように、ラプラス近似での分散にも再利用される。

$$\begin{aligned}
\nabla\nabla\Psi &= -C_N^{-1} - \nabla\sigma(a_N) \\
&= -C_N^{-1} - \left(\frac{\partial}{\partial a_N}\sigma(a_1)^T \quad \frac{\partial}{\partial a_N}\sigma(a_2)^T \quad \cdots \quad \frac{\partial}{\partial a_N}\sigma(a_N)^T\right)^T \\
&= -C_N^{-1} - \begin{pmatrix} \frac{d}{da_1}\sigma(a_1) & & 0 \\ & \ddots & \\ 0 & & \frac{d}{da_N}\sigma(a_N) \end{pmatrix} \\
&= -C_N^{-1} - \begin{pmatrix} \sigma(a_1)(1-\sigma(a_1)) & & 0 \\ & \ddots & \\ 0 & & \sigma(a_N)(1-\sigma(a_N)) \end{pmatrix} \quad \left(\because \frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))\right) \\
&= -C_N^{-1} - W_N
\end{aligned}$$

得られた勾配とヘッセ行列をニュートン法の更新式に代入すると次の更新式を得る。（演習6.25）

$$\begin{aligned}
a_N^{new} &= a_N - (\nabla\nabla\Psi(a_N))^{-1} \nabla\Psi(a_N) \\
&= a_N + (C_N^{-1} + W_N)^{-1}(t_N - \sigma(a_N) - C_N^{-1} a_N) \\
&= (C_N^{-1} + W_N)^{-1}\{(C_N^{-1} + W_N) a_N + t_N - \sigma(a_N) - C_N^{-1} a_N\}
\end{aligned}$$

$$a_N^{new} = \left( \mathbb{I} C_N^{-1} + W_N C_N C_N^{-1} \right)^{-1} \left\{ C_N^{-1} a_N + W_N a_N + t_N - \sigma(a_N) - C_N^{-1} a_N \right\}$$

（$\mathbb{I}$ を作り出す）

$$= \left\{ (\mathbb{I} + W_N C_N) C_N^{-1} \right\}^{-1} \left\{ t_N - \sigma(a_N) + W_N a_N \right\}$$

$$(AB)^{-1} = B^{-1} A^{-1}$$

$$= C_N (\mathbb{I} + W_N C_N)^{-1} \left\{ t_N - \sigma(a_N) + W_N a_N \right\} \quad (6.83)$$

これより、ニュートン法（今回は反復再重み付け最小二乗法でもある）の更新式を得られたので繰り返し計算によって、モード $a_N^*$ と負のヘッセ行列 $H = -\nabla\nabla\Psi = W_N + C_N^{-1}$ が算出される。これより、目的の確率分布 $p(a_N | t_N)$ は次のガウス分布となる。（上巻 p.215, (4.132)式）

$$p(a_N | t_N) = \mathcal{N}(a_N | a_N^*, H^{-1})$$

最後に、予測分布が得られるまでのロードマップを振り返る。我々は(6.77) 式の $p(a_N | t_N)$ を求めることができたので、予測分布 $p(t_{N+1} | t_N)$ を得ることができた。

$$p(t_{N+1} | t_N) = \int p(t_{N+1} = 1 | a_{N+1}) p(a_{N+1} | t_N) \, da_{N+1} \quad (6.76)$$

（$p(t_{N+1} = 1 | a_{N+1})$：ベルヌーイ分布なので $\sigma(a_{N+1})$ ／ $p(a_{N+1} | t_N)$：ガウス分布で近似する。）

重畳積分の近似公式（4.153）を使う。

$$\Downarrow$$

$$\int p(a_{N+1} | a_N) p(a_N | t_N) \, da_N \quad (6.77)$$

（$p(a_{N+1} | a_N)$：ガウス分布（∵回帰のガウス過程）(6.66)(6.67) ／ $p(a_N | t_N)$：ガウス分布（∵ラプラス近似））

ガウス分布の周辺化の公式を使う（2.115）

※ 最終形を露わな形で書きたい気持ちはあったけど、答え合わせができないのでここで止めます…

最後にハイパーパラメータ $\theta$ の最適化を扱う。基本的なアイデアとしては「6.4.3 超パラメータの学習」同様に、尤度関数を最大にする最尤推定量を求める。尤度関数は次のように表される。

$$p(t_N | \theta) = \int p(t_N | a_N) p(a_N | \theta) \, da_N$$

（$p(t_N | a_N)$：ベルヌーイ ／ $p(a_N | \theta)$：ガウス）

（感想）

6.90 式の導出できませんでした。困った点は次の３つ

1. (6.90) に $H = W_N + C_N^{-1}$ や $a_N^*$ が現れているので $p(a_N \mid t_N)$ を使いそうだが、積分形式以外に $p(a_N \mid t)$ の分布が現れる形に変形できない。
2. 積分の代わりにラプラス近似を行うとこのとだが、$t_N$ の分布はベルヌーイ分布なのでガウス分布に近似して良いのか怪しい。
3. 積分の中身の $p(a_N \mid \theta)$ にのみラプラス近似を適用しそうに思えたが、重畳積分の近似公式（4.151 - 4.155）の形が出てこないので違うっぽい。

6.90式が次のように導出できたとする。

$$\ln p(t_N \mid \theta) = \Psi(a_N^*) - \frac{1}{2}\ln\left|W_N + C_N^{-1}\right| + \frac{N}{2}\ln(2\pi)$$

$$\text{ここで } \Psi(a_N^*) = \ln p(a_N^* \mid \theta) + \ln p(t_N \mid a_N^*)$$

これの $\theta$ に関する停留点を求めることで、超パラメータの最適化を行う。

$$\left(\frac{\partial \ln p(t_N \mid \theta)}{\partial \theta_i}\right) = \sum_{n=1}^{N} \frac{\partial \Psi(a_N^*)}{\partial a_n^*}\cdot\frac{\partial a_n^*}{\partial \theta_i} - \frac{1}{2}\sum_{n=1}^{N}\frac{\partial \ln\left|W_N + C_N^{-1}\right|}{\partial a_n^*}\frac{\partial a_n^*}{\partial \theta_i} + 0$$

$\frac{\partial \Psi(a_N^*)}{\partial a_n^*} = 0$ になる？（理由）$a_N^*$ は $\frac{\partial \Psi(a_N)}{\partial a_N} = 0$ より得た

(6.91) 式で扱う。

$$= -\frac{1}{2}\sum_{n=1}^{N}\frac{\partial \ln\left|W_N + C_N^{-1}\right|}{\partial a_n^*}\frac{\partial a_n^*}{\partial \theta_i}$$

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

$$= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left\{(W_N + C_N^{-1})^{-1}\frac{\partial(W_N + C_N^{-1})}{\partial a_n^*}\right\}\frac{\partial a_n^*}{\partial \theta_i}$$

$C_N$ は $\{C_N\}_{ij} = k(x_i, x_j) + \beta^{-1}\delta_{ij}$ より、$a_n^*$ に依存しない。

$$= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left\{(W_N + C_N^{-1})^{-1}\frac{\partial W_N}{\partial a_n^*}\right\}\frac{\partial a_n^*}{\partial \theta_i}$$

$\{C_N^{-1}(I + C_N W_N)\}^{-1} = (I + C_N W_N)^{-1} C_N$

$\frac{\partial}{\partial a_n^*}\begin{pmatrix}\sigma_1^*(1-\sigma_1^*) & & 0\\ & \ddots & \\ 0 & & \sigma_N^*(1-\sigma_N^*)\end{pmatrix} = \begin{pmatrix}0 & & & 0\\ & \ddots & & \\ & & \frac{\partial}{\partial a_n^*}\sigma_n^*(1-\sigma_n^*) & \\ & & & \ddots \\ 0 & & & 0\end{pmatrix}$ $(n,n)$ の成分以外は 0

ただし、$\sigma_n^* = \sigma(a_n^*)$ とおく。

$\mathrm{Tr}(\,)$ をとると $(n,n)$ 成分のみ残る。

$$= -\frac{1}{2}\sum_{n=1}^{N}\left[(I + C_N W_N)^{-1} C_N\right]_{nn}\frac{\partial}{\partial a_n^*}\sigma_n^*(1-\sigma_n^*)\frac{\partial a_n^*}{\partial \theta_i}$$

$$\left( \frac{\partial \ln p(t_N|\theta)}{\partial \theta_i} \right) = -\frac{1}{2} \sum_{n=1}^{N} \left[ (I + C_N W_N)^{-1} C_N \right]_{nn} \frac{\partial}{\partial a_n^*} (\sigma_n^* - \sigma_n^{*2}) \frac{\partial a_n^*}{\partial \theta_i}$$

$$= -\frac{1}{2} \sum_{n=1}^{N} \left[ (I + C_N W_N)^{-1} C_N \right]_{nn} \sigma_n^* (1 - \sigma_n^*)(1 - 2\sigma_n^*) \frac{\partial a_n^*}{\partial \theta_i} \quad (6.92)$$

(6.94)で示す。

（感想）

勾配を (6.92) で示してしまったけど、教科書だと (6.91) 式を挟んでいる点で異なる。(6.91) をどのように導出したかわからなかったので相談したい。（6.92）で現れた $\frac{\partial a_n^*}{\partial \theta_i}$ を更に計算する。

$$\frac{\partial a_N^*}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left\{ C_N (t_N - \sigma_N) \right\} \quad (\because 6.84 : a_N^* = C_N (t_N - \sigma_N))$$

$$= \frac{\partial C_N}{\partial \theta_i} (t_N - \sigma_N) - C_N W_N \frac{\partial a_N^*}{\partial \theta_i} \quad (6.93)$$

(6.93)ポイント抜けてますね？

$$\Leftrightarrow \frac{\partial a_N^*}{\partial \theta_i} + C_N W_N \frac{\partial a_N}{\partial \theta_i} = \frac{\partial C_N}{\partial \theta_i} (t_N - \sigma_N)$$

$$\Leftrightarrow (I + C_N W_N) \frac{\partial a_N}{\partial \theta_i} = \frac{\partial C_N}{\partial \theta_i} (t_N - \sigma_N)$$

$$\Leftrightarrow \frac{\partial a_N}{\partial \theta_i} = (I + C_N W_N)^{-1} \frac{\partial C_N}{\partial \theta_i} (t_N - \sigma_N) \quad (6.94)$$

ここも教科書と異なる

全く教科書通りにうまく行かなかったが、とりあえず勾配が得られたのでこれを元に何らかの最適化法を適用して最適なハイパーパラメータを取得する。

### 6.4.7 ニューラルネットワーク（NN）との関係

カラム的な内容なので、各々読むこととする。主張としては次の通り。

■ NNはデータサイズに応じてユニット数を変える必要があるが、ガウス過程にはない。

■ NNの事前分布を広いクラス（？）に設定すると、予測分布がガウス分布に近くなる。