【ことわり】
なぜか表内の注釈が表示されません。適宜口頭で補うので勘弁ください。
【Abbreviation の約束】
・NN；Neural Network

1　ベイズニューラルネットワーク

ネットワークの重みを求めるにあたってのフレームワークについて、これまでの振り返り（単純な線形回帰モデル）と今回やること（多層 NN モデル）を以下の表 1 で整理する[*1]。

表 1　モデルの比較

変更点	単純なモデル	多層 NN モデル
手法	最尤推定	ベイズ的取り扱い
事後分布	厳密な評価可能[*2]	厳密な評価不可
予測分布	閉形式で表示可能[*3]	閉形式で表示不可

この表から多層 NN モデルについて、事後分布や予測分布を求める際にネットワーク関数のパラメタについて近似が必要となることがわかる。そのため、まずパラメタの事後分布に対してラプラス近似を適用することを起点として議論を進めていくことにする。

そしてラプラス近似に加えて適宜近似を導入すると回帰モデル及びクラス分類モデルが得られこれが本日のアウトプットである。今回近似のポイントは 2 つ。

1. ラプラス近似で作ったガウス分布の分散は十分小さい。
2. ネットワーク関数がパラメタに関して線形[*4]

この二つのポイントについては後で詳細な式変形を辿る際に触れたいと思う。

今回の流れは次の通りである。まず回帰モデルについて予測分布の表式をベイズ的アプローチによって得る。その後クラス分類モデルについて前者のモデルをベースに必要な修正を施すことで、その表式を得る。なお、議論の都合上ひとまず超パラメタについては固定しておく。必要に応じて超パラメタの最適化も行っていく。

1.1　パラメタの事後分布

ここでは 1 次元の連続な目標変数 t を入力ベクトル $\boldsymbol{x}$ から予測する問題を考える[*5]。つまり予測分布モデルの表式を得ることにする。

予測分布を得るためのフレームワークはこれまで習ってきたように以下の通りである。

1. パラメタについて事前分布と尤度関数を定義

[*1] ちなみに今回範囲はセクション 5.6 との繋がりはないと認識している。
[*4] のちにわかるがこれはちょっと語弊があるケースが生じるので注意。
[*5] このサブセクションで予測分布まで取り扱うと言っているのに、サブセクションの名前が「パラメタの事後分布」というのはナンセンスだと思う。とはいうものの本レジュメでは教科書に合わせることにしています。

2. ベイズの定理を用いて定義した事前分布と尤度関数から事後分布を計算
3. パラメタについて周辺化[*6]を行うことで予測分布を計算

本セクションの冒頭でも述べたが上記フレームワークにおいて必要な近似を適宜施していく。近似を使うタイミングは次の通りである。まずステップ 2. でラプラス近似を用いる。それからステップ 3. において、セクション冒頭に追加で導入すると述べた 2 つの近似（ガウス分布の分散が十分小さい/ネットワーク関数がパラメタに関して線形）を用いる。ここから上記フレームワークにしたがいベイズニューラルネットワークの予測分布の表式（5.172）まで導いてみせる。

1. 事前分布と尤度関数の定義

読み合わせとするが必要な式のみ導入しておく。

・事前分布

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

ここで $\boldsymbol{w}$ は重みパラメタであり α は超パラメタである。

・尤度

$$p(D|\boldsymbol{w}, \beta) = \Pi_{n=1}^{N} N(t_n|y(\boldsymbol{x}_n, \boldsymbol{w}), \beta^{-1}) \tag{2}$$

ここで $\boldsymbol{x}_i$ はデータ集合であり i.i.d. とする。$D = \{t_1, \cdots, t_N\}$ は目標値の集合である。

2. 事後分布の計算（ベイズの定理の利用）

ベイズの定理を用いて事後分布を求める。式 1 と 2 によって事後分布の表式は

$$p(\boldsymbol{w}|D, \alpha, \beta) \propto p(\boldsymbol{w}|\alpha)p(D|\boldsymbol{w}, \beta) \tag{3}$$

となる。尤度の部分に $y(\boldsymbol{x}_n, \boldsymbol{w})$ が入り込んでおり、この部分が非線形となるため[*7]事後分布がガウス分布にならない。ここでラプラス近似を施すことによってガウス分布的な取り扱いができるようにしたい[*8]。

ラプラス近似のフレームワークは基本的には下記の通り。

・元の事後分布のモードを見つける。

・事後分布の 2 階微分の行列を評価する。

・p.215 式 (4.134) を用いて近似分布を計算する。

上記フレームワークのもとで事後分布の近似形を得るところまでたどり着いてみせる。

まず事後分布のモードについてはその分布の対数を最大化することを考えてやれば良い[*9]。対数は以下のように表せる[*10]。

$$\ln p(\boldsymbol{w}|D) = -\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} - \frac{\beta}{2}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2 + const. \tag{4}$$

この最大化は適当な最適化アルゴリズムを用いて実行される（らしい）。この最適化計算によって我々は事後分布のモード $\boldsymbol{w}_{MAP}$ を得る。

[*6] 先に断っておくと周辺化積分を解析的に実行することはできないので工夫が必要になる。その工夫は本文でちゃんと説明します。

[*7] y は p.226 式（5.1）のような形なので当然非線形です。

[*8] ラプラス近似の詳細については p.213〜215 にて振り返りを行う。

[*9] 教科書に「これは正則化二乗和誤差の最大化に相当する」と書いてあるが、ここは「最小化に相当する」の間違いだろうか？私の教科書だけが間違えている可能性はあるが...

[*10] 式 3 に対数をかませるだけ。詳細は割愛。

次に事後分布の 2 階微分を求める[*11]。式 4 に重みについての微分演算子を 2 回かます。この微分計算によって得られる行列を $\boldsymbol{A}$ としておく。y の重みに関する微分について連鎖公式を使うなどすれば

$$\begin{aligned}\boldsymbol{A} &= -\nabla\nabla \ln p(\boldsymbol{w}|D, \alpha, \beta) \\ &= \alpha\boldsymbol{I} + \beta\sum_{n=1}^{N} \nabla y_n (\nabla y_n)^T + \sum_{n=1}^{N} (y_n - t_n)\nabla\nabla y_n \end{aligned} \tag{5}$$

となる。ここで第二項について

$$\boldsymbol{H} = \sum_{n=1}^{N} \nabla y_n (\nabla y_n)^T + \sum_{n=1}^{N} (y_n - t_n)\nabla\nabla y_n \tag{6}$$

と約束する。これは p.252 式（5.83）の結果そのものである。

最後に p.215 式 (4.134) を用いると事後分布は次のようなガウス分布で近似できる。

$$q(\boldsymbol{w}|D) = N(\boldsymbol{w}|\boldsymbol{w}_{MAP}, \boldsymbol{A}^{-1}) \tag{7}$$

平均は近似していない元の分布のモードを採用し分散共分散行列はヘシアンを含んだ行列を採用する。以上がラプラス近似を用いた事後分布の近似形の導出である。そして次のステップでは我々が知りたかった予測分布を求めることにする。

3. 予測分布の計算

ステップ 2. で求めた事後分布について重み $\boldsymbol{w}$ に関する周辺化積分を施す。これが求めたい予測分布（1 次元の連続な目標変数 t を入力ベクトル $\boldsymbol{x}$ から予測する分布）であった。予測分布は重みに関する条件付き確率 $p(t|\boldsymbol{x}, \boldsymbol{w}, \beta)$ を用いて[*12]

$$p(t|\boldsymbol{x}, D) = \int p(t|\boldsymbol{x}, \boldsymbol{w}) q(\boldsymbol{w}|D) d\boldsymbol{w} \tag{8}$$

のように表せる。

残念なことに、この積分は被積分関数の $p(t|\boldsymbol{x}, \boldsymbol{w})$ の中に非線形関数 y が入り込んでおり、解析的に扱いづらいため別のアプローチで予測分布の表式を得ることを考える。そのために追加で近似を用いることにする。この近似についてはセクションの冒頭でも述べたが改めてポイントを以下に記しておく。

・事後分布の分散が十分に小さい。つまり $\boldsymbol{w} - \boldsymbol{w}_{MAP}$ が微小。
・分散が小さいことからネットワーク関数について w_{MAP} まわりでテイラー展開ができる。

$\boldsymbol{w} - \boldsymbol{w}_{MAP}$ が微小であるため、ネットワーク関数 $y(\boldsymbol{x}, \boldsymbol{w})$ についてテーラー展開を行うと線形項のみ残すことができる。

$$y(\boldsymbol{x}, \boldsymbol{w}) \simeq y(\boldsymbol{x}, \boldsymbol{w}_{MAP}) + \boldsymbol{g}^T(\boldsymbol{w} - \boldsymbol{w}_{MAP}) \tag{9}$$

ここで $\boldsymbol{g}$ は

$$\boldsymbol{g} = \nabla_{\boldsymbol{w}} y(\boldsymbol{x}, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{MAP}} \tag{10}$$

と定義している。これを条件付き確率の表式に代入してやると

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) \simeq N(t|y(\boldsymbol{x}, \boldsymbol{w}_{MAP}) + \boldsymbol{g}^T(\boldsymbol{w} - \boldsymbol{w}_{MAP}), \beta^{-1}) \tag{11}$$

[*11] 正確には事後分布の負の対数ですが。
[*12] p.281 式（5.161）を参照。

というガウス分布の表式が得られる。

ここから予測分布を求めるために使える便利な公式（2.115）[*13]を思い出そう。それは条件付き分布と事後分布が与えられているときに周辺分布、つまり今回でいう予測分布を一般的に与えるものであった。

実際に今から公式（2.115）を使うが、実用的な観点から使い方について注意を記載しておく。注意するポイントは「(2.115）内の変数やパラメタと自分が今計算している式の中身の変数やパラメタの対応関係を考える」ということであった。その対応関係について表 2 で整理しておこう。この対応関係を考えることで予測分

表 2　変数・パラメタの対応関係

変更点	p.90 式（2.115）	今回
確率変数	$\boldsymbol{y}$	t
期待値の係数行列	$\boldsymbol{A}$	$\boldsymbol{g}^T$
期待値	$\boldsymbol{\mu}$	$\boldsymbol{w}_{MAP}$
期待値の切片部分	$\boldsymbol{b}$	$y(\boldsymbol{x}, \boldsymbol{w}_{MAP})$
条件付き分布の精度行列	$\boldsymbol{L}$	β
精度行列	$\boldsymbol{\Lambda}$	$\boldsymbol{A}(= \alpha \boldsymbol{I} + \beta \boldsymbol{H})$

布の式は

$$p(t|\boldsymbol{x}, D, \alpha, \beta) = N(t|y(\boldsymbol{x}, \boldsymbol{w}_{MAP}), \sigma^2(\boldsymbol{x})) \tag{12}$$

となる。ここで分散については

$$\sigma^2(\boldsymbol{x}) = \beta^{-1} + \boldsymbol{g}^T \boldsymbol{A}^{-1} \boldsymbol{g} \tag{13}$$

ように定義する。

この分布の特徴量である期待値と分散について以下のポイントを整理しておく。

・平均；ネットワーク関数 $y(\boldsymbol{x}, \boldsymbol{w}_{MAP})$ で与えられる。重みは MAP 推定値となる。
・分散；線形回帰モデルの結果 p.155 式（3.59）と対応する形で与えられる[*14]。

1.2　超パラメタ最適化

先のサブセクションでは超パラメタについては固定するという前提で議論を行った。本サブセクションではエビデンス理論[*15]を導入し、超パラメタについてセルフコンシステント[*16]な方程式を導くことで、超パラメタの推定を行うところまで議論する。

ここでエビデンス理論を用いた超パラメタの決定に関するフレームワークを簡単に整理しておく。

[*13] p.90 下部参照。

[*14] 教科書に書かれている「対比的」という日本語はミスリーディングだと思う。「あれ？なんか日本語おかしいぞ？」と思ったら原文を確認するということを教訓としたい。

[*15] p.164〜168 を参照。

[*16] 「セルフコンシステント」とは「問題を解くために, 問題の答えが先に必要になってしまう」という状況を指す表現です。今で言えば α を求めるために α が必要になってしまうという状況に対応しています。このような方程式の問題を解くための基本的なアプローチとしては、「それっぽい」解を持ってきて次の解を計算する。今度はその計算した解を再び方程式に代入し新しい解を求める。この操作を繰り返し（iterative）実行し入力と出力の差が十分に小さくなったところで、それを方程式の解にしてやるという流れを辿ります。計算終了にあたって、例えば入力と出力の差のオーダーが小数点第何位までといった閾値を設けるといった対応をします。

1. モデルエビデンスを定義（5.174）。
2. エビデンスの対数をとる。
3. 超パラメタごとに対数エビデンスを偏微分しセルフコンシステント方程式を導出。
4. セルフコンシステント方程式を iterative に解くことで超パラメタを決定（今回は詳細説明なし[*17]）

今回はステップ 1.〜3. について式変形を追う中で理解してもらう。一度学んでいる内容だが、だいぶ前の内容なので途中計算はなるべく丁寧に追うことにする。

1．モデルエビデンスを定義

超パラメタのモデルエビデンスは次のように同時分布 $p(D|\boldsymbol{w},\beta)$ を重み $\boldsymbol{w}$ で周辺化積分することによって得られる。

$$p(D|\alpha,\beta)=\int p(D|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)d\boldsymbol{w} \tag{14}$$

2．対数エビデンスの表式を確認

まず上で定義したエビデンスの近似形を p.216 式（4.135）を用いて表す。（4.135）内における Z がエビデンスと等しくなることに注意すると[*18]

$$\begin{aligned}
Z &= p(D|\alpha,\beta) \\
&\simeq f(\boldsymbol{w}_{map})\frac{(2\pi)^{\frac{W}{2}}}{|\boldsymbol{A}|^{1/2}} \ (\because \text{ ラプラス近似}) \\
&= p(D|\boldsymbol{w}_{MAP})p(\boldsymbol{w}_{MAP})\frac{(2\pi)^{\frac{N}{2}}}{|\boldsymbol{A}|^{1/2}} \ (\because \text{ p.216 中段の尤度や事前確率のルール})
\end{aligned} \tag{15}$$

[*17] 方程式を解くのはコンピューターの仕事なので。
[*18] 等しくなるようにルールづけをしている。

式 15 の対数を取ると

$$\begin{aligned}
\ln[p(D|\boldsymbol{w}_{MAP})p(\boldsymbol{w}_{MAP})\frac{(2\pi)^{\frac{W}{2}}}{|\boldsymbol{A}|^{1/2}}] &= \ln p(D|\boldsymbol{w}_{MAP}) + \ln p(\boldsymbol{w}_{MAP}) + \frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{A}| \\
&= \ln[\Pi_{n=1}^{N} N(t_n|y(\boldsymbol{x}_n, \boldsymbol{w}_{MAP}), \beta^{-1})] + \ln[N(\boldsymbol{w}_{MAP}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})] \\
&+ \frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{A}| \\
&= \sum_{n=1}^{N} \ln[N(t_n|y(\boldsymbol{x}_n, \boldsymbol{w}_{MAP}), \beta^{-1})] + \ln[N(\boldsymbol{w}_{MAP}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})] \\
&+ \frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{A}| \\
&= \sum_{n=1}^{N} [\ln(\frac{\beta^{1/2}}{\sqrt{2\pi}}) + \ln\{\exp(-\beta\frac{(t_n - y_{MAP})^2}{2})\}] \\
&+ \ln\{\frac{\alpha^{W/2}}{\sqrt{2\pi}^{W}}\exp(-\alpha\frac{\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP}}{2})\} + \frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{A}| \\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \sum_{n=1}^{N}\frac{\beta}{2}(t_n - y_{MAP})^2 \\
&+ \frac{W}{2}\ln\alpha - \frac{W}{2}\ln 2\pi - \frac{\alpha}{2}\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP} + \frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{A}| \\
&= -\sum_{n=1}^{N}\frac{\beta}{2}(t_n - y_{MAP})^2 - \frac{\alpha}{2}\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP} \\
&- \frac{1}{2}\ln|\boldsymbol{A}| + \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi \\
&= -E_{MAP} - \frac{1}{2}\ln|\boldsymbol{A}| + \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi \qquad (16)
\end{aligned}$$

式 16 が求めるべき対数エビデンスである。ここで正則化誤差関数 E_{MAP} は次のように定義している。

$$E_{MAP} = \sum_{n=1}^{N}\frac{\beta}{2}(t_n - y_{MAP})^2 + \frac{\alpha}{2}\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP} \qquad (17)$$

実は対数エビデンスも正則化誤差関数も p.166 式（3.86）の線形回帰モデルの結果に対応する形になっていることが確かめられる。

3. セルコンシステント方程式の導出

超パラメタ α, β のそれぞれで対数エビデンスの式 16 を偏微分しイコール 0 とおいてやれば目的のセルフコンシステント方程式が得られる。ただしステップ 3. についてはレジュメ【210206 輪講】の頁 3〜4 を大いに参考にしながら説明する。したがって本レジュメに詳細を記すことはしない。ただし変数の対応関係だけ以下に記しておく。

・$\boldsymbol{\Phi}^t\boldsymbol{\Phi} \rightarrow \boldsymbol{H}$

・$\boldsymbol{m}_N \rightarrow \boldsymbol{w}_{MAP}$

計画行列の積をヘシアンに、重みの事後分布の期待値をモードに変更すれば良い。

ここで有効パラメタ内の固有値の取り扱いについて、線形回帰モデルの場合とニューラルネットワークの場合での違いを述べておく。

前者では、固有値は計画行列をベースに求めていた。計画行列は学習データで作られており一度作ってしまえば不変となりセルフコンシステント計算の中で固有値が変わるということはなかった。それゆえに p.168 式（3.91）内の λ は厳密に不変である。

一方後者では、固有値はヘシアンをベースに求めていた。ヘシアンは $\boldsymbol{w}_{MAP}$ で作られており、学習の中で更新され変化するためセルフコンシステント計算の中で固有値も変わってしまう。したがって p.284 式（5.179）内の λ を不変とみなして有効パラメタを計算することは厳密さには欠けていることに注意されたい。

1.3 クラス分類のためのベイズニューラルネットワーク

先のサブセクションでニューラルネットワーク関数に対して回帰モデルを作った。このモデルに適当な修正を加えることで 2 クラス分類のモデルを構築する。以下の表 3 で修正点を簡単に整理しておく。なお、出力は 1 つでロジスティックシグモイド関数で与えられるものとする。

表 3　モデルの相違点まとめ

変更点	線形回帰	クラス分類
1. 誤差関数	二乗和	交差エントロピー[*19]
2. 正則化誤差関数	式（5.165）	式（5.182）
3. 事後分布の近似	本レジュメ p.3 式 7	左に同じ[*20]
4. 出力の近似	出力関数 y を線形近似	出力ユニットの活性 a を線形近似

変更点 4. について、今回予測分布の近似表式を求めるにあたっては活性化関数を経由することになる。実はこの手法は p.219 サブセクション 4.5.2 で既に学んでいる[*21]。これらの修正点を踏まえたうえで次の二つの結果を導く。

・クラス分類のためのエビデンス関数（修正点 1. と 2. が関係）
・クラス分類のための予測分布（修正点 3. と 4. が関係）

エビデンス関数についてはまず誤差関数を新しく定義し、ラプラス近似によって重みの周辺分布を計算する。その後超パラメタについてのセルフコンシステント方程式を導く。予測分布については目標変数の事後分布をデルタ関数で近似する方法と、活性化関数に関する適当な近似をいくつかを施す方法を紹介する。そして活性化関数に対する近似から予測分布の式（5.190）を導く。

・エビデンス関数

式 16 の 1 行目を利用する。今回は第一項を交差エントロピー関数に置き換えてあげれば良いのでモデルエビ

[*21] プロビット関数の逆関数の利用というやつでした。

デンスの式は次のようになる。

$$\begin{aligned}\ln p(D|\alpha) &\simeq \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \ln\{\frac{\alpha^{W/2}}{\sqrt{2\pi}^W}\exp(-\alpha\frac{\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP}}{2})\} \\ &+ \frac{W}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{A}| \\ &= \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} - \frac{\alpha}{2}\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP} \qquad (18) \\ &+ \frac{W}{2}\ln\alpha - \frac{W}{2}\ln(2\pi) + \frac{W}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{A}| \\ &= -E_{MAP} + \frac{W}{2}\ln\alpha - \frac{1}{2}\ln|\boldsymbol{A}| \qquad (19)\end{aligned}$$

ここで正則化誤差関数 E_{MAP} は次のように定義する。

$$E_{MAP} = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \frac{\alpha}{2}\boldsymbol{w}_{MAP}^T\boldsymbol{w}_{MAP} \qquad (20)$$

ここから式 18 を α について偏微分しイコール 0 とすることでセルフコンシステントな方程式が導かれる。

教科書の図 5.22 には 2 クラス分類問題に対してそれぞれ最尤推定とエビデンス理論に基づいてフィッティングした曲線を描いている。前者（黒線）に比べて後者（赤線）は過学習を抑えていることが読み取れる。

・予測分布

予測分布は式（5.168）で定義されている。

$$p(t|\boldsymbol{x}, D) = \int p(t|\boldsymbol{x}, \boldsymbol{w})q(\boldsymbol{w}|D)d\boldsymbol{w} \qquad (21)$$

これは解析的に取り扱うことができない。そのため右辺内の事後分布 $q(\boldsymbol{w}|D)$ が $\boldsymbol{w} = \boldsymbol{w}_{MAP}$ 周りで鋭く尖っているデルタ関数であると近似する。この近似によって予測分布の式は

$$p(t|\boldsymbol{x}, D) \simeq p(t|\boldsymbol{x}, \boldsymbol{w}_{MAP}) \qquad (22)$$

と表すことができるが、これはかなりナイーブな近似方法なので別のアプローチを考えたい。

ここで取る手法は出力ユニットの活性化関数を線形近似するというものである。$\boldsymbol{w} = \boldsymbol{w}_{MAP}$ 周りでの展開を行うことで

$$a(\boldsymbol{x}, \boldsymbol{w}) \simeq a_{MAP}(\boldsymbol{x}) + \boldsymbol{b}^T(\boldsymbol{w} - \boldsymbol{w}_{MAP}) \qquad (23)$$

が導かれる。

ここから a の分布について考えたのちロジスティックシグモイド関数とのたたみ込み積分によって予測分布を求めていく。この議論は p.218 の 4.5.2 の議論を大いに参考にしながら説明する。したがって本レジュメでは詳細は割愛。最後にロジスティックシグモイド関数をプロビット関数の逆関数で近似してやれば式（5.190）が導かれる。

$$p(t = 1|\boldsymbol{x}, D) = \sigma(\kappa(\sigma_a^2)a_{MAP}) \qquad (24)$$

以上からベイズニューラルネットワークにおけるクラス分類の予測分布の表式が得られた。ラプラス近似を用いていることから $\boldsymbol{w}_{MAP}$ が（間接的に）数式内に現れることがポイントとなる[*22]。

[*22] 露わに出てきているわけではありませんが、活性化関数 a_{MAP} は $\boldsymbol{w}_{MAP}$ によって定まっています。

最後に今回学習内容を一言でまとめるとベイズの枠組みで予測を行う際にはパラメタを適切に周辺化することがポイントになるということである（**Chap.5 及び上巻終わり**）[*23]。

[*23] まずは半年間お疲れ様でした。スケジュールを見たら 11 月 1 日が初回になっていました。これだけ難しい教科書を投げ出すことなくひとまずやり抜くことができたのはよしきの多大なるサポートあってのことだと思っています。ありがとうございます。老婆心ながら次回に向けて思い出しておいた方がいい項目を洗い出しておきます。次回範囲はカーネル法の導入になります。なのでカーネル関数ってなんだっけということは復習しておいた方がいいでしょう。上巻の p.157 あたりを見るのが良いと思います。他には私が見た限りで関連する項目についてキーワードを並べていくと、「正則化二乗和誤差」、「フィッシャー情報量/行列」でしょうか。「フィッシャー情報量/行列」については「自然科学の統計学」p.120 の 4.3 節を見るのが良いでしょう。下巻はさらにハードな内容かと思いますがくじけずに一緒に頑張りましょう。私たちなら乗り越えられると信じています。なんとか来年の 3 月までには終わらせたいですね。