5.5.1 無矛盾なガウス事前分布（続）

前回は、ネットワークの学習における「無矛盾」の定義と、どのような正則化項であれば無矛盾となるかを学んだ。

～おさらい～

無矛盾とは、入力変数や目標変数を線形変換した場合に、変換前と等価な出力をするネットワークが得られるような性質である。次式のように、ネットワークの層ごとに正則化係数を定めるとき、無矛盾性を持つことができる。

$$\frac{\lambda_1}{2}\sum_{w\in W_1} w^2 + \frac{\lambda_2}{2}\sum_{w\in W_2} w^2 \qquad \left(\begin{array}{l} W_1,\ W_2 \text{は各層における} \\ \text{バイアス項を除いたパラメータ} \end{array}\right) \qquad (5.121)$$

～おさらい終了～

教科書5.5.1の残りの部分で主張している内容は、主に下記2点である。

1. (5.121) 式のようなバイアス項を除いた正則化はベイズ学習に適さないので、通常はバイアス項も含めた正則化項を使用する。
2. 正則化係数の決定が、ネットワークにどのような影響を及ぼすのかを図5.11より視覚的に理解する。

これら2点について順に説明する。

1. (5.121) 式のようなバイアス項を除いた正則化はベイズ学習に適さない

もともと、正則化が出てきた背景として、p.30 (1.66) でMAP推定を行なった場合に、事前分布の対数項から2次正則化が出てきたという背景があった。今回のケースでも事前分布を次のように設定することで、(5.121) 式のような2次の正則化項が導き出される。

$$p(w|\alpha_1,\alpha_2) \propto \exp\left(-\frac{\alpha_1}{2}\sum_{w\in W_1} w^2 - \frac{\alpha_2}{2}\sum_{w\in W_2} w^2\right) \qquad (5.122)$$

事後分布：$p(w|x,t) = p(t|x,w,\alpha_1,\alpha_2)\,p(w|\alpha_1,\alpha_2)$

の対数の最大化を考えると、赤、下線部分から、(5.121)式の正則化項が出てくる。

この事前分布は、バイアス項に関する制約を置いていないので、バイアス項のパラメータについては、無情報事前分布の仮定の元で、$-\infty$ から $+\infty$ まで等しい確率で出現する。このような事前分布は正しく正規化できないため、変則事前分布である。（変則事前分布についての説明はp115）

このとき、モデルエビデンスは次のようになる。

$$p(D|M_i) = \int p(D|w, M_i)\, p(w|M_i)\, dw \quad (\because \text{P.161 (3.68)})$$

$$= 0$$

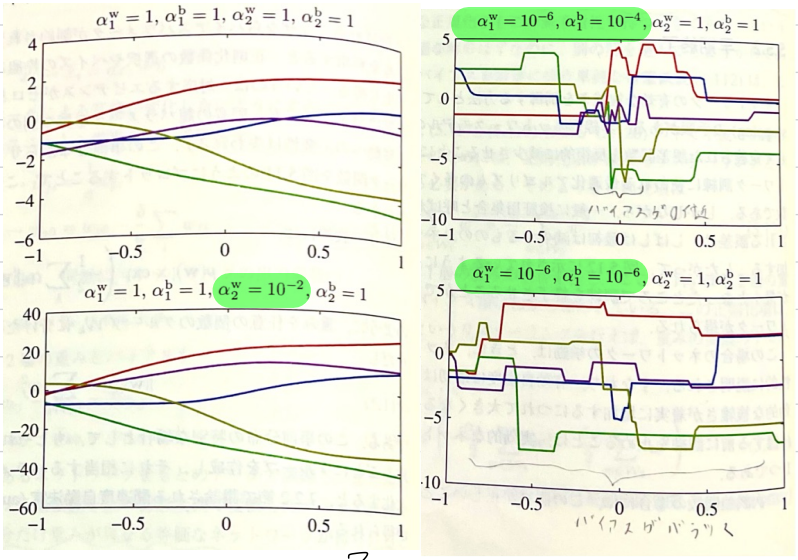
（となるらしい。$p(w|M_i)$ が変則事前分布であることを使って示したいが、良い方針が思いつかなかった）

モデルエビデンスが0となると、モデル比較できないため、これはベイズ学習に適さない。

したがって、**通常はバイアス項も含めた正則化項を使用する**。

2. 正則化係数の設定が、ネットワークにどのような影響を及ぼすか

正則化係数 λ_1, , λ_2 は、学習時にこちらで決定するものである。そこで、どのような正則化係数を設定するべきかについて興味が沸く。λ_1, , λ_2 は、（5.122）でみるように、ガウス事前分布の精度パラメータに相当する。精度パラメータをいくつか変更したネットワークについて出力をプロットしたものが図5.11である。



y軸のスケールが変わる

$\alpha_1^w = 10^{-6}$ とすると、線の変化が急峻になる。

α_1^b を変えると変化の位置が変わる。

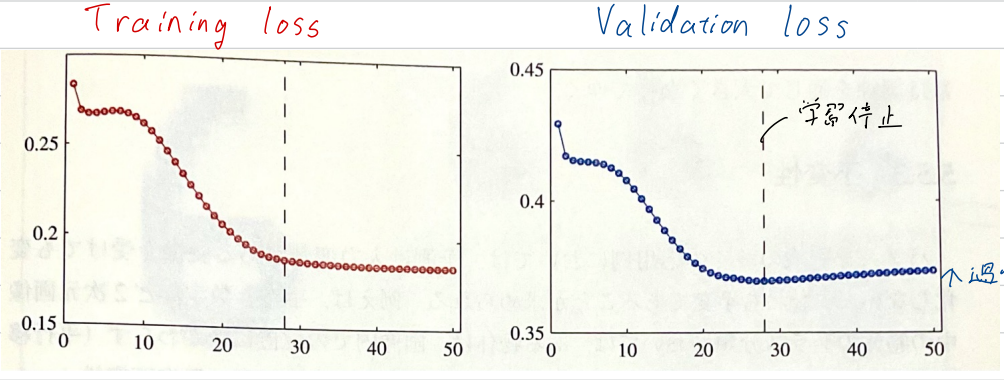
（感想）この図で言いたいことは、おそらく正則化パラメータを適当に選んではいけないよ、ということだと思っている。正則化パラメータがガウス事前分布の精度パラメータの役割を果たしている以上、ある程度自然なネットワークが生成されるに正則化パラメータを選択する必要がありそう。

7.2.2 節で議論される関連度自動決定では、正則化パラメータ α_k $(= \lambda_k)$ を尤度最大化によって決定するらしい。詳細は7章に回す。

正則化の他にも、過学習を抑制するような手法が存在する。その一つが次に扱う早期終了である。

5.5.2 早期終了

過学習を抑制するために、検証データに対する誤差を評価する手法については1章で学んだ。ニューラルネットワークの反復的な学習においても、検証データの誤差が減少しなくなった時点で学習を停止することで、過学習を抑制できる。（下図）この方法を早期終了と呼ぶ。



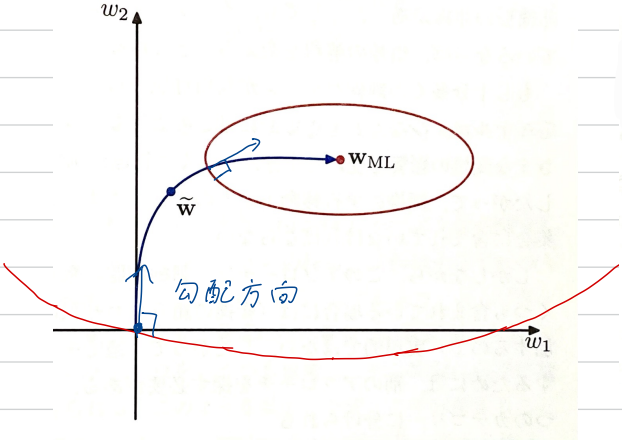
また、早期終了は荷重減衰を用いる場合と類似の挙動をする。これを確かめる。

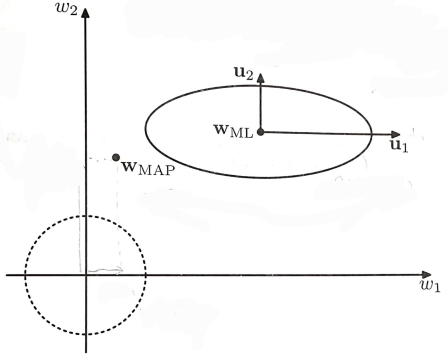
> （感想）ここの教科書の例は非常に限定的な例に思える。というのも、誤差関数が重み w の二次関数としている点、パラメータの更新方向に勾配のみを考慮しており、2次微分を考慮してない点などが限定的である。

左下図は2次誤差関数（赤線の等高線）に対して、勾配を用いたパラメータ更新を反復して行う場合のパラメータの軌跡（青矢印線）を示す。2次誤差関数の形状から、初めは w_2 方向への勾配が大きく、後半は w_1 方向への勾配が大きくなる。
早期終了した結果、求まるパラメータ $\tilde{w}$ は投稿線の楕円の長軸側に位置する。

一方で、荷重減衰の正則化を用いた学習でも、求まるパラメータは楕円の長軸側に位置する（右下図）。荷重減衰で求まるパラメータは、事前分布にガウス分布を用いた場合のMAP推定量に相当するため、w_{MAP} で示している。

P169 図 3.15





5.5.3 不変性

前項、前々項と、過学習を抑制する方法を紹介してきた。ここからはガラッと話題が変わって、不変性の話となる。

パターン認識において、入力変数がある変換を受けても変化しない、すなわち不変であることが求められることがある。例えば画像データにおいて、対象の位置の変化に対する不変性（平行移動不変性）や、サイズに対する不変性（尺度不変性）などがその例である。
学習データに十分なバリエーションのデータがあれば、不変性をもつモデルを学習できるが、そのようなケースは稀であるため、通常はいくつか工夫を行う。5.5.3 項から5.5.6 項までは不変性を保つための工夫を紹介する。以降、下記4つのアプローチを紹介する。

① データ増強
学習データを人工的に変換したものを学習データに加える。画像データの例だと、画像を拡大、縮小したものや、傾けたり反転させたりしたデータを新たに学習データに追加する。データ増強（Data augmentation）などと呼ばれる。（5.5.5 項）

② 接線伝搬法
正則化項に誤差関数を加えることで、入力の変換に対して出力が変化した場合にペナルティーを与える。（5.5.4 項）

③ 不変性を保つような特徴抽出
特徴抽出の段階で不変性を担保する。例えば、画像中で対象物が写っていそうな部分を切り抜くような前処理を行うことで、画像中のどの位置に物体が写っていても特徴が不変となり、予測の値も不変となる。職人芸のような処理が必要になるため、教科書ではこれ以上扱わない。

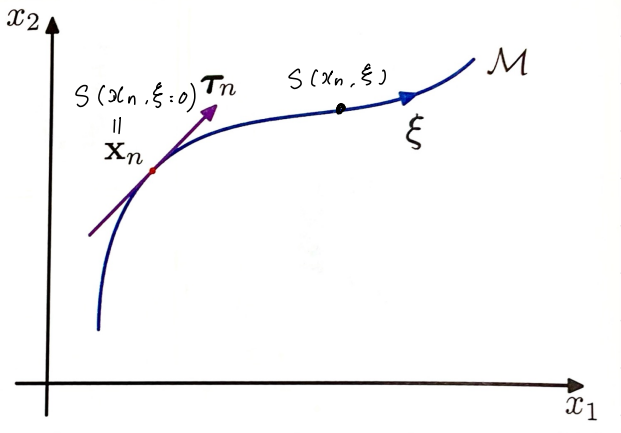
④ たたみ込みニューラルネットワーク
ニューラルネットワークの構造に不変性を構築する。局所的受容野を利用して、重みパラメータを入力特徴量間で重みを共有する。（5.5.6 項）

5.5.4 接線伝搬法

接線伝搬法は、誤差関数の正則化項を調整することで、不変性のあるモデルを学習する方法である。変換があるパラメータで支配されるとき、パラメータを変化に対して出力の変化が大きいほどペナルティを課すような正則化項を用いる。

ここでは、連続な変換（移動や回転を含むが、鏡像は含まない）についてのみ扱う。教科書に合わせて、2次元のデータを例とする。ある変換が単一のパラメータ ξ で支配されるとすると、変換先の集合はある1次元の集合となる（下図青線）。（この青線上の空間を多様体と呼んでいるが、多様体がなんなのか詳しくは理解していない。） x_n にこの変換を作用させて得られるベクトルを $s(x_n, \xi)$ と表現すると、青線上の点は全て $s(x_n, \xi)$ で表現される。$s(x_n, \xi)$ の $\xi = 0$ における接ベクトルは次のように表現できる。（接ベクトルはこの後の正則化項に出てくるのでここで紹介している）

$$\tau_n = \left. \frac{\partial s(x_n, \xi)}{\partial \xi} \right|_{\xi = 0} \qquad (5.125)$$



接線伝搬法では、 ξ を変化させたときの y の変化量に対してペナルティを課す。正則化項を Ω で表現すると、誤差全体は次のように表現される。

$$\tilde{E} = E + \lambda \Omega$$

接線伝搬法では、正則化項に重みベクトルのノルムではなく次のものを使用する。

$$\Omega = \frac{1}{2} \sum_n \sum_k \left(\frac{\partial y_{nk}}{\partial \xi} \Big|_{\xi=0} \right)^2$$

ξ を変化させたときの y の変化量

$$= \frac{1}{2} \sum_n \sum_k \left(\sum_{i=1}^{D} \frac{\partial y_{nk}}{\partial x_i} \frac{\partial x_i}{\partial \xi} \Big|_{\xi=0} \right)^2 \quad (\because \text{Chain rule})$$

$$= \frac{1}{2} \sum_n \sum_k \left(\sum_{i=1}^{D} J_{ki} \tau_n \right)^2 \quad (5.128)$$

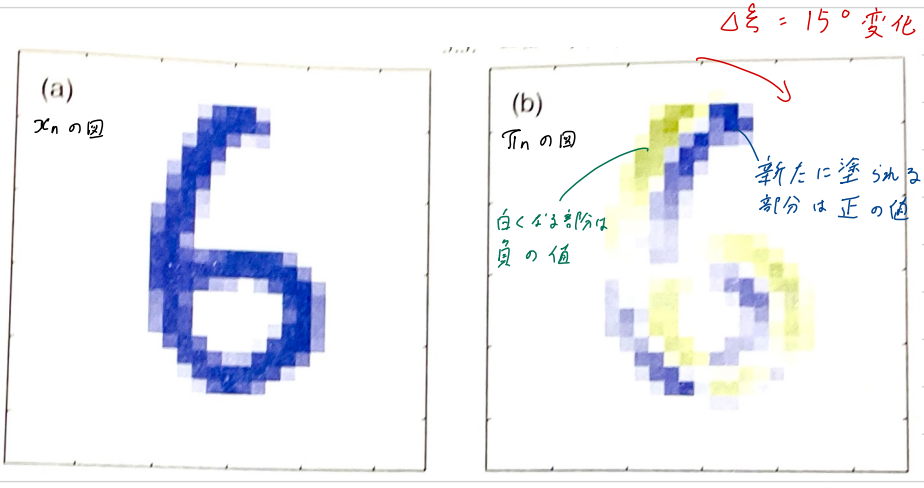
J_{ki}：ヤコビ行列　　τ_n：$(\because 5.125)$

また、誤差関数 $\tilde{E}$ の最小化には、Ω の勾配 $\frac{\partial \Omega}{\partial w_{kl}}$ が必要となる。これは、これまでの逆伝搬を応用することで計算可能である（らしい）。演習5.26の内容だが今回は割愛する。

τ_n の計算だが、実装の際は有限幅の差分によって近似できる。

$$\tau_n = \frac{S(x_n, \Delta\xi) - x_n}{\Delta\xi} + O(\Delta\xi)$$

手描き文字画像を例に、回転角 ξ について τ_n を求めたものが下図 b である。微小に回転させた際に、ピクセルが新たに塗られる部分については値が大きく、逆にピクセルが白塗りに戻る部分は値が小さくなることが図示されている。



次回以降の 5.5.5 , 5.5.6 は、不変性を担保するためのアプローチ1とアプローチ4の手法の内容となる。少しきりが悪くなってしまったが何卒よろしくお願いします m(_ _)m