本レジュメの最後に App. をつけています。

1　ニューラルネットワークの正則化

本節の役割はニューラルネットワークの学習において、過学習を防ぎ汎化性能を向上させる手法について学ぶというものであった。4 章まではニューラルネットワークを用いない学習について、汎化性能を保証する手法を取り扱ってきた。しかしながら、今後はニューラルネットワークによる学習をメインの武器に据えていくことから、ここでその手法を学ぶことには意義がある[*1]。

モデルが汎化性能を保つということは、入力が何らかの変換を受けた時についてもそうなっていて欲しい。入力の変換後も汎化性能が保たれ、予測結果が不変であることを「不変性を持つ」と呼ぶ。このような不変性を保証するためのアプローチは以下の 4 つであると学んだ[*2]。

1. データ増強(←今回取り扱い対象)
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。データ増強 (Data augmentation) などと呼ばれる (5.5.5 項)。

2. 接線伝搬法(←前回学習済)
正則化項に誤差関数を加え、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(←取り扱いなし)
特徴抽出の段階で不変性を担保する。

4. たたみ込みニューラルネットワーク(←今回取り扱い対象)
ニューラルネットワークの構造に不変性を構築する (5.5.6 項)。

前回はアプローチ 2. について取り扱った。そして上記の通り今回はアプローチ 1. と 4. について順番にその手法を理解していく。本レジュメの構成は次の通り。

・変換されたデータを用いた訓練（アプローチ 1.）では「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こと」↔「アプローチ 2. において正則化項を加えること」という対応関係について式変形を通じて理解してもらう。

・「たたみ込みニューラルネットワーク（アプローチ 4.）」では、優れた汎化性能を持つ手法であるたたみ込みニューラルネットワークについてその原理を学ぶ。

・「ソフト重み共有」では、アプローチ 4. を成り立たせるいくつかの機構のうち、その一部分（重みに対する条件）を置き換える手法について学ぶ。この置き換えは誤差関数に正則化項を導入することに相当している。このようにして置き換えた手法によって重みの分布（平均値・分散）がどのような学習過程を経て決められるかについて式変形を通じて理解する。

[*1] これは教科書の下巻にも繋がるから、ということも勿論あるのですが、機械学習のムーブメントとしてそうなっているという意味合いが強いです。とは言うものの現場の IT エンジニアの感触は、ちょっと違ったものかもしれないので、認識が違うと思う場合はコメントください。
[*2] 前回レジュメを参照。「頁 4；210410 輪講.pdf」

1.1 変換されたデータを用いた訓練

ここではアプローチ 1. と 2. の等価性について以下のようなフレームワークで示していく。

・各データ点に対してパラメタ ξ で定まる変換を施す[*3]。
・変換後の誤差関数を約束しテーラー展開を用いてパラメタ ξ について有意な項のみを残す[*4]。
・有意な項を残すと、今回約束した誤差関数は（5.127）とアナロジーを持った形で表現される（5.131-32)。
・上記で表現した誤差関数は ξ の次数に関する考察を行うことで簡単化した形で表せる。
・簡単化された正則化項として得られた量が（5.128）に等しいことから等価性の結論を得る。

等価性の結論からデータ補強というアプローチは誤差関数に正則化項を加えて学習させるという考え方に基づいており、理論的に正当性が保証された手法であると言える。以下では上記のフレームワークに従って、(5.129）を起点に（5.134）までたどり着いて見せる。

まず入力データを 5.5.4 で約束した方法で変換を施すと、変換前の誤差関数（5.129）は次のような形で書ける[*5]。

$$\tilde{E} = \frac{1}{2}\int\int\int (y(\boldsymbol{s}(\boldsymbol{x},\xi)) - t)^2 p(t|\boldsymbol{x})p(\boldsymbol{x})p(\xi)d\boldsymbol{x}dtd\xi \tag{1}$$

分布 $p(\xi)$ は平均ゼロで分散は小さいとする[*6]。

ここで変換後の入力 $\boldsymbol{s}(\boldsymbol{x},\xi)$ について ξ に関するテーラー展開を行う。

$$\begin{aligned}
\boldsymbol{s}(\boldsymbol{x},\xi) &= \boldsymbol{s}(\boldsymbol{x},0) + \xi\frac{\partial}{\partial\xi}\boldsymbol{s}(\boldsymbol{x},\xi)|_{\xi=0} + \frac{\xi^2}{2}\frac{\partial^2}{\partial\xi^2}\boldsymbol{s}(\boldsymbol{x},\xi)|_{\xi=0} + O(\xi^3) \\
&= \boldsymbol{x} + \xi\boldsymbol{\tau} + \frac{\xi^2}{2}\frac{\partial}{\partial\xi}\frac{\partial}{\partial\xi}\boldsymbol{s}(\boldsymbol{x},\xi)|_{\xi=0} + O(\xi^3)\ (\because (5.125)) \\
&= \boldsymbol{x} + \xi\boldsymbol{\tau} + \frac{\xi^2}{2}\boldsymbol{\tau}' + O(\xi^3)
\end{aligned} \tag{2}$$

ここから入力変換後の出力 $y(\boldsymbol{s}(\boldsymbol{x},\xi)$ は p.268 最下部の式のようになる[*7]。これを式 1 に代入する。ξ について次数に注目して整理すると p.269 上部 $\tilde{E}$ についての式が得られる（> Go to App.)。分布 $p(\xi)$ は平均ゼロと設定したから期待値 $E(\xi) = 0$ となる。また $E(\xi^2) = \lambda$ として $O(\xi^3)$ の項を無視すると誤差関数は次の表式でまとめられる。

$$\tilde{E} = E + \lambda\Omega \tag{3}$$

ここで Ω は式（5.132）に従うものとする。式 3 は（5.127）と共通の形式で表現されている。

次に共通な形式で表せているところからさらに踏み込み、等価性を示していく。実は誤差関数を最小にする関数 $y(\boldsymbol{x})$ は

$$y(\boldsymbol{x}) = E(t|\boldsymbol{x}) + O(\xi) \tag{4}$$

[*3] 変換のやり方は 5.5.4 で議論済み。
[*4] 結論を言えばパラメタ ξ について二次まで残すことになる。こんなものは後知恵だが仕方あるまい。
[*5] この形で書ける根拠が明確にわかっていない。こんな形になりそうくらいのイメージしか持てていない。
[*6] 元の入力ベクトル $\boldsymbol{x}$ を大きく変えないという前提で考えるため。とはいうものの「どのくらい大きく」なのか定量的にはわからない。
[*7] この導出はお手上げ。申し訳ない。

と書けるので正則化項の式（5.132）の第一項は 0 になる。こうすると正則化項の式は

$$\Omega = \frac{1}{2}\int (\boldsymbol{\tau}^T \nabla y(\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x} \tag{5}$$

のようになる。これは接線伝播法における正則化項（5.128）と等価なものになる[*8]。

1.2 たたみ込みニューラルネットワーク

ここではたたみ込みニューラルネットワークとは何かということと、その原理についての説明を行う。このたたみ込みニューラルネットワークという手法を用いることで入力に変換があっても不変性を持つモデルを構築することができる[*9]。ここは基本的に読み合わせとしたいが、教科書本文の段落ごとにポイントのみ本レジュメに記載しておく。

・第一段落；特になし。上記にて説明済み。
・第二段落；特になし。
・第三段落；たたみ込みってどんなことをするんですか（直観的な理解）？局所的な特徴抽出→画像全体の情報取得という考え方について
・第四段落；たたみ込みってどんなことをするんですか（フレームワーク紹介＋視覚的な理解）？画像輝度のたたみ込み計算との等価について[*10]。また、ここでは 5.5.7 との対比手法として「重み共有」という考え方を取り扱う。
・第五段落；ここからがよくわからなかった。[*11]

1.3 ソフト重み共有

本サブセクションで取り扱うソフト重み共有を用いたとき、学習によって得られる重みについて、先のサブセクションとどのようにアウトプットが異なるかまず結論を抑える。
・5.5.6.；同じ特徴マップに属するユニットは同一の重みを持つ。
・5.5.7.；同じ特徴マップに属するユニットは似たような重みを持つ（ソフト重み共有）。
このソフト重み共有においては異なる重みの値を持たせるため、重みの事前分布として混合ガウス分布を用いることがキーポイントとなる[*12]。このように混合ガウス分布を導入したときに、誤差関数（正則化項あり）を最小化することを考えるが、この誤差関数を支配するパラメタは次の 4 つ。

1. 重み w_i
2. ガウス分布の中心 μ_j
3. ガウス分布の分散 σ_j
4. 混合係数 π_j

以下では誤差関数を上記 4 種類のパラメタでそれぞれ微分を行い、各パラメタがどのような振る舞いを示すか

[*8] 厳密に言えば等価ではありません。データ点について（5.128）は離散形で、（5.134）は連続形で表現されているのだから。
[*9] ただし、本レジュメではその不変性を保証してやるわけではない。あくまで手法の仕組みの紹介にとどまる。
[*10] たたみ込みについてのお決まりの説明のことです。フィルター（ここで言う特徴マップのこと）を動かして元画像と掛け算するやつ。フィルターをいくつか用意しておいてそれらを入力画像上で動かしてたたみ込みを実行。
[*11] ここでの話がいわゆる「入力に変換が起こっても予測結果に影響ないよ、不変性保たれているよ」という話なのだろうが、この結論と教科書 p.271 下部（7 行〜15 行のところ）の説明とのつながりがわからなかった。
[*12] ベイズ的取り扱いは前提の上で議論を進めている。

調べることにする。

まず、準備として上記の通り重み $\boldsymbol{w}$ の事前分布を次のように混合ガウス分布で表す。ここで $\boldsymbol{w}$ を構成する各 w_i は i.i.d であるとする。

$$p(w_i) = \sum_{j=1}^{M} \pi_j N(w_i|\mu_j, \sigma_j^2) \tag{6}$$

π_j がいわゆる混合係数である。$\boldsymbol{w}$ 事前分布の負の対数尤度が誤差関数の正則化項になることは既に学んでいて、これが議論を進める上で便利なことも知っているからここで準備しておく。

$$\Omega(\boldsymbol{w}) = -\sum_i \ln(\sum_{j=1}^{M} \pi_j N(w_i|\mu_j, \sigma_j^2)) \tag{7}$$

誤差関数全体はこれまでと同様に（5.139）の形で表現され、この右辺第二項が式 7 であると考えれば良い。

ここからは、このサブセクションの冒頭で整理したように、誤差関数全体の最小値を求めるために各パラメタによる微分を考え、その結果に対する解釈を与える。

・重みw_i

ここでは誤差関数 $\tilde{E}$ を重み w_i で微分する。便利のため次の量 $\gamma(w_i)$ を定義しておく。

$$\gamma(w_i) = \frac{\pi_j N(w|\mu_j, \sigma_j^2)}{\sum_k \pi_k N(w|\mu_k, \sigma_k^2)} \tag{8}$$

式（5.139）の両辺を w_i で微分すると

$$\frac{\partial \tilde{E}(\boldsymbol{w})}{\partial w_i} = \frac{\partial E(\boldsymbol{w})}{\partial w_i} + \lambda \frac{\partial \Omega(\boldsymbol{w})}{\partial w_i} \tag{9}$$

ポイントは上式の第二項を評価することである。係数 λ はひとまず無視すると

$$\begin{aligned}
\frac{\partial \Omega(\boldsymbol{w})}{\partial w_i} &= -\frac{\partial}{\partial w_i} \ln(\sum_j^M \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}}) \exp(-\frac{(w_i - \mu_j)}{2\sigma_j^2}) \\
&= -\frac{\sum_j^M \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{(w_i-\mu_j)}{2\sigma_j^2}) \cdot \frac{-2(w_i-\mu_j)}{2\sigma_j^2}}{\sum_j^M \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{(w_i-\mu_j)}{2\sigma_j^2})} \\
&= \frac{\sum_j^M \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{(w_i-\mu_j)}{2\sigma_j^2}) \cdot \frac{(w_i-\mu_j)}{\sigma_j^2}}{\sum_k^M \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(w_i-\mu_k)}{2\sigma_k^2})} \quad (\because \text{分母の } j \text{ と } k \text{ は独立}) \\
&= \sum_j \gamma_j(w_i) \frac{w_i - \mu_j}{\sigma_j^2}
\end{aligned} \tag{10}$$

これを式 9 に代入すると

$$\frac{\partial \tilde{E}(\boldsymbol{w})}{\partial w_i} = \frac{\partial E(\boldsymbol{w})}{\partial w_i} + \lambda \sum_j \gamma_j(w_i) \frac{w_i - \mu_j}{\sigma_j^2} \tag{11}$$

となる。右辺第二項が正則化項の効果を表していて、そのポイントは次の二つ。

1. 各重み w_i を混合ガウス分布の平均 μ_j に近づけようとする[*13]。

[*13] これは第二項イコールゼロということから来ている。念のため確認。

2. 上記について近づける度合いは重み w_i の事後分布 $\gamma(w_i)$ に比例する大きさで決定される。

・ガウス分布の中心μ_j

以下、重みによる微分の場合と同じような計算を辿るだけなので計算過程は省略する。期待値による微分については（5.142）を得る[*14]。

・ガウス分布の分散σ_j

分散による微分についても合成関数の微分公式を適当に用いることで、簡単に（5.143）を導くことができる。ここでは分散の動かし方について、分散 σ_j に対応する期待値 μ_j まわりの分散の方に重みをつけて動かすということを行っている。重みの付け方はやはり事後分布 $\gamma(w_i)$ で与えられる。

・混合係数π_j

まず準備として混合係数 π_j を事前分布と解釈し次のような条件付けを行う。

$$\sum_j \pi_j = 1,\ 0 \leq \pi_j \leq 1 \tag{12}$$

これを満たす π_j として以下のようなソフトマックス関数を用いる[*15]。

$$\pi_j = \frac{\exp(\eta_j)}{\sum_{k=1}^{M} \exp(\eta_k)} \tag{13}$$

ここで η は補助変数である[*16]。

ここから誤差関数を η について微分することを考える（対応；演習 5.32)。まず準備として（5.208）の導出を行う。まず式 13 を η_j で微分すると

$$\begin{aligned}
\frac{\partial \pi_k}{\partial \eta_j} &= \frac{\partial}{\partial \eta_j} \frac{\exp(\eta_k)}{\sum_{k=1}^{M} \exp(\eta_k)} \\
&= \frac{\exp(\eta_k)\frac{\partial \eta_k}{\partial \eta_j} \sum_k \exp(\eta_k) - \exp(\eta_k)\exp(\eta_j)}{(\sum_k \exp(\eta_k))^2} \\
&= \frac{\exp(\eta_k)\delta_{kj}}{\sum_k \exp(\eta_k)} - \frac{\exp(\eta_k)\exp(\eta_j)}{(\sum_k \exp(\eta_k))^2} \\
&= \frac{\exp(\eta_j)\delta_{jk}}{\sum_k \exp(\eta_k)} - \frac{\exp(\eta_k)}{\sum_k \exp(\eta_k)} \frac{\exp(\eta_j)}{\sum_k \exp(\eta_k)} \\
&= \pi_j \delta_{jk} - \pi_j \pi_k
\end{aligned} \tag{14}$$

次にこの後の計算のため以下の微分計算をしておく。

$$\begin{aligned}
\frac{\partial}{\partial \eta_j} \sum_k \pi_k N_k &= \frac{\partial}{\partial \pi_k} \sum_k \pi_k N_k \frac{\partial \pi_k}{\partial \eta_j} \ (\because \text{Chain Rules}) \\
&= \sum_k N_k (\pi_j \delta_{jk} - \pi_j \pi_k) \ (\because \text{式 } 14) \\
&= \pi_j N_j - \pi_j \sum_k \pi_k N_k
\end{aligned} \tag{15}$$

[*14] この式の解釈について、重みで微分した場合との違いがよくわからなかった。

[*15] p.d.f. の候補は色々あるだろうに、なぜソフトマックス関数が唐突に出てきたのか正直よくわかっていない。

[*16] 教科書では ξ となっているが、おそらく η の間違いだと思う。

最後に誤差関数の微分を考えるが、実際のところ Ω の部分にのみ注目すれば良い。

$$
\begin{aligned}
\frac{\partial \Omega(\boldsymbol{w})}{\partial \eta_j} &= -\sum_i \left(\frac{\frac{\partial}{\partial \eta_j}\sum_j \pi_j N_j}{\sum_j \pi_j N_j}\right) \\
&= -\sum_i \left(\frac{\frac{\partial}{\partial \eta_j}\sum_k \pi_k N_k}{\sum_j \pi_j N_j}\right) \\
&= \sum_i \left(\frac{\pi_j N_j - \pi_j \sum_k \pi_k N_k}{\sum_j \pi_j N_j}\right) \ (\because \text{式 } 15) \\
&= \sum_i \left(\pi_j - \frac{\pi_j N_j}{\sum_j \pi_j N_j}\right) \\
&= \sum_i (\pi_j - \gamma(w_i)) \qquad (16)
\end{aligned}
$$

これは π_j が事後分布の方向へ引き寄せられることを意味している。以上より正則化項の導入によって重みの期待値や分散が学習過程を通じて決定されることがわかる。

[p.269 上部 式導出]

(5.130) の $g(\phi(x,\xi))$ に p.268 下部 $g(\phi(x,\xi))$ の式を代入する。

$$\tilde{E} = \frac{1}{2}\iiint \{y(x) + \xi^T \Delta y(x) + \frac{\xi^2}{2}[\Delta^T \Delta y(x) + (\Delta y(x))^T \Delta^T \Delta] + O(\xi^3) - \tau\}^2 \times p(\tau|x)p(x)p(\xi)\, d\xi\, dx\, d\tau$$

以下カンタンのため3次のオーダーより以上のところは書かない。最後に $O(\xi^3)$ として復活させる。

$$\tilde{E} = \frac{1}{2}\iiint \{y(x) - \tau + \xi^T \Delta y(x) + \frac{1}{2}\xi^T [\Delta (\Delta y(x))^T + (\Delta y(x)) \Delta^T]\xi\}^2 \times p(\xi)p(x)p(\tau|x)\, d\xi\, dx\, d\tau$$

$$= \frac{1}{2}\iiint (y(x) - \tau)^2 p(\tau|x)p(x)p(\tau)\, d\xi\, dx\, d\tau$$

$d\xi\, dx\, d\tau$

$$+ \frac{1}{2}\iiint (y(x) - \tau)\{\xi^T \Delta y(x) + \frac{1}{2}\xi^T [\Delta (\Delta y(x))^T + (\Delta y(x)) \Delta^T]\xi\} p(\xi)p(x)p(\tau|x)\, d\xi\, dx\, d\tau +$$

$$+ \frac{1}{2}\iiint (\xi^T \Delta y(x))^2 p(\xi)p(x)p(\tau|x)\, d\xi\, dx\, d\tau +$$

$$= E + \iiint (y(x) - \tau)\xi^T \Delta y(x)\, p(\xi)p(x)p(\tau|x)\, d\xi\, dx\, d\tau \quad (\because (5.129))$$

$d\xi\, dx\, d\tau$

$$+ \frac{1}{2}\iiint [(y(x) - \tau)\{\frac{1}{2}\xi^T [\Delta (\Delta y(x))^T + (\Delta y(x)) \Delta^T]\xi\} + (\xi^T \Delta y(x))^2]\, p(\xi)p(x)p(\tau|x)\, d\xi\, dx\, d\tau +$$

$$= E + E(\xi) \iint (y(x) - \tau)\Delta^T \Delta y(x)\, p(x)p(\tau|x)\, dx\, d\tau$$

$dx\, d\tau\, d\xi$

$$+ \frac{1}{2}E(\xi^2) \iint [(y(x) - \tau)\{\frac{1}{2}\xi^T [\Delta (\Delta y(x))^T + (\Delta y(x)) \Delta^T]\xi\} + (\xi^T \Delta y(x))^2]\, p(x)p(\tau|x)\, dx\, d\tau +$$

$$+ O(\xi^3)$$