〜これまで〜

5.1でニューラルネットワークモデルの構造の説明、5.2、5.3でモデルの学習方法の説明、5.4では汎化性能を上げる話をしてきた。ここでは、回帰問題のうち特殊な問題を解くためのモデルを扱う。

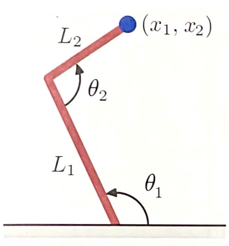
5.6 混合密度ネットワーク

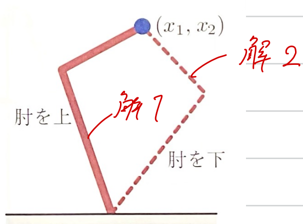
5.1- 5.5 節までは回帰問題の誤差関数に最小二乗誤差を用いてきた(5.11式)。これは誤差項にガウス分布を仮定しているが、現実にはそのような仮定をおけることは珍しい。例えば確率密度が多峰性を持つような目的変数に対しては、貧弱な予測しか得られない。混合密度ネットワークはこのような目的変数に対して有効な予測ができる。本節では、混合密度ネットワークで解決できる問題の提示と、学習方法の紹介、そして評価を行う。

【今回扱いたい問題】

混合密度ネットワークがなんであるかを紹介する前に、今回解決したい問題を明確にする。結論から述べると、逆問題に対する予測を行う際に、これまでのガウスノイズを仮定したニューラルネットワークモデルがうまく機能しなくなる。

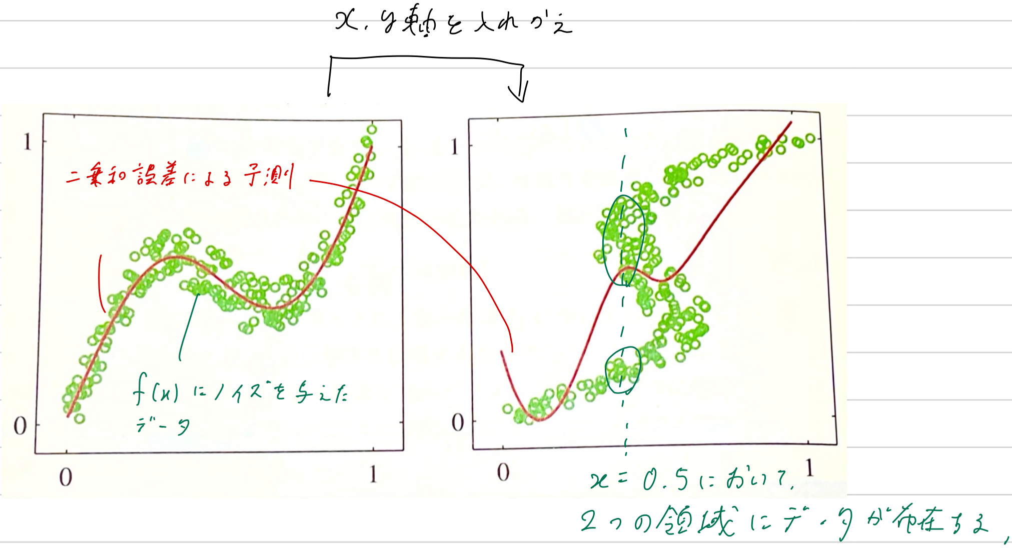
逆問題の例として、ロボットアームの例を紹介する。下図のように2本の連結したアームが、ある関節角で繋がれているときことを考える。もし、関節角が得られていて、アームの終端位置を知りたい場合、終端位置は一意に特定することができる。これは原因から結果を求める問題で、順問題と呼ばれ、比較的容易に解ける。一方で、アームの終端位置が得られていて、対応する関節角を知りたい場合、結果から原因を求める問題となり、逆問題と呼ばれる。逆問題は、複数の関節角の組み合わせから終端位置が合うような組み合わせを探す問題となり難易度が高くなり、解が非連続な複数の値となりうる。





逆問題は因果関係のある変数のうち，結果の変数から原因の変数を求める際に出現する。ロボットアームの例の他にも、例えば病人の症状から人体の疾患を予測する例などがある。本節ではこのような問題を取り扱うに当たって、非常に簡略化した逆問題を考える。すなわち、下図のように $f(x) = x + 0.3\sin(2\pi x)$ の値にノイズを加えたサンプルのx軸とy軸を入れ替えたデータに対して、新しいy軸の値を予測することを考える。

この例では、f(x)（結果）をもとに、x（原因）の値を予測するため逆問題に相当する。二乗和誤差を用いた学習では、このような問題に良い予測を与えることができない（図中赤線）。というのも、二乗和誤差を用いた学習はノイズに単峰性のガウス分布を仮定しており、複数の分離した領域にまたがって存在するデータに対してはフィットしないためである。



【モデルの構造】

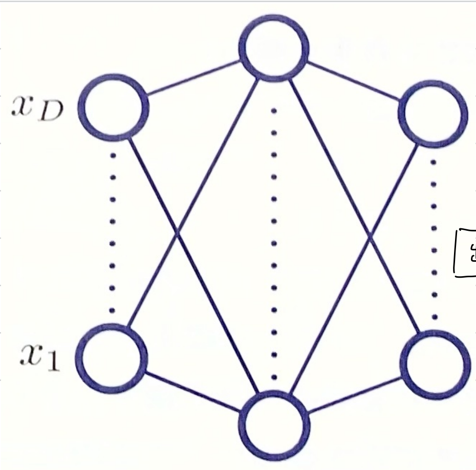
混合密度ネットワークは、条件付き密度関数 $p(t|x)$ に多峰性の分布を仮定することでこのような問題に対して良い予測を与える。これは、与えられた x に対して、取りうる予測値が複数の領域に分布することを表現できるためである。混合密度ネットワークの例として、条件付き密度分布に混合ガウス分布を用いるモデルを考える。

$$p(t|x) = \sum_{k=1}^{K} \pi_k(x) \mathcal{N}(t \mid \mu_k(x), \sigma_k^2(x) I)$$

入力ベクトル x の関数

このとき、パラメータ $\pi_k(x), \mu_k(x), \sigma_k^2(x)$ を、ニューラルネットワークの出力に対応させる。つまり、ニューラルネットワークを順伝搬させた際に得られる出力が、混合モデルのパラメータとなる。混合モデルが K 個の要素を持ち、t が L 個の要素を持つとき、ネットワークの出力の総数は $(L+2)K$ 個となる。（後述）

実際には、パラメータが取るべき値の制約に応じて、出力を変換したものをパラメータに使用する。（教科書 p.277 の式 5.149 - 5.152 の読み合わせとする）
式中の変数 a はネットワークの出力をパラメータに対応させたものに相当する。



出力

π にまつわる出力

$$\left.\begin{matrix} a^{\pi}_1 \\ \vdots \\ a^{\pi}_K \end{matrix}\right\} \quad \pi_k = \frac{\exp(a_k^{\pi})}{\sum_{l=1}^{K} \exp(a_l^{\pi})}$$

σ にまつわる出力

$$\left.\begin{matrix} a^{\sigma}_1 \\ \vdots \\ a^{\sigma}_K \end{matrix}\right\} \quad \sigma_k = \exp(a_k^{\sigma})$$

μ にまつわる出力

$$\left.\begin{matrix} a^{\mu}_{11} \\ \vdots \\ a^{\mu}_{KL} \end{matrix}\right\} \quad \mu_{kj} = a^{\mu}_{kj}$$

ネットワークの出力の数について考えてみる。

まず、 a_k^{π} の個数は、混合分布におけるグループの個数に相当するので、K個である。 a_k^{σ} の個数は、各ガウス分布につき一つの分散パラメータが存在するので、こちらもグループの個数だけ存在し、K個である。最後に、a_{kl}^{μ} については、各ガウス分布につき L 個の平均パラメータが存在するので、合計で KL 個である。

よって、出力全体の数は、$K + K + LK = (L+2)K$ である。

【混合密度ネットワークの学習】

ここまで、今回解くべき問題と、混合密度ネットワークの構造について見てきた。最後にモデルの学習方法を学ぶ。方針としては負の対数尤度を定義して、誤差逆伝播法を適用することで学習を進める。

誤差逆伝播法の学習方法がどのようなものであったか、過去のレジュメからおさらいすることから始めよう。（210327輪講.pdf の 1頁より）

5.3.2 単純な例

5.3.1 では、誤差逆伝播法の具体的な手続きについて学んだ。その手続きとは次のようなものであった。

1. 入力ベクトルをネットワークに入れ順伝搬させ、全てのユニットの出力を求める。
2. 出力ユニットにおける誤差 $\delta_k = y_k - t_k$ を算出する。
3. 隠れユニットにおける誤差 $\delta_j = h'(a_j) \sum_k w_{kj} \delta_k$ を逆伝搬により求める。
4. 全ユニットで微分値 $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$ を評価する。

ステップ1については、これまでの構造上問題なく実施できる。ステップ2はおいておいて、ステップ3についてもステップ2が完了すれば問題なく実施できる。ステップ4も、ステップ2、3が実施できれば実施可能である。

問題のステップ2に着目する。このときの例では、ネットワークの最終出力を $y_k = a_k$ で表現しており、誤差関数には二乗和誤差 $(y_k - t_k)^2$ を用いていたため、誤差を次のように算出していた。

$$\delta_k = \frac{\partial E}{\partial a_k} = \frac{\partial}{\partial y_k} \sum_{l=1}^{K} (y_l - t_l)^2 = y_k - t_k$$

（ステップ2の式に相当）

一方で、今回の混合密度ネットワークの例では、誤差関数に負の対数尤度を使用する。ステップ2における $\delta_k = \frac{\partial E}{\partial a}$ さえ計算できれば、ステップ3、4についても実施できるため、誤差逆伝播法を適用できる。

データが1つ得られた際の負の対数尤度は次のような式である。

$$E = -\ln\left\{\sum_{k=1}^{K}\pi_k(x_n, w)\,\mathcal{N}(t_n \mid \mu_k(x_n, w),\ \sigma_k^2(x_n, w)I)\right\}$$

$$= -\ln\left\{\sum_{k=1}^{K}\pi_k \mathcal{N}_{nk}\right\}$$ (略記法を用いる)

以下、$\dfrac{\partial E}{\partial a_k^\pi}$、$\dfrac{\partial E}{\partial a_{kl}^\mu}$、$\dfrac{\partial E}{\partial a_k^\sigma}$ を順に求めることとする。

①
$$\frac{\partial E}{\partial a_k^\pi} = \sum_{l=1}^{K}\left\{\frac{\partial E}{\partial \pi_l}\right\}\cdot\left(\frac{\partial \pi_l}{\partial a_k^\pi}\right)$$

ソフトマックスの微分 P.250 (5.76)

$$= \sum_{l=1}^{K}\left\{-\frac{\mathcal{N}_{nl}}{\sum_{k=1}^{K}\pi_k\mathcal{N}_{nk}}\right\}(\delta_{kl}\pi_l - \pi_l\pi_k)$$

$$= \sum_{l=1}^{K}\left\{-\frac{\mathcal{N}_{nl}}{\sum_{k=1}^{K}\pi_k\mathcal{N}_{nk}}\right\}\delta_{kl}\pi_l - \sum_{l=1}^{K}\left\{-\frac{\mathcal{N}_{nl}}{\sum_{k=1}^{K}\pi_k\mathcal{N}_{nk}}\right\}\pi_l\pi_k$$

$$= -\frac{\mathcal{N}_{nk}}{\sum_{k=1}^{K}\pi_k\mathcal{N}_{nk}}\cdot\pi_k + \pi_k\sum_{l=1}^{K}\left\{\frac{\pi_l\mathcal{N}_{nl}}{\sum_{k=1}^{K}\pi_k\mathcal{N}_{nk}}\right\}$$

$l = k$ のみのこる。

$$= -\frac{\pi_k\mathcal{N}_{nk}}{\sum_{k=1}^{K}\pi_k\mathcal{N}_{nk}} + \pi_k\sum_{l=1}^{K}\left\{\frac{\pi_l\mathcal{N}_{nl}}{\sum\pi_k\mathcal{N}_{nk}}\right\}$$

$$= -\gamma_{nk} + \pi_k\sum_{l=1}^{K}\gamma_{nl}$$

$$= \pi_k - \gamma_{nk}$$

② $$\frac{\partial E}{\partial a^{\mu}_{kl}} = \frac{\partial E}{\partial \mu_{kl}} \cdot \frac{\partial \mu_{kl}}{\partial a^{\mu}_{kl}} = \frac{\partial E}{\partial \mu_{kl}} \cdot \frac{\partial a^{\mu}_{kl}}{\partial a^{\mu}_{kl}}$$

$$= \frac{\partial E}{\partial \mu_{kl}}$$

おきかえ

$$= -\frac{1}{\Sigma \pi_k N_{nk}} \cdot \frac{\partial}{\partial \mu_{kl}} \left\{ \sum_{l=1}^{K} \pi_l N_{nl} \right\}$$

グループkのガウス分布の平均なので、$l=k$ のみのこるはず。

$$= -\frac{\pi_k}{\Sigma \pi_k N_{nk}} \cdot \frac{\partial N_{nk}}{\partial \mu_{kl}}$$

$$= -\frac{\pi_k}{\Sigma \pi_k N_{nk}} \cdot \frac{\partial}{\partial \mu_{kl}} \left[\frac{1}{(2\pi)^{L/2}} \frac{1}{|\sigma_k^2 I|^{1/2}} \exp\left\{ -\frac{1}{2\sigma_k^2} \sum_{s=1}^{L} (t_{ns} - \mu_{ns})^2 \right\} \right]$$

$$= -\frac{\pi_k}{\Sigma \pi_k N_{nk}} \cdot N_{nk} \cdot \frac{\partial}{\partial \mu_{kl}} \left\{ -\frac{1}{2\sigma_k^2} \sum_{s=1}^{L} (t_{ns} - \mu_{ns})^2 \right\}$$

$$= -\gamma_{nk} \cdot \left(-\frac{1}{2\sigma_k^2} \right) \cdot 2 (t_{nl} - \mu_{nl})(-1)$$

$$= \gamma_{nk} \left\{ \frac{\mu_{nl} - t_{nl}}{\sigma_k^2} \right\}$$

③ $$\frac{\partial E}{\partial a^{\sigma}_{k}} = \frac{\partial E}{\partial \sigma_k} \cdot \frac{\partial \sigma_k}{\partial a^{\sigma}_{k}} = \frac{\partial E}{\partial \sigma_k} \cdot \frac{\partial e^{a^{\sigma}_{k}}}{\partial a^{\sigma}_{k}}$$

$$= \frac{\partial}{\partial \sigma_k} \left\{ -\ln \sum_{k=1}^{K} (\pi_k N_{nk}) \right\} \cdot e^{a^{\sigma}_{k}}$$

$$= -\frac{\pi_k}{\Sigma \pi_k N_{nk}} \cdot \frac{\partial}{\partial \sigma_k} \left[\frac{1}{(2\pi)^{L/2}} \frac{1}{|\sigma_k^2 I|^{1/2}} \exp\left\{ -\frac{1}{2\sigma_k^2} \sum_{s=1}^{L} (t_{ns} - \mu_{ns})^2 \right\} \right] \cdot \sigma_k$$

$$= -\frac{\pi_k}{\Sigma \pi_k N_{nk}} \cdot \frac{1}{(2\pi)^{L/2}} \left[\frac{\partial}{\partial \sigma_k} \left(\frac{1}{\sigma_k^L} \right) \exp\{\cdots\} + \left(\frac{1}{\sigma_k^L} \right) \frac{\partial}{\partial \sigma_k} \exp(\cdots) \right] \sigma_k$$

$$
\begin{aligned}
&= -\frac{\pi_k}{\sum \pi_k N_{nk}} \cdot \frac{1}{(2\pi)^{L/2}} \Big[(-L)\, \sigma^{-L-1} \exp(\cdots) \cdot \sigma \\
&\qquad\qquad + \sigma^{-L} \exp(\cdots) \frac{\partial}{\partial \sigma_k} \left\{ -\frac{1}{2\sigma_k^2} \sum_{s=1}^{L} (t_{ns} - \mu_{ns})^2 \right\} \cdot \sigma \Big] \\
&= -\frac{\pi_k}{\sum \pi_k N_{nk}} \cdot \frac{1}{(2\pi)^{L/2}} \sigma^{-L} \exp(\cdots) \left\{ (-L) - \frac{\|t_n - \mu_n\|^2}{2} \cdot (-2)\, \sigma_k^{-3} \cdot \sigma \right\} \\
&= -\frac{\pi_k N_{nk}}{\sum \pi_k N_{nk}} \left\{ -L + \frac{\|t_n - \mu_n\|^2}{\sigma_k^2} \right\} \\
&= \gamma_{nk} \left(L - \frac{\|t_n - \mu_n\|^2}{\sigma_k^2} \right)
\end{aligned}
$$

以上で、誤差関数に対する出力の微分値が求まった。再度やりたいことを振り返ると、誤差逆伝播法のステップ2を実行したいのであった。$\frac{\partial E}{a_k^{\pi}}$, $\frac{\partial E}{\partial a_{kl}^{\mu}}$, $\frac{\partial E}{\partial a_k^{\sigma}}$ は δ_k に相当するのでこれにより残りのステップについても実行できるようになった。

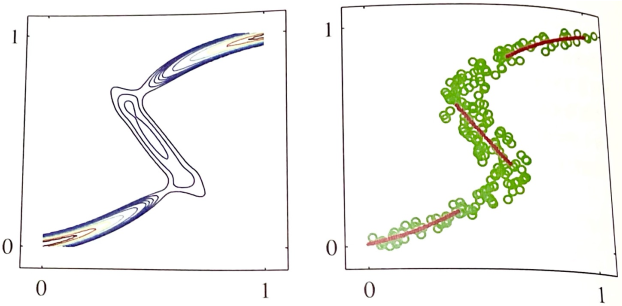
1. 入力ベクトルをネットワークに入れ順伝搬させ、全てのユニットの出力を求める。
2. 出力ユニットにおける誤差 $\delta_k = y_k - t_k$ を算出する。
3. 隠れユニットにおける誤差 $\delta_j = h'(a_j) \sum_k w_{kj} \delta_k$ を逆伝搬により求める。
4. 全ユニットで微分値 $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$ を評価する。

以上をもって、混合密度ネットワークは誤差逆伝播法で学習できることが示された。

【混合密度ネットワークの評価】

これまでで、本節で解きたい問題と、それに対する予測モデルの構造、学習の手続きについて紹介してきた。最後に、簡易的なサンプルデータに対して学習を行なった結果を見て、その有用性を考察する。

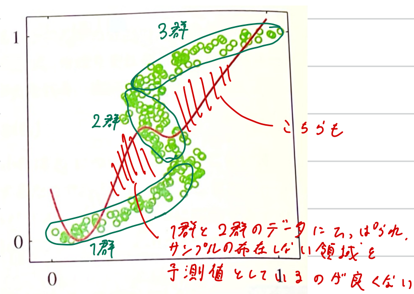
先ほど挙げた簡易的な例について混合密度ネットワークを学習させた結果、条件付き確率密度 $p(t|x)$ の等高線をプロットしたものが下図の（c）である。



(c)　　　(d)

混合密度ネットワークの出力としては予測分布が求まるが、あえてここからある1点を予測値として取り出す場合の応用について考えてみると、今回の手法の「うまみ」がわかる。

ガウス分布をノイズに仮定した学習の場合、サンプルが存在する領域の間を予測値とするケースがあった（右図）。いわゆる、データが両方のサンプルに引っ張られて、どっちつかずな予測をしている状態と言える。一方で、混合密度ネットワークでは、条件付き確率の最頻値を予測値とすることで、サンプルの存在する領域を予測値とすることができており（上図 (d) ）、こちらの方がよっぽど良い予測と言える。



まとめると、今回は以下のことを学んだ。

- 逆問題については、これまでのガウスノイズを仮定した学習モデルが機能しなくなる
- 混合密度ネットワークは、条件付き確率を混合分布で表現するモデルで逆問題に対する予測に有効である。
- 混合密度ネットワークは誤差逆伝播法によって学習可能である。