階層型ネットワークの階層型ネットワーク

浅川伸一 <asakawa@twcu.ac.jp>

1 はじめに

Mixture of experts (以下 ME と略記) と呼ばれる階層型のネットワークを紹介します。ME は入力データ空間をいくつかの小領域に分割し、その分割された各領域に対してひとつのニューラルネットワークを割り当てることによって複雑な問題の解を求めるため手法です。複雑な問題を分割して小領域に区切ることによって、1つの大きなニューラルネットワークを学習させるよりも効率の良い学習をさせようというモデルです。このような手法を分割統治 divide–and–conquer と言ったりします。ME における学習とは、入力空間の分割の仕方と、各小領域に属するデータに対する最適な答えとを見つけ出すことです。

分割統治 は科学における一般原理であるといって良いでしょう。この分割統治を自動的に行ない学習させようと言うのが ME の発想です。ME はローカルネットワークとゲートネットワークから成り立っています。ME は教師あり学習の一手法です。

2 エキスパートの階層

入力ベクトルが m 次元の実数値をとるベクトル $\mathcal{R}^m$ で、出力が n 次元の実数値をとる ベクトル $\mathcal{R}^n$ で定義された要素であるとします。観測されたデータを $\mathcal{X} = \left\{\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}\right\}$ とします。ME の木構造において非末端部にはゲーティングネットワークがあり、1つの数値 (確率と考えて良い) を出力します。一方、末端部分にはエキスパートネットワークがあります。各エキスパートネットワークは入力ベクトル $\boldsymbol{x}$ に対して出力ベクトル $\boldsymbol{\mu}_{ij}$ を出力します。各々の出力ベクトルは木を登って行きゲーティングネットワークによって混合されます。

エキスパートネットワーク (i, j) は入力ベクトル $\boldsymbol{x}$ に対する出力としてベクトル $\boldsymbol{\mu}_{ij}$ を出力します。

$$\boldsymbol{\mu}_{ij} = f\left(\mathcal{U}_{ij}\boldsymbol{x}\right) \tag{1}$$

ここで $\mathcal{U}_{ij}$ は結合係数行列で、入力 $\boldsymbol{x}$ は切片項を表す常に 1 を出力する固定の要素を含みます。関数 f のことを入力と出力を結びつけるという意味で

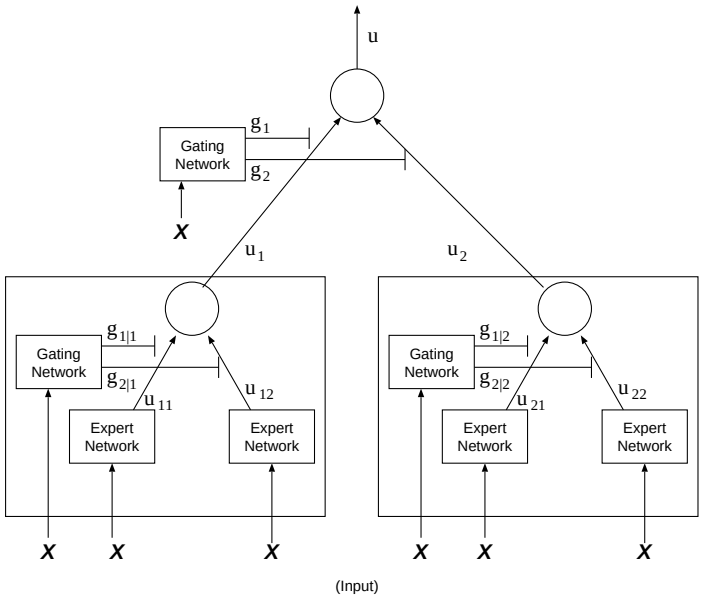

図 1: 2 段階の ME

リンク関数と言います。f は問題によって線形でもシグモイド関数でも、あるいはガウシアン関数でもなんでも良いのです。

ゲーティングネットワークは線形であり ξ を変数とする以下の式のように表されます。

$$\xi_i = \boldsymbol{v}_i^T \boldsymbol{x} \tag{2}$$

ここで $\boldsymbol{v}_i$ は結合係数ベクトルです。このとき最上位のゲーティングネットワークの i 番目の出力は ξ_i のソフトマックス関数として定義されます。

$$g_i = \frac{e^{\xi_i}}{\sum_k e^{\xi_k}} \tag{3}$$

g_i は正であり、すべての g_i について合計すると必ず 1 になります。すなわちソフトマックス関数は入力空間を「柔らかく」分割するものだといえます (第 3 節参照)。

同じようにして下位のゲーティングネットワークも線形であり、出力 ξ_{ij} は以下のように定義されます。

$$\xi_{ij} = \boldsymbol{v}_{ij}^T \boldsymbol{x} \tag{4}$$

このとき

$$g_{j|i} = \frac{e^{\xi_{ij}}}{\sum_k e^{\xi_{ik}}} \tag{5}$$

は 2 段目の i 番目のゲーティングネットワークにおける j 番目のゲーティングネットワークの出力です。ここでも $g_{j|i}$ は正であり各 $\boldsymbol{x}$ に対して足し合わ

せると 1 になります。非末端部 (ノード) における出力は、そのノードの下にあるエキスパートネットワークの出力に各ゲーティングネットワークの重みをかけた出力になります。すなわち第 2 層のノード部における出力は

$$\boldsymbol{\mu}_i = \sum_j g_{j|i}\boldsymbol{\mu}_{ij} \tag{6}$$

となり、木の最上位層ゲーティングネットワークの出力は

$$\boldsymbol{\mu} = \sum_i g_i\boldsymbol{\mu}_i = \sum_i g_i \sum_j g_{j|i}\boldsymbol{\mu}_{ij} \tag{7}$$

となります。g も $\boldsymbol{\mu}$ も入力 $\boldsymbol{x}$ に依存しています、故に、全体の出力は入力ベクトル $\boldsymbol{x}$ の非線型関数となります。g_i は $\boldsymbol{v}_i$ と $\boldsymbol{x}$ とに依存する関数であることを明記するために $g_i(\boldsymbol{v}_i, \boldsymbol{x})$ と表記することもあります。

3 リッジ回帰

ソフトに分割された領域とは、領域間に重なりがあると言う意味です。この重複の性質を理解するために 2 つのエキスパートからなる 1 つの階層を考えてみることにします。この場合、ゲーティングネットワークは g_1 と g_2 との 2 つの出力を持ちます。ゲーティング出力 g_1 は

$$g_1 = \frac{e^{\xi_1}}{e^{\xi_1} + e^{\xi_2}} \tag{8}$$

$$= \frac{1}{1 + e^{-(\boldsymbol{v}_1 - \boldsymbol{v}_2)\boldsymbol{x}}} \tag{9}$$

と表されます。これは $\boldsymbol{v}_1 - \boldsymbol{v}_2$ というベクトルの方向によって決まるシグモイド関数です。ゲーティング出力 g_2 は $1 - g_1$ に等しいことになります。あ

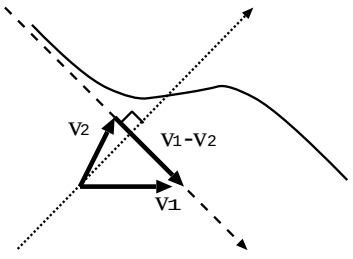

図 2: ソフトマックスとリッジ

る $\boldsymbol{x}$ が与えられると全体の出力 $\boldsymbol{\mu}$ は $g_1\boldsymbol{\mu}_1 + g_2\boldsymbol{\mu}_2$ によって与えられます。

これはエキスパートの重みづけ平均となり、重みはリッジ関数の値によって決まります。$g_1 = g_2 = 1/2$ というリッジに沿って 2 つのエキスパートは等しく貢献します。このリッジから離れるに従って 1 つのエキスパートの影響が大きくなり、他方の影響は小さくなります。

$$\begin{aligned} g_1 = g_2 = \frac{1}{2} &= \frac{1}{1 + e^{-(\boldsymbol{v}_1 - \boldsymbol{v}_2)\boldsymbol{x}}} \\ 1 &= \frac{2}{1 + e^{-(\boldsymbol{v}_1 - \boldsymbol{v}_2)\boldsymbol{x}}} \\ e^{-(\boldsymbol{v}_1 - \boldsymbol{v}_2)\boldsymbol{x}} &= 1 \\ -(\boldsymbol{v}_1 - \boldsymbol{v}_2)\,\boldsymbol{x} &= 0 \end{aligned}$$

という計算により $\boldsymbol{v}_1 - \boldsymbol{v}_2$ というベクトルに直交するベクトルの方向では 2 つのエキスパートの影響は等しくなります。リッジを横切る滑らかさの度合は $\boldsymbol{v}_1 - \boldsymbol{v}_2$ というベクトルの大きさによって決まります。もし $\boldsymbol{v}_1 - \boldsymbol{v}_2$ が大きければリッジ関数は 2 つの領域をハッキリと分離することになり、エキスパートの重み付された出力は線形に近くなります。もし、$\boldsymbol{v}_1 - \boldsymbol{v}_2$ が小さいのなら各エキスパートはリッジの両側において意味のある程度まで貢献し、なめらかなマップになります。$\boldsymbol{v}_1 - \boldsymbol{v}_2$ の値が大きい程分離の程度が大きくなり、この重みが 0 になる極限では両エキスパートは等しく貢献することになります。

4 ME の確率的解釈

ME の出力ベクトル $\boldsymbol{y}$ を、入力データベクトル $\boldsymbol{x}$ とデータの密度関数を表すパラメータ $\boldsymbol{\theta}$ とを用いて条件付確率として定式化することができます。

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_i g_i(\boldsymbol{v}_i, \boldsymbol{x}) \sum_j g_{j|i}(\boldsymbol{v}_{ij}, \boldsymbol{x})\, P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_{ij}) \tag{10}$$

ここで $\boldsymbol{\theta}_{ij}$ は分布の密度関数を定めるパラメータベクトルです。例えば $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ が正規分布に従うという仮定の下では $\boldsymbol{\theta}$ の成分には平均と分散が含まれます。データが多次元正規分布に従い、分散共分散行列が $\Sigma = \sigma^2 \boldsymbol{I}$ (ただし $\boldsymbol{I}$ は単位行列) となるならば、データ $\boldsymbol{x}$ と $\boldsymbol{\theta}$ とが与えられたときの密度関数は次のように書くことができます。

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \sum_i g_i \sum_j g_{j|i}\, e^{-(1/2\sigma^2)(\boldsymbol{y} - \boldsymbol{\mu}_{ij})^T (\boldsymbol{y} - \boldsymbol{\mu}_{ij})} \tag{11}$$

分散パラメータ σ^2 は $\boldsymbol{\mu}_{ij}$ を中心とする円の大きさ (半径) を定めるものと解釈することが出来ます。$\sigma^2 \to 0$ の極限ではディラックのデルタ関数となります。

y が離散的な 2 値関数、すなわち多次元ベルヌーイ試行ならば

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_i g_i \sum_j g_{j|i} u_{ij}^y (1 - u_{ij})^{1-y} \tag{12}$$

となります。

5 パラメータの推定

5.1 事後確率

ME ニューラルネットワークの学習を考えるために ME の木構造のノードにおける事後確率 h を定義します。g_i と $g_{j|i}$ とは事前確率と呼ぶことにします。なぜなら g_i, $g_{j|i}$ は入力データベクトル $\boldsymbol{x}$ だけで決まり、対応する出力ベクトル $\boldsymbol{y}$ の知識なしに定まるからです。事後確率は入力ベクトル $\boldsymbol{x}$ と出力ベクトル $\boldsymbol{y}$ の両方が分かったとき定義されます。ベイズの定理を用いて i 番目のノードの事後確率は次のように定義されます。

$$h_i = \frac{g_i \sum_j g_{j|i} P_{ij}(\boldsymbol{y}|\boldsymbol{x})}{\sum_i g_i \sum_j g_{j|i} P_{ij}(\boldsymbol{y}|\boldsymbol{x})} \tag{13}$$

$$h_{j|i} = \frac{g_{j|i} P_{ij}(\boldsymbol{y}|\boldsymbol{x})}{\sum_j g_{j|i} P_{ij}(\boldsymbol{y}|\boldsymbol{x})} \tag{14}$$

g_i, と $g_{j|i}$ は (3), (5) 式で定義されています。

また、h_i と $h_{j|i}$ との積、すなわち結合事後確率 h_{ij} を定義しておきます。この量はエキスパートネットワーク (i, j) がデータを生成したと見なすことができる確率です。

$$h_{ij} = \frac{g_i g_{j|i} P(\boldsymbol{y}|\boldsymbol{x})}{\sum_i g_i \sum_j g_{j|i} P_{ij}(\boldsymbol{y}|\boldsymbol{x})} \tag{15}$$

5.2 尤度

N 個のデータ集合 $\mathcal{X} = \left\{\left(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}\right)\right\}_1^N$ から次の対数尤度関数が得られます。

$$l\left(\boldsymbol{\theta}; \mathcal{X}\right) = \sum_t \log \sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} P_{ij}(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}) \tag{16}$$

密度関数 P がガウシアンで共分散行列が単位行列であり、リンク関数が恒等関数である場合を考えます。このとき $l\left(\boldsymbol{\theta}; \mathcal{X}\right)$ を各々のパラメータで微分することで以下のような結合係数行列の勾配上昇学習則を得ます。

$$\Delta U_{ij} = \rho \sum_t h_i^{(t)} h_{j|i}^{(t)} \left(\boldsymbol{y}^{(t)} - \boldsymbol{\mu}_{ij}^{(t)}\right) \boldsymbol{x}^{(t)T} \tag{17}$$

ここで ρ は学習率です。一番上のゲーティングネットワークにおける i 番目の結合係数ベクトルの勾配上昇学習則は

$$\Delta \boldsymbol{v}_i = \rho \sum_t \left(h_i^{(t)} - g_i^{(t)}\right) \boldsymbol{x}^{(t)} \tag{18}$$

で与えられます。i 番目の下位レベルのゲーティングネットっワークにおける j 番目の結合係数ベクトルの勾配上昇学習則は

$$\Delta \boldsymbol{v}_{ij} = \rho \sum_t h_i^{(t)} \left(h_{j|i}^{(t)} - g_{j|i}^{(t)} \right) \boldsymbol{x}^{(t)} \tag{19}$$

式 (17),(18),(19) で与えられるアルゴリズムはバッチ学習アルゴリズムです。対応するオンラインアルゴリズムは単純に加算記号を落して、刺激提示毎にパラメータを更新することで得られます。例えば

$$U_{ij}^{(t+1)} = U_{ij}^{(t)} + \rho h_i^{(t)} h_{j|i}^{(t)} \left(\boldsymbol{y}^{(t)} - \boldsymbol{\mu}^{(t)} \right) \boldsymbol{x}^{(t)T} \tag{20}$$

は t 番目の刺激パターンに基づく (i,j) 番目のエキスパートネットワークの結合係数に対する確率的更新則です。

5.3 EM アルゴリズム

EM アルゴリズムによって ME ネットワークのパラメータを求めることを考えます。まず尤度関数が簡単になるような隠れ変数を定義します。変数 z_i はすべての i の中のたった 1 つの z_i だけが 1 となり、他の全ては 0 であるとします。同様に変数 $z_{j|i}$ はたった 1 つの $z_{j|i}$ だけが 1 であり、他はすべて 0 であるとします。これらの隠れ変数は確率モデルにおける決定に対応したラベルであると解釈できます。また z_i と $z_{j|i}$ との積である z_{ij} を定義しておきます。z_{ij} は特定のエキスパートネットワークを示すラベルであると解釈できます。もし、ラベル z_i, $z_{j|i}$, z_{ij} が分かれば尤度最大化問題は各エキスパートネットワークの回帰問題に分離され、ゲーティングネットワークについては分割された集合の分類問題に切り離され、これらの問題は互いに独立に解くことができます。もちろん隠れ変数は未知なのですが、隠れ変数と観測可能なデータとを結びつける確率モデルを特定することができます。この確率モデルは z_{ij} の項を用いて

$$\begin{aligned} P\left(\boldsymbol{y}^{(t)}, z_{ij}^{(t)} | \boldsymbol{x}^{(t)}, \boldsymbol{\theta} \right) &= g_i^{(t)} g_{j|i}^{(t)} P_{ij}(\boldsymbol{y}^{(t)} | \boldsymbol{x}^{(t)}) & (21) \\ &= \prod_i \prod_j \left\{ g_i^{(t)} g_{j|i}^{(t)} P_{ij}(\boldsymbol{y}^{(t)} | \boldsymbol{x}^{(t)}) \right\}^{z_{ij}^{(t)}} & (22) \end{aligned}$$

と書けます。この確率の対数をとると次の完全データ尤度が求められます。

$$\begin{aligned} l_c\left(\boldsymbol{\theta}; \mathcal{Y}\right) &= \sum_t \sum_i \sum_j z_{ij}^{(t)} \log \left\{ g_i^{(t)} g_{j|i}^{(t)} P_{ij}(\boldsymbol{y}^{(t)}) \right\} & (23) \\ &= \sum_t \sum_i \sum_j z_{ij} \left\{ \log g_i^{(t)} + \log g_{j|i}^{(t)} + \log P_{ij}(\boldsymbol{y}^{(t)}) \right\} & (24) \end{aligned}$$

上式で表現された完全データ尤度と (16) で定義された不完全データ尤度の関係に注意して下さい。指示変数 z_{ij} を用いることで対数を加算記号の内側

に持って来ることが可能になりました。このことによって最大化問題を簡単にすることができます。完全データ尤度の期待値をとることで EM アルゴリズムの E ステップを

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(p)}) = \sum_t \sum_i \sum_j h_{ij}^{(t)} \left\{ \log g_i^{(t)} + \log g_{j|i}^{(t)} + \log P_{ij}(\boldsymbol{y}^{(t)}) \right\} \tag{25}$$

と定義します。ここで以下の事実を使いました。

$$\begin{aligned} E\left[z_{ij}^{(t)}|\mathcal{X}\right] &= P\left(z_{ij}^{(t)} = 1|\boldsymbol{y}^{(t)}, \boldsymbol{x}^{(t)}, \boldsymbol{\theta}^{(p)}\right) && (26) \\ &= \frac{P\left(\boldsymbol{y}^{(t)}|z_{ij}=1, \boldsymbol{x}^{(t)}, \boldsymbol{\theta}^{(p)}\right) P\left(z_{ij}^{(t)} = 1|\boldsymbol{x}^{(t)}, \boldsymbol{\theta}^{(p)}\right)}{P\left(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}, \boldsymbol{\theta}^{(p)}\right)} && (27) \\ &= \frac{P\left(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}, \boldsymbol{\theta}_{ij}^{(p)}\right) g_i^{(t)} g_{j|i}^{(t)}}{\sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} P\left(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}, \boldsymbol{\theta}^{(p)}\right)} && (28) \\ &= h_{ij}^{(t)} && (29) \end{aligned}$$

すなわち z_{ij} の期待値は事後確率 h_{ij} になります。同様にして $E\left[z_i^{(t)}|\mathcal{X}\right] = h_i^{(t)}$、$E\left[z_{j|i}^{(t)}|\mathcal{X}\right] = h_{j|i}^{(t)}$ が導出できます。

EM アルゴリズムの M ステップでは $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(p)})$ をエキスパートネットワークパラメータとゲーティングネットワークパラメータに関して最大化する必要があります。式 (25) を見ると、エキスパートネットワークパラメータは $h_{ij}^{(t)} \log P(\boldsymbol{y}^{(t)})$ をとおしてだけ $\mathcal{Q}$ に影響し、ゲーティングネットワークパラメータは $h_{ij}^{(t)} \log g_i^{(t)}$ と $h_{ij}^{(t)} \log g_{j|i}^{(t)}$ とをとおしてだけ $\mathcal{Q}$ に影響しすることが分かります。従って M ステップは以下のような個別の最大化問題に帰結します。

$$\begin{aligned} \boldsymbol{\theta}_{ij}^{(p+1)} &= \arg\max_{\boldsymbol{\theta}_{ij}} \sum_t h_{ij}^{(t)} \log P_{ij}(\boldsymbol{y}^{(t)}) && (30) \\ \boldsymbol{v}_i^{(p+1)} &= \arg\max_{\boldsymbol{v}_i} \sum_t \sum_k h_k^{(t)} \log g_k^{(t)} && (31) \\ \boldsymbol{v}_{ij}^{(p+1)} &= \arg\max_{\boldsymbol{v}_{ij}} \sum_t \sum_k h_k^{(t)} \sum_l h_{l|k}^{(t)} \log g_{l|k}^{(t)} && (32) \end{aligned}$$

個々の最大化問題はそれ自身尤度最大化問題です。式 (30) は単に密度関数 P_{ij} の重みつき尤度最大化問題です。