EMアルゴリズムによるパラメータ推定に関する一考察*

2J－4

金田 亮一　鈴木 誠　平澤 茂一†

早稲田大学理工学部 経営システム工学科

1 はじめに

実世界から得られる複雑で不完全な情報を処理するために，確率統計的なモデル化が有効な手段として用いられている．ニューラルネットモデル，音声認識や画像処理において用いられるマルコフモデルなども，その代表的なものである．従来，ニューラルネットの学習ではバックプロパゲーションに代表される最急降下法が用いられてきた．しかしながら，一般に複雑な確率モデルに関する推論は，困難な非線形最適化問題に帰着されがちであり，学習に時間を要したり，局所最適解に陥りやすいという問題がある[1]，[2]．本稿では不完全データからの学習アルゴリズムであるEMアルゴリズムについての考察を行う．EMアルゴリズムは，最急降下法と同様，解を逐次改良していく繰返し探索のアルゴリズムであるが，その評価はKL情報量で与えているられている．すなわち，アルゴリズムを繰返し探索することにより，経験分布と推定分布との差がKL情報量の意味で最小に向かっている．これは尤度が最大に向かうのと同じことを意味するので，一般に最尤推定量を求めていることになる[1]．本稿では，尤度最大という評価基準に代わり，二乗誤差損失に対するEMアルゴリズムを与え，具体的な計算例を示す．

2 従来研究

2.1 不完全データの最尤推定

EMアルゴリズムでは，完全データ$x \in \mathcal{X}$というものが背後にあって，観測されるデータyはその完全データxの一部が欠落した不完全データであると考える．すなわち，$y = y(x)$なる，既知の多対1写像が存在しているとする．実際の応用では，ベクトルxが$x = (y, z)$のように分離できて，yだけが観測され，zが隠れた変数となっている場合が多い．完全データは実際に想定されるデータであることもあるし，推定を容易にするために仮想的に導入される場合もある．ここで，パラメータΨをもつ完全データ$x \in \mathcal{X}$の確率分布$f(x;\Psi)$を考える．$f(x;\Psi)$に対応してyの確率分布$g(y;\Psi)$が次のように与えられる．

$$g(y;\Psi) = \int_{X(y)} f(x;\Psi)dx \qquad (1)$$

ただし，$X(y) \subset \mathcal{X}$はyの逆写像集合である．観測データyが与えられた時に$g(y;\Psi)$をΨのもっともらしさを示す関数（尤度）として考え，その最大値をとるΨを求める方法が最尤推定法であり，データyと確率モデル$g(y;\Psi)$が与えられた時にΨを推論するために用いられる典型的な方法である．また，尤度gのかわりに対数尤度$\log g$を最大化しても同じであるので，対数尤度を用いることが多い．

2.2 EステップとMステップ

EMアルゴリズムはパラメータをある適当な初期値に設定し，Eステップ (Expectation step) とMステップ (Maximization step) と呼ばれる2つの手続きを繰り返すことによりΨの値を逐次更新する方法であり，次のように定式化される．

(1) パラメータの初期値を適当な点$\Psi = \Psi^{(0)}$にとる．

(2) $p = 0,1,2,\cdots$に対して次の2つのステップを繰り返す．

(2-a)**Eステップ** ： 完全データの対数尤度$\log f(x;\Psi)$の，データyとパラメータ$\Psi^{(p)}$のもとでの条件付平均を求める．すなわち，

$$\begin{aligned} Q^{(p)}(\Psi) &= E_{\Psi^{(p)}}[\log f(X^n;\Psi)|y^n] \\ &= \int_{X(y^n)} f(x^n;y^n,\Psi^{(p)}) \log f(X^n;\Psi)dx \end{aligned} \qquad (2)$$

を計算する．

(2-b)**Mステップ** ： $Q^{(p)}(\Psi)$を最大化するΨを$\Psi^{(p+1)}$とおく．なお，不完全データyが与えられたときの完全データxの条件付き分布は，Bayesの公式から

$$f(x^n;y^n,\Psi) = \begin{cases} f(x^n;\Psi)/g(y^n;\Psi) & x^n \in \mathcal{X}(y^n) \\ 0 & x^n \notin \mathcal{X}(y^n) \end{cases} \qquad (3)$$

で与えられる．このEステップとMステップの繰り返しによって尤度が単調に増加することが証明されている．[2]

したがって局所的には最適解に収束し，少なくとも初期解よりは，良い解が得られる．すなわち，Eステップにおいて完全データの対数尤度の条件付き期待値を求める変わりに完全データの損失の期待値をとってもEステップとMステップの繰り返しによって二乗誤差が単調に減少することが示せれば，局所的には最適解に収束し，少なくとも初期解よりは，良い解が得られることが分かる．以下においては損失関数が二乗誤差損失の場合についてEMアルゴリズムを拡張し，

*A note on parameter estimation by EM algorithm

†Ryouichi Kanada Makoto Suzuki Shigeichi Hirasawa Dep. of Industrial and Management Systems Engineering Waseda Univercity

具体例に対する計算により，その性質について考察を与える．

3 結果

3.1 正規混合モデル

例として，二つの正規分布 $N(\mu_1,1),N(\mu_2,1)$ の混合を考える．ここで，μ_1，μ_2 は平均であり，分散は1に固定する．この例において観測データ y は，$N(\mu_1,1),N(\mu_2,1)(k=1,2)$ のうちどちらかから選ばれるものとする．さらに k 番目の正規分布が選ばれる確率を，θ_k であるとする．このとき μ,θ を推定する問題を考える．

$$g(y;\theta,\mu)=\sum_{k=1}^{2}\theta_k\phi(y|\mu_k,1) \tag{4}$$

ここで，ϕ は正規分布の密度関数

$$\phi(y;\mu,1)=\frac{1}{\sqrt{2\pi}}\exp\{\frac{-(x-\mu)^2}{2}\} \tag{5}$$

である．次に，完全データを x とし，x が観測データ y を用いて $x=(y,z)$ と与えられたとする．ここで z は，$z(z=1,2)$ 番目の正規分布を表すもので，観測できない変数と考えるものとする．このとき完全データ x の分布は次のように書ける

$$f(x;\theta,\mu)=\theta_z\phi(y;\theta_z) \tag{6}$$

3.2 定式化

EMアルゴリズムと同じ手順を行うが，Eステップにおいて完全データの対数尤度の条件付き期待値を求める代わりに完全データの二乗誤差損失の期待値をとって推定を行うことにする．

(1) パラメータの初期値を適当な点 $\Psi=\Psi^{(0)}$ にとる．

(2) $p=0,1,2,\cdots$ に対して次の2つのステップを繰り返す．

(2-a)**Eステップ**：経験分布と推定分布の二乗誤差の，データ y^n とパラメータ $\Psi^{(p)}$ に関する条件付平均を求める．

$$\begin{aligned}Q^{(p)}(\Psi) &= E_{\Psi_{(p)}}[\{q(x_n)-f(x_n;\Psi)\}^2]\\ &= \int_{X(y_n)} f(x_n;\Psi^{(p)})\{q(x_n)-f(x_n;\Psi)\}^2(dx)\end{aligned}$$

を計算する．ただし，完全データ x の経験分布を以下のように定義する．

$$q(x)=\frac{1}{n}\sum_{k=1}^{n}\delta(x=x_k) \tag{8}$$

(2-b)**Mステップ**：$Q^{(p)}(\Psi)$ を最小化する Ψ を $\Psi^{(p+1)}$ とおく．

3.3 正規混合分布モデルの場合の計算結果

例で与えたモデルを提案したアルゴリズムで解くと以下のようになる．観測データ $y_1,y_2\ldots,y_n$ が与えられたとき，

$$Q(\theta^{(p)},\mu)=\sum_{i=1}^{n}\sum_{k=1}^{2}\frac{\theta_k^{(p)}\phi(y_i;\mu_k^{(p)},1)\{\theta_k q(y_i)-\theta_k\phi(y_i;\mu_k,1))}{g(y_i;\theta^{(p)},\mu^{(p)},1)} \tag{9}$$

この(9)式の左辺を最小にする θ,μ の関係を求めて

$$\theta^{(p+1)}=\frac{\sum_{i=1}^{n}c_2^i\{q(y_i);\phi(y_i;\mu_2,1)\}^2}{\sum_{i=1}^{n}[c_1^i\{q(y_i)-\phi(y_i;\mu_1,1)\}^2+c_2^i\{q(y_i)-\phi(y_i;\mu_2,1)\}^2]} \tag{10}$$

ただし，c_1^i,c_2^i は以下のようになる．

$$c_1^i=\theta_1^{(p)}\phi(y_i;\mu_1,1),\ c_2^i=(1-\theta_1^{(p)})\phi(y_i;\mu_2,1) \tag{11}$$

また，このときの μ_1，μ_2 の満たすべき条件は以下のようになる．

$$\sum_{i=1}^{n}c_1^i\{q(y_i)-\phi(y_i;\mu_1,1)\}^2\frac{\partial\phi(y_i;\mu_1,1)}{\partial\mu_1}=0 \tag{12}$$

$$\sum_{i=1}^{n}c_2^i\{q(y_i)-\phi(y_i;\mu_2,1)\}^2\frac{\partial\phi(y_i;\mu_2,1)}{\partial\mu_2}=0 \tag{13}$$

4 考察

EMアルゴリズムでは，K-L情報量の意味で経験分布と推定分布との差が小さくなるようにしている．本研究では，他の評価基準である二乗誤差損失に対する拡張アルゴリズムを示した．(10)式の形は，損失の比率を与えており，通常のアルゴリズムの解を拡張した形が得られている．(12)，(13)式は，各正規分布の重み付け平均された損失を最小にすることを示している．このことによりEMアルゴリズムとはいろいろな評価基準を与えることのできるアルゴリズムであり，これにより目的に合わせた推定量を求めることができる事が示唆される．

5 まとめ

今回は二乗誤差損失という評価基準についてアルゴリズムを示したが，今後は一般の損失関数について反復を繰り返すごとに損失が単調に小さくなること示すこと課題といえる．このことがいえれば，EMアルゴリズムの適応範囲がさらに広がるといえる．

参考文献

[1] 赤穂昭太郎，"EMアルゴリズムの幾何学"，情報処理，Vol. 37，No. 1 pp. 43-51，1996.

[2] Geoffrey J McLachlan, Thriyambakam Krishnan, *The EM Algorithm and Extensions*,