アテンション領域における EM アルゴリズム

赤穂 昭太郎 (PY)
電子技術総合研究所

EM Algorithm in an Attention Region

Shotaro Akaho (akaho@etl.go.jp)
Electrotechnical Laboratory

Abstract — We consider the learning in which data points are restricted in an attention region. The learning is treated as a maximum likelihood estimation with missing values. We present an EM algorithm for the learning in an attention region and show its effectiveness by some simple simulation results.

1. はじめに

学習や認識を効率良く行なうためには, 適当なアテンション (注視) をかけることが有効である. ところが, アテンションを行なうということは制約した領域の外のデータは捨ててしまうということにもなる.

本論文では, データが欠落した状況で学習や認識を行なうための手法として EM アルゴリズムを用いる. EM アルゴリズム [2] はボルツマンマシンの学習則とも深い関係にあり, またモジュール化されたニューラルネットの学習を高速に行なうことができるなど [3], 新しい学習則として注目されている [1].

2. 問題設定

ある決められたアテンションの領域を C とする. データ x はある未知の確率分布に従って発生されるとし, x は C の中に入った時のみ観測されるとする.

この観測にしたがってデータ $\underset{(1)}{x}, \ldots, \underset{(n)}{x}$ が得られたとして, x の従う確率分布を, $f(x \mid \xi)$ (ξ はパラメータ) で近似することが学習の目的である.

3. 学習アルゴリズム

上記の問題設定において欠測しているデータは, 観測されなかったデータを含めた総データ数および観測されなかったデータそのものの値である. ここではそれぞれを 2 つのステップに分けて扱うこととし, 以下のアルゴリズムで学習を行なう.

学習アルゴリズム

1. パラメータ ξ を適当な初期値 $\xi^{(0)}$ に固定する.
2. $p = 0, \ldots$ について以下の手続きを繰り返す.
   (a) ξ を $\xi^{(p)}$ に固定した時の欠測データ数 $m^{(p)}$ を推定する.
   (b) n 個の観測データと $m^{(p)}$ 個の欠測データに基づく EM アルゴリズムによって ξ の最尤推定値を求め, $\xi^{(p+1)}$ とする.

以下では具体的に上記の手続きの内容を示す.

4. 欠測データ数の推定

ξ を固定した時, 欠測されるかされないかは, C に属するか属さないかの 2 項分布であり, x が C に属する確率

$$P_C(\xi) \equiv \int_C f(x \mid \xi)\, \mathrm{d}x \tag{1}$$

を使うと, m の最尤推定値は

$$m_{\mathrm{mle}} = \frac{P_{\bar{C}}(\xi)}{P_C(\xi)} n \tag{2}$$

で求められる (ただし, 右辺は一般に整数ではないので丸める必要がある).

5. 指数分布族の混合モデルの場合

次に m を固定した時の EM アルゴリズムを指数型分布族の混合分布の場合に具体的に示す.

もともと欠測値のない混合モデルも EM アルゴリズムを適用して最尤推定量を高速に求めることができる. この場合はさらに欠測値を含んだ場合について考慮する必要がある. また特に, 指数型分布族の混合モデルの場合は簡単な形で書けるのでここでは指数型分布族の場合のアルゴリズムを述べる.

指数型分布族の混合モデルはパラメータ ξ として π, θ をもち,

$$f(x \mid \pi, \theta) = \sum_k \pi_k g_k(x \mid \theta_k)\,, \tag{3}$$

で表される. ただし, g_k は

$$g_k(x \mid \theta_k) = \exp[C_k(x) + \theta_k \cdot t_k(x) - \psi_k(\theta_k)] \tag{4}$$

という形の分布である.

Keywords— Attention, Learning, Adaptation, Statistical inference, EM algorithm

このとき, EM アルゴリズムの各ステップは

$$\eta_k^{(p+1)} = \frac{A_k^{(p)}[t_k]}{A_k^{(p)}[1]} \tag{5}$$

となる. ただし, η_k は期待値パラメータ $\mathrm{E}[t_k(x)]$ であり, $A_k^{(p)}$ は次で与えられる汎関数である.

$$\begin{aligned} A_k^{(p)}[\mathcal{X}] \equiv & \sum_{i=1}^{n} \{ \frac{g_k(\underset{(i)}{x} \mid \theta_k^{(p)})}{f(\underset{(i)}{x} \mid \pi^{(p)}, \theta^{(p)})} \cdot \mathcal{X}(\underset{(i)}{x}) \} \\ & + \frac{m}{P_{\bar{C}}(\pi^{(p)}, \theta^{(p)})} \int_{\bar{C}} g_k(x \mid \theta_k^{(p)}) \cdot \mathcal{X}(x)\, \mathrm{d}x. \end{aligned} \tag{6}$$

6. 正規混合モデルの場合

上記の計算を実際に行なう場合には $\int_C g_k(x \mid \theta_k)\, \mathrm{d}x$ および $\int_C g_k(x \mid \theta_k) t_k\, \mathrm{d}x$ を計算する必要があるが, 各 $g_k(x \mid \theta_k)$ が正規分布

$$h(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{7}$$

の場合は,

$$H(x) \equiv \int^x h(x \mid \mu, \sigma^2)\, \mathrm{d}x \tag{8}$$

を誤差関数などを用いて計算すれば,

$$\int^x x h(x \mid \mu, \sigma^2)\, \mathrm{d}x = \mu H(x) - \sigma^2 h(x) \tag{9}$$

$$\int^x x^2 h(x \mid \mu, \sigma^2)\, \mathrm{d}x = (\mu^2+\sigma^2) H(x) - (\mu+x)\sigma^2 h(x) \tag{10}$$

などによって陽に計算できる.

多次元の場合も, C が超直方体 (の和集合) で, 各次元が独立である正規分布モデルに対しては同じように計算できる.

7. 実験結果

図 1 に, 1 次元および 2 次元の正規混合モデルに対して行なったシミュレーションの結果を示す. 繰り返し数はいずれも 10 ステップ行なった.

1 次元は mode 数 2 でそれぞれの π, μ, σ の値は次の通りで $n = 1000$.

π	μ	σ
0.5	0	0.2
0.5	1	0.2

2 次元の場合は mode 数 3 で, π, μ, σ の値は次の通りで $n = 1000$.

π	μ	σ
0.5	(0,0.5)	(0.2,0.2)
0.25	(0,1)	(0.2,0.2)
0.25	(1,0)	(0.2,0.2)

いずれも高速に収束しているが, 欠測データを仮定しない場合は端の影響がでている.

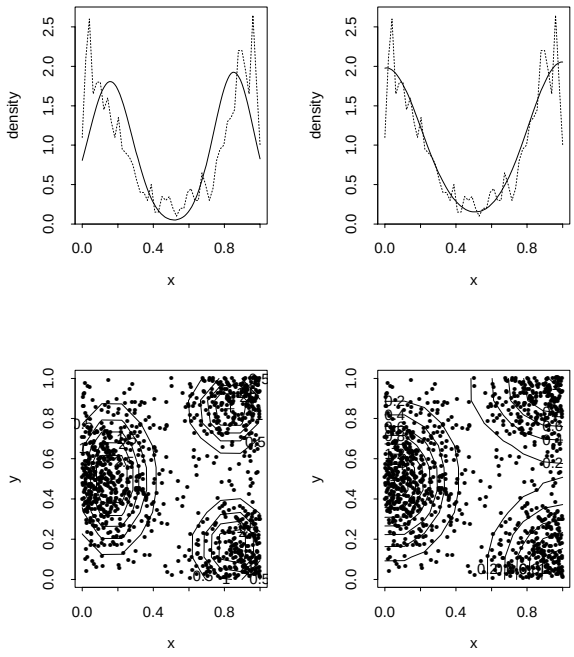

図 1. シミュレーション結果. 上段: 1 次元データ (実線: あてはめた密度, 点線: 度数分布). 下段: 2 次元データ ($*$: データ点, 実線: あてはめた密度の等高線, $+$ 各正規分布の中心). 左: 欠測データを仮定しない場合. 右: 提案アルゴリズム.

8. おわりに

アテンション領域に制限されたデータからの学習を, 欠測値を含む統計的学習ととらえて EM アルゴリズムを用いた計算法を示し, 簡単なシミュレーションによって有効性を示した.

今後の課題としては, より複雑な対象に対する認識や学習を, Jordan の modular network などを用いて同様な計算法を使って構成することが考えられる (その場合一般に最急降下法を援用することが必要となる). また, 今回は mode の数を固定したが, 通常は対象の個数などはあらかじめわからないのが普通である. それも自動的に行なうような機構を考える必要がある.

References

[1] Amari, S.: Information geometry of the EM and em algorithms for neural networks. (to appear).

[2] Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B, Vol. 39, pp. 1–38, 1977.

[3] Jordan, M. I. and Jacobs, R. A.: Hierarchical mixtures of experts and the EM algorithm. In *Proc. of IJCNN'93*, pp. 1339–1344, Nagoya, 1993.