EM アルゴリズム考察

Shin Saito
shinsa@jp.ibm.com

2013 年 4 月 4 日

概要

この文書では、**期待値最大化法**、通称 ***EM* 法**または ***EM* アルゴリズム** *(Expectation-Maximization Algorithm)* についての理解を述べる。

1 はじめに

文献 [1, 2] を参考にした。

2 パラメータ推定法

確率モデルにおけるパラメータの推定法について述べる。

2.1 最尤推定法

N 個の観測値の集まりであるサンプルデータ $D = \{x_1, \ldots, x_N\}$ が、**独立に、同一の確率分布に従って生成される (i.i.d.)** と仮定する。また、観測値としては d 次元ベクトルまたは d 個の変数の並びを仮定する。この確率分布が未知のパラメータを持つとした場合に、観測データから適切なパラメータを推定するのがパラメータ推定の問題である。

パラメータを θ とすると、確率関数は $p_\theta(X)$ や $p(X;\theta)$ と書かれることが多い。ここでは後者を採用する。すると独立性の仮定により、データ D の生起確率は

$$P(D) = \prod_{x_i \in D} p(x_i;\theta)$$

となり、これは θ の関数となる。これを**尤度** *(likelihood)* といい、$L(\theta)$ で表す。さらに、この尤度の対数を**対数尤度** *(log likelihood)* とよび、多くの場合 $l(\theta)$ で表す。

$$l(\theta) = \log P(D) = \sum_{x_i \in D} \log p(x_i;\theta)$$

最尤推定法 *(maximum likelihood estimation)* とは、対数尤度を最大化することによりパラメータを求める方法である。

2.2 EM 法

EM 法とは、パラメータ推定において、観測値の一部が得られない、つまり観測値が**不完全データ**である場合に、**非観測変数** *(unobserved variable)* または**隠れ変数** *(latent variable)* の値とともにパラメータを推定する方法である。基本的にはパラメータに仮の値を割り当てて最尤推定を行い、パラメータの値を更新していくと共に対数尤度を増加させる、という方法をとる。例を挙げる。

2.2.1 例: 混合モデル

隠れ変数は、内部状態を表す変数、と呼ばれることもある。そのネーミングによくマッチする例をあげる。

確率関数 $p_i, \ldots, p_M$ から得られる混合モデル

$$P(x;\boldsymbol{\lambda}) = \sum_{j=1}^{M} \lambda_j p_j(x) \quad (\text{ただし、} \sum_j \lambda_j = 1)$$

に対して、観測データ $x_1, \ldots, x_N$ からパラメータ $\boldsymbol{\lambda}$ $(= \lambda_1, \ldots, \lambda_M)$ を推定する問題を考える。この混合モデルはまず確率 λ_j で j を選び、p_j に従って x_i を生成する。観測データからはどの j が選ばれたかわからないため、単純な最尤推定法は使えない。x_i を生成する際にどの j が選ばれたかを q_i で表すと、これが隠れ変数となる。 この場合、データ $x_1, \ldots, x_N$ は不完全データであり、対応する完全データは $(x_1, q_1), \ldots, (x_N, q_N)$ となる。

3 クラスタリングと EM 法

そもそも EM 法は様々な分野で用いられていたアルゴリズムの一般化として 1977 年に提案された。ここで、EM 法の直感的な理解への助けとして、クラスタリングに用いられる **k-平均法** *(k-means)* を発展させることによって EM 法との関連を見ていくことにする。

3.1 クラスタリング

クラスタリングとはおおまかに言うと、インスタンス (観測値; observation ともいう) の集合 $D = \{x_1, \ldots, x_N\}$ を適当なクラスタ (部分集合) に分割することである。各インスタンスは d 次元ベクトルのようなものを想定する。クラスタリングの問題は、分割数があらかじめ指定されている場合とそうでない場合があるが、今回は指定されているものとする。さらに言うと、D を指定された数の集合に直和分割する問題を解くものとする。

3.2 k-means

k-means は、与えられたインスタンス集合 D を k 個のクラスタ $C_1, \ldots, C_k$ に直和分割するアルゴリズムであり、始めて公開されたのは 1965 年であると言われている。クラスタリングは各クラスタ C_j の重心 m_j がわかれば簡単に求まるがそれは未知である。k-means では m_j に適当な初期値を代入することから始め、各インスタンスの属するクラスタと、各クラスタの重心を交互に更新することによって最適なクラスタリングを求める。具体的なアルゴリズムは以下のようになる:

1. $\forall j.\ m_j$ に適当な初期値を代入し、以下の 2 と 3 を終了条件を満たすまで繰り返す:
2. (Assignment Step) $\forall i.\quad j^{(i)} = \operatorname*{argmin}_j d(x_i, m_j)$ (ただし $d(x,m)$ は d と m との距離) として、x_i をクラスタ $C_{j^{(i)}}$ に加える。つまり、重心が最も近いクラスタに加える。このとき、クラスタへの割り当てが前回のループと全く変わらなかった場合、終了とする。
3. (Update Step) $\forall j.\ m_j = \dfrac{\sum_{x_i \in C_j} x_i}{|C_j|}$

(参考にした文献には明記されていないが) このアルゴリズムはいずれ終了し、ある種の意味で局所最適なクラスタリングを与える。局所最適であるので、初期値によって結果が異なる可能性がある。

あとで見るように、これは (ちょっと特殊な) 確率モデルの元での局所最尤推定になっていることがわかる。

3.3 確率モデルとしてのクラスタリング

前述のモデルでは各インスタンスは最も重心が近いクラスタに属するとした。これをより確率的な考え方にマッチするよう、モデルとその見方を変更する。

確率モデルとして捉えた場合、各インスタンスは、まずどのクラスタから生成されるかが (確率的に) 決められた後、そのクラスタから確率的に生成される、と考える。つまり、インスタンス x_i がクラスタ C_j によって生成されるという事象の確率は

$$p(x_i, j) = p(j)p(x_i|j)$$

となる。ここでの $p(x_i|j)$ は混合モデルに現れた $p_j(x_i)$ と同じものの表記であると思ってよい。さらにこれを全てのクラスタにわたって加えることにより、x_i の (周辺確率としての) 生成確率が求まる。

$$p(x_i) = \sum_j p(j)p(x_i|j)$$

つまりこれは、不完全データ $\{x_i\}$ に対する完全データ (とパラメータ) $\{(x_i, j)\}$ の推定問題であることがわかる。

3.4 混合正規分布によるクラスタリング

混合正規分布 (Gaussian mixture) では、クラスタからインスタンスを生成する確率分布として正規分布を仮定する。つまり、クラスタの重心 m_j およびインスタンスの偏差 σ をパラメータとする正規分布 $N(m_j, \sigma)$ に従うとする。すると、インスタンスを d 次元ベクトルであるとして、

$$p(x_i|j; \boldsymbol{m}) = p(x_i|j; m_j) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{|x_i - m_j|^2}{2\sigma^2}\right)$$

が成り立つ[*1]。ここでは単純化のために、σ を定数とし、クラスタによって変わらないとする[*2]。

前節のモデルをこの混合正規分布モデルを用いて拡張すると、k-means におけるループでの処理は以下のようになる。

1. $\forall j.\ m_j$ に適当な初期値を代入し、以下の 2 と 3 を終了条件を満たすまで繰り返す:
2. (Assignment Step) $\forall i, j.\ x_i$ がクラスタ C_j に属する事後確率 $p(j|x_i)$ を求める。前節のモデルではインスタンスはいずれか 1 つのクラスタに属したが、ここではそれぞれのクラスタに属している確からしさを求めることになる。

$$p(j|x_i; \boldsymbol{m}) = \frac{p(x_i, j; \boldsymbol{m})}{p(x_i; \boldsymbol{m})} = \frac{p(x_i, j; \boldsymbol{m})}{\sum_j p(x_i, j; \boldsymbol{m})} = \frac{p(j)p(x_i|j; \boldsymbol{m})}{\sum_j p(j)p(x_i|j; \boldsymbol{m})}$$

   ここでクラスタの生成確率が一様であるとすると、$p(j)$ を省くことができ、

$$p(j|x_i) = \frac{p(x_i|j; \boldsymbol{m})}{\sum_j p(x_i|j; \boldsymbol{m})}$$

   と計算できる形になる。なお、ループの終了条件については後で述べる。
3. (Update Step) $\forall j.$ ここでは前ステップで求めた事後確率を用いて、重み付きでクラスタの重心を計算し値を更新する。

$$m_j = \frac{\sum_{x_i \in D} p(j|x_i; \boldsymbol{m}) x_i}{\sum_{x_i \in D} p(j|x_i; \boldsymbol{m})}$$

これをもとに考えると、前述の k-means は混合正規分布におけるアルゴリズムを、ある確率分布に適用したものであることがわかる。つまり、k-means は前述のアルゴリズムで、$p(j|x_i)$ を、x_i が $\boldsymbol{m}$ の中で m_j に最も近いとき 1 となり、その他の場合に 0 となるように取ったものである。

また、これを一般化することにより、EM 法のアルゴリズムが得られる。ここで述べたアルゴリズムの終了条件は EM 法の終了条件を流用することができる。

3.5 EM 法

ここでクラスタリングから一度離れて、パラメータ推定の一般的な問題を考える。不完全な観測データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ を確率モデル $p(\boldsymbol{x}; \boldsymbol{\theta})$ から生成された観測データであるとし、 対応する完全データは $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)$ であるとする。このときパラメータ $\boldsymbol{\theta}$ を推定する問題を考える。

EM 法は繰り返しにより、現在のパラメータ $\boldsymbol{\theta}$ から対数尤度 $l(\boldsymbol{\theta})$ を増加させるような新しいパラメータ $\bar{\boldsymbol{\theta}}$ を求める方法である。ここで、パラメータを更新した際の対数尤度の差を求めてみる。すると、

$$\begin{aligned} l(\bar{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}) &= \sum_{\boldsymbol{x}_i} \log p(\boldsymbol{x}_i; \bar{\boldsymbol{\theta}}) - \sum_{\boldsymbol{x}_i} \log p(\boldsymbol{x}_i; \boldsymbol{\theta}) \\ &= \text{(中略)} \\ &\geq Q(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) - Q(\boldsymbol{\theta}, \boldsymbol{\theta}) \end{aligned}$$

が成り立つ。ただし、

$$Q(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{\boldsymbol{x}_i} \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}_i; \boldsymbol{\theta}) \log p(\boldsymbol{x}_i, \boldsymbol{y}; \bar{\boldsymbol{\theta}})$$

である。この関数を Q 関数とよぶ。Q 関数を最大化するような $\bar{\boldsymbol{\theta}}$ を求めることにより、対数尤度が増加していくことがわかる。EM 法の終了条件は、Q 関数または対数尤度の増加が十分小さくなったとき、とすればよい。なお、EM 法で求まるのは局所最適解であることに注意する。

なお、実際の計算の際には同時確率を展開して

$$\begin{aligned} Q(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{x}_i} \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}_i; \boldsymbol{\theta}) \log \left\{ p(\boldsymbol{y}; \bar{\boldsymbol{\theta}}) p(\boldsymbol{x}_i|\boldsymbol{y}; \bar{\boldsymbol{\theta}}) \right\} \\ &= \sum_{\boldsymbol{x}_i} \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}_i; \boldsymbol{\theta}) (\log p(\boldsymbol{y}; \bar{\boldsymbol{\theta}}) + \log p(\boldsymbol{x}_i|\boldsymbol{y}; \bar{\boldsymbol{\theta}})) \end{aligned}$$

とすることが多い。これは確率モデルの表現方法によって異なってくる。

3.5.1 混合正規分布モデルとの関係

前述の Q 関数の式を、混合正規分布モデルでのアルゴリズムと比べてみると、$p(\boldsymbol{y}|\boldsymbol{x}_i; \boldsymbol{\theta})$ の部分が、インスタンスが各クラスタに属する事後確率を求めている部分に相当することがわかる。また、Q 関数は $\log p(\boldsymbol{x}_i; \bar{\boldsymbol{\theta}})$ に関する期待値 (つまりデータの対数尤度) を求めていることがわかる。これを最大化することにより、同じくある種の期待値である重心の更新式が得られるのは興味深いことである。

[*1] $p(x_i|j; \boldsymbol{m})$ とは $\boldsymbol{m}$ をパラメータとする条件付き確率 $p_{\boldsymbol{m}}(x_i|j)$ のことである。同様に、$p(x_i, j; \boldsymbol{m})$ とは $p_{\boldsymbol{m}}(x_i, j)$ のことである。

[*2] これを、クラスタごとに異なる、定数でないパラメータ σ_j である、としても EM 法で解くことができる。

3.5.2 (予想) 不完全データの一部がわかっている場合

不完全データの一部がわかっている場合がある。例えば隠れ変数 y_1, y_2, y_3 のうち、いずれか 1 つだけは常にわかっている場合、などである。参考文献には載っていなかったが、この場合、ある組み合せに対しては $\sum_{\boldsymbol{y}} \cdots$ が確実に計算でき、その場合は推定の精度がより高くなるか、繰り返しの回数が減少することが予想される。

参考文献

[1] 高村大也. 言語処理のための機械学習入門. 自然言語処理シリーズ 1. コロナ社, 2010.

[2] 北研二. 確率的言語モデル. 言語と計算 4. 東京大学出版会, 1999.