東京大学 工学部 計数工学科 応用音響学

D4 - 混合正規分布と EM アルゴリズム

嵯峨山 茂樹 <sagayama@hil.t.u-tokyo.ac.jp> 他

東京大学 工学部 計数工学科

資料所在 http://hil.t.u-tokyo.ac.jp/~sagayama/applied_acoustics/

謝辞：システム情報第一研究室勉強会資料を利用 (山本 担当分)

- 混合正規分布
- 不完全データ問題
- EMアルゴリズム

参考文献

References

[1] 渡辺美智子，山口和範 編著，“EM アルゴリズムと不完全データの諸問題,” 多賀出版，2000. ISBN4-8115-5701-8 C1033 ￥6600E.

[2] 北研二, “確率的言語モデル,” 東京大学出版会, 1999. ISBN4-13-065404-7 C3304 ￥3800E.

[3] 北研二, 中村哲, 永田昌明, “音声言語処理,” 森北出版, 1996.

[4] L. Rabiner, B. H. Juang, “音声認識の基礎,” 1995.

[5] 上田修功，“ベイズ学習,” 電子情報通信学会誌，Vol. 85，No. 4，pp. 265-271, 2002.

[6] 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.

[7] 甘利俊一，“情報幾何学,” 応用数理，Vol. 2，No. 1，pp. 37-56, 1992.

正規分布

確率変数: x
平均: μ
分散: σ^2
1次元正規分布:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

対数表現:

$$L(\mu,\sigma^2;x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma - \underbrace{\frac{(x-\mu)^2}{2\sigma}}_{\textit{Mahalanobis}\text{ 距離}}$$

n 次元正規分布

確率変数ベクトル: $\boldsymbol{x}$
平均ベクトル: $\boldsymbol{\mu}$
分散行列: Σ
n **次元正規分布:**

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp -(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

対数表現:

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}_{\textit{Mahalanobis}\text{ 距離}}$$

混合正規分布

M 混合 n 次元正規分布:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{m=1}^{M} \lambda_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \\
&= \sum_{m=1}^{M} \lambda_m \frac{1}{(2\pi)^{n/2}|\Sigma_m|^{1/2}} \exp\left\{-(\boldsymbol{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_m)\right\} \\
&\quad \left(\sum_{m=1}^{M} \lambda_m = 1\right)
\end{aligned}
$$

混合正規分布とEMアルゴリズム 1/6: 不完全データ

例： 成人男女の身長のデータがある。これをそれぞれ単一の正規分布でモデル化しよう。全ての身長データについて男女どちらのデータかが分かっていれば、男女構成割合および男女それぞれの平均と分散を求めることは容易である。しかし、男女の区別が分からない不完全なデータが含まれている場合はどうすればよいか。このような問題を不完全データ問題と呼ぶ。

最も簡単な対処法は、不完全なデータは無効としてノーカウントにする。つまり、捨ててしまうわけだ。何も考えずにそうしている人は実際多い。不完全データが少なければこれでも良いが、そうでないときは困ってしまう。測定にコストがかかっている場合は、やはり不完全だと言うだけでせっかくの情報を捨てるのはもったいない。極端な場合、すべてのデータについて男女が分からないようなときは、この方法は取れない。

混合正規分布とEMアルゴリズム 2/6: 完全化

もし問題になっている不完全データの値が、男女全体の分布の中でかなり小さい値であれば、これは女データである確率が高い。逆に、もしかなり大きい値であれば、男データの確率が高い。

しかし、男女のいずれとも明らかでない、全体の平均あたりの値であればどうか。確かにそのデータは男女のいずれかなのだが、ここで無理矢理、男女を決定する方針は取らずに、その不完全データが男に属する確率と女に属する確率を考えることにしよう。つまり、不完全データは、例えば男0.5人、女0.5人のデータとしてカウントして、男女どちらの分布の推定にも使うのである。もちろん、不完全データ値が、男女の中間とは限らないから、男0.7人、女0.3 人のデータとしてカウントする方がよい場合もあるだろう。

混合正規分布とEMアルゴリズム 3/6: 事後確率

では、その比率はどのように求めればよいか。これは、もし男女それぞれの分布が分かっていれば、それらに男女の構成割合を掛けてやった男の分布からそのデータ値が出る確率 (あるいは確率密度値) が分かるし、女の分布からそのデータ値が出る確率 (あるいは確率密度値) も分かる。

逆にデータ値を与えて、それが男である確率と女である確率 (両者の和は 1) は、上の確率値を和が 1 になるように正規化した値、すなわち事後確率によって求められる。これで、不完全データ値があれば、それを男 α 人、女 $1-\alpha$ 人としてカウントすることができる。

しかし、問題は、このためには、男女の分布が分かっていなければならないということである。男女の分布を求めるために不完全データを使おうとしているのに、これでは鶏と卵のようなものだ。

混合正規分布とEMアルゴリズム 4/6: 反復計算

では、こうしよう。先に仮の男女の平均と分散を決めて分布を仮定しておこう。いろいろな仮定の仕方がありそうだが、あくまで仮の分布である。そして、これによって不完全データを男女に1以下の小数のカウントを決める。不完全データが複数ならば、それぞれについて行う。こうすると、たとえば男61.3人、女59.7人のデータ、などとなるかもしれない。こうして、データに人数重みづけをして平均と分散を求める。これで、男女の分布が最初の仮定から更新された。分布が更新されたならば、不完全データの小数カウントも変動するだろう。だから計算し直す。そうすると分布も求め直す。つまり、これを繰り返すのだ。

この手順をまとめると、まずなんらかの方法で、男女それぞれに初期分布が求められている場合、

(1) 男女それぞれの分布を用いて、不完全データそれぞれについて、男女所属確率(事後確率)を求める。

(2) 不完全データをその小数カウントで加えた男女それぞれのデータから、男女それぞれの分布を求め直す。

を繰り返すことになる。

混合正規分布とEMアルゴリズム 5/6: DLR論文

初期分布を求めるには、具体的には、男女別の分からないデータ(不完全データ)が少量なら除外するか、あるいは不完全データはずべて男女各0.5人とカウントするか、さまざまな手法があり得るだろう。

以上の方法は極めて素朴な直観に基づく方法で、なんとなく妥当な感じもするが、はたして収束するのだろうか。また、収束するとして推定した分布は真の分布に近付くのだろうか、あるいはもっと正確に言えば、どういう性質の分布推定になっているのだろうか。

これを理論づけたのが、有名な Dempster-Laird-Rubin の論文(1977)である。この中で彼らは上のような推定は尤度最大の分布推定(最尤推定)になることを証明した。

混合正規分布とEMアルゴリズム 6/6: データ欠測値問題

上の手順は、さらに一般化して考えることができる。→ E-step M-step

以上で問題とした混合正規分布以外にも、不完全データ問題は多い。データ欠測値問題と呼ばれている不完全データ問題の一つの例を挙げよう。電球の寿命を調べてその分布を求めるとしよう。全部の電球が切れるまで待っているわけに行かないから、試験期間に寿命が測定できない電球が残る。だからといって、切れなかった電球を除外して寿命の分布を求めたのでは、全く誤った結果になってしまう。これもやはり不完全データ問題として、切れなかった電球も含めて全体の分布を推定するべきである。

確率モデルの統計的学習

- 確率モデル
  - 現象をモデル化すると扱いやすい
  - 確率モデル：統計的に情報源をモデル化
    - モデルの構造決定
    - モデルの学習
    - 認識アルゴリズム
    - etc.
- モデルの学習
  - 学習：過去の経験を未来に活かすための手段
  - モデル：観測データ（標本）の発生メカニズム
- 統計的学習：標本からモデルのパラメタを推定する問題
  - 手持ちのデータに対する整合性（過去）
  - 未学習データに対する予測（未来）

統計的学習の分類（1）

学習データの形態

- **教師なし学習**（unsupervised）
  - **学習：入力xの背後の確率分布推定**
  - **学習データ：入力x**
  - **例）混合正規分布推定**
- **教師あり学習**（supervised）
  - **学習：入出力写像$f : x \rightarrow y$の学習**
  - **学習データ：入力xとそれに対する理想出力y（教師信号）**
  - **例）関数近似（回帰）**

統計的学習の分類（2）

パラメタ推定法

- 最尤学習
  - 未知パラメタは確定的変数→点推定
  - 観測データとパラメタの尤度関数を最大化するパラメタを求める
  - EMアルゴリズムもこの一種
- ベイズ学習
  - 未知パラメタは確率変数→分布推定
  - ベイズの定理により，事前分布と観測データから事後分布を得る
  - 事後分布に基づく予測分布の算出が可能
- etc.

最尤学習

- 最尤推定（Maximum Likelihood）
  - 尤度関数：観測データ D の同時分布の対数

$$L(\theta; D) = \log p(D; \Theta) = \log \prod_n p(d_n; \Theta)$$

  - 最尤推定値：観測データ D の対数尤度関数を最大化する値

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, L(\theta; D)$$

  - 真の分布と推定分布のKL距離の最小化と等価

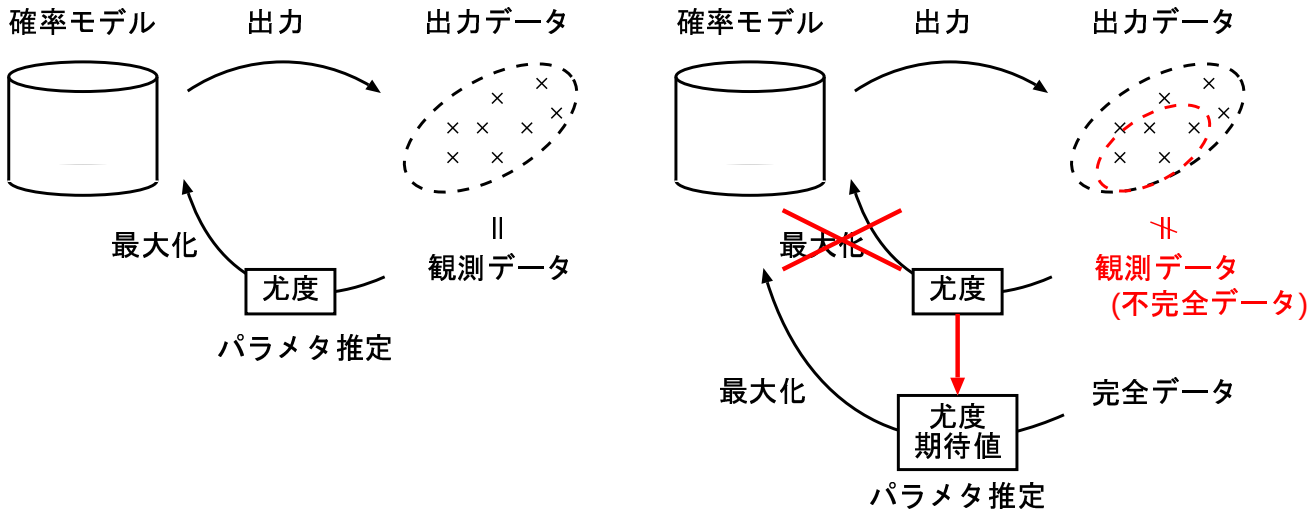
$$\theta_{ML} = \underset{\theta}{\operatorname{argmin}}\, \mathrm{KL}(p(D; \theta_0), p(D; \theta))$$

$$\mathrm{KL}(p(D; \theta_0), p(D; \theta)) = \int p(D; \theta_0) \log \frac{p(D; \theta_0)}{p(D; \theta)}\, dD$$

  - 解析解が求められない場合が多い

EMアルゴリズム

- EMアルゴリズム（Expectation Maximization）
  - 最尤法（解析的）よりも汎用な数値解法
  - 逐次的に局所最適値に近づく
  - 不完全データからの最尤推定値を求める理論的枠組み
  - 観測できない変数（隠れ変数，潜在変数）を含めてもよい
  - 尤度関数の最大化の代わりに完全データ尤度関数の**期待値を最大化**



例）HMMでは$x = (y, q)$が完全データ

EMアルゴリズムの数式

観測データ D のときの尤度関数（$\theta \in \Theta$，Z：非観測データ）

$$\begin{aligned} L(\theta; D) &= \log p(D \mid \theta) \\ &= \log p(D, Z \mid \theta) - \log p(Z \mid D, \theta) \end{aligned} \tag{1}$$

ある θ^t と D を条件とした，Z についての尤度関数の期待値

現在の状態

$$\begin{aligned} E_{\theta^t}\left[\log p(D \mid \theta)\right]_{Z|D} &= \sum_Z p(Z \mid D, \theta^t) \log p(D \mid \theta) \\ &= \log p(D \mid \theta) \end{aligned} \tag{2}$$

(1)式と(2)式より

$$\begin{aligned} L(\theta; D) &= E_{\theta^t}\left[\log p(D, Z \mid \theta)\right]_{Z|D} - E_{\theta^t}\left[\log p(Z \mid D, \theta)\right]_{Z|D} \\ &= Q(\theta \mid \theta^t) - H(\theta \mid \theta^t) \end{aligned} \tag{3}$$

次にくる θ が θ^t より最尤であるためには

$$\underbrace{L(\theta; D) - L(\theta^t; D)}_{\text{尤度関数を大きくするには}} = \{\ \underbrace{Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)}_{\text{完全データの期待値を大きくする}}\ \} - \{\ \underbrace{H(\theta \mid \theta^t) - H(\theta^t \mid \theta^t)}_{\text{ここは Jensen の不等式により負}}\ \} \geq 0$$

EMアルゴリズム

1. 初期値θ^0を設定，$0 \rightarrow t$
2. 以下を収束するまで繰りかえす
    1. $Q(\theta \,|\, \theta^t)$を計算
    2. $\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta \,|\, \theta^t)$とし，$t+1 \rightarrow t$

Q関数

■ Q関数

$$Q(\theta \mid \theta^t) = E_{\theta^t}\left[\log p(D, Z \mid \theta)\right]_{Z|D} = \sum_Z p(Z \mid D, \theta^t) \log p(D, Z \mid \theta)$$

完全データの尤度の Z に関する期待値

＝ 完全データの尤度関数に隠れ変数 Z の可能な値を全て代入

→ 観測データの完全化 (Z は埋めたので D のみ)

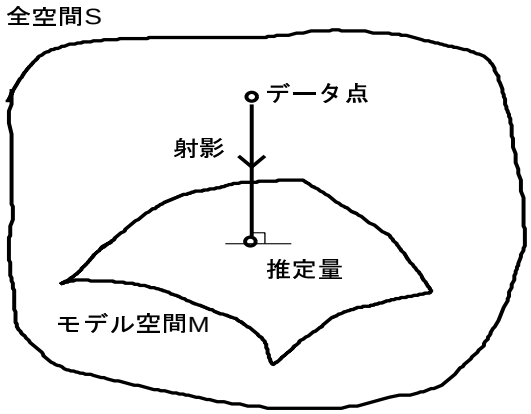
ただし，Z には現在の状態に基づく確率的情報をつける

■ Jensenの不等式

$$\begin{aligned}
&H(\theta^t \mid \theta^t) - H(\theta \mid \theta^t) \\
&\quad = E_{\theta^t}\left[\log p(Z \mid D, \theta^t)\right]_{Z|D} - E_{\theta^t}\left[\log p(Z \mid D, \theta)\right]_{Z|D} \\
&\quad = \sum_Z p(Z \mid D, \theta^t) \log p(Z \mid D, \theta^t) - \sum_Z p(Z \mid D, \theta^t) \log p(Z \mid D, \theta) \\
&\quad = \sum_Z p(Z \mid D, \theta^t) \log \frac{p(Z \mid D, \theta^t)}{p(Z \mid D, \theta)} \\
&\quad \geq \left(\sum_Z p(Z \mid D, \theta^t)\right) \log \frac{\Sigma_Z\, p(Z \mid D, \theta^t)}{\Sigma_Z\, p(Z \mid D, \theta)} = 0
\end{aligned}$$

幾何学的な見方 [6, 7]

- 確率分布の空間
  - $S = \{p(x \mid \theta)\}$：パラメタθを持つ確率分布の集合
  - Sはθを局所座標系とする空間（多様体）と見なせる
- 統計的な推定の幾何学的イメージ
  - モデルの空間Mを考える（Sの部分空間）
  - データはまた別の座標系の１点
  - 推定：データ点からモデル空間Mへの射影



- EMアルゴリズムの幾何学的イメージ
  - D：データ多様体（Sの部分空間）
  - データが不完全な分自由度を持つ
  - EM：２つの空間間の射影のくり返し