概要

前節ではEP法の手続きを学んだ。EP法は変分法と異なり forward のKLダイバージェンスを最小化する手法であり、分解した因子を一つずつ更新していくことで求めたい近似分布を得る。残りの節はこれらを実際の問題に適用していく例を見る。

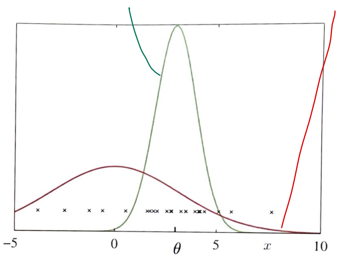
10.7.1 節では比較的簡単な例として、雑音を含む混合ガウス分布のデータを等方ガウス分布で近似する例をみる。前節で学んだEP法の手続きとの対応を明らかにしながら理解を深めることを目標とする。

10.7.2 節ではより一般的な確率分布に対してEP法を適用するために、EP法をグラフィカルモデル上で実行する方法を学ぶ。8章では変数の周辺分布をメッセージパッシングで求めていたが、EP法では更新後の因子をメッセージパッシングにより求める。

10.7.1 例：雑音データ問題

ここでは雑音データに対してEP法を適用する例を確かめる。観測データは次のような混合ガウス分布を仮定する。

$$p(x|\theta) = \underbrace{(1-w)\mathcal{N}(x|\theta, I)}_{\text{推定したい分布}} + \underbrace{w\mathcal{N}(x|0, \alpha I)}_{\text{背景雑音}}$$



その他の細かい設定は教科書を参照とし、式のみ下記に記す。

事前分布： $p(\theta) = \mathcal{N}(\theta|0, bI)$

同時分布： $p(D, \theta) = p(\theta)\prod_{n=1}^{N} p(x_n|\theta)$

$$\left(= f_0(\theta)\prod_{n=1}^{N} f_n(\theta) \right)$$

近似分布： $q(\theta) = \mathcal{N}(\theta|m, vI)$

$$\left(\begin{aligned} &= \tilde{f}_0(\theta)\prod_{n=1}^{N} \tilde{f}_n(\theta) \\ &= p(\theta)\prod_{n=1}^{N} s_n \mathcal{N}(\theta|m_n, v_n I) \quad \because (*) \end{aligned} \right)$$

(*) 部分について

等方ガウス分布は指数二次関数の総積の形に分解できる。 $\mathcal{N}(\theta|\mathbf{m}_n, v_n \mathbf{I})$ は厳密にガウス分布を示しているのではなく、指数二次関数を表すための都合の良い略記法である。 $\tilde{f}_0(\theta)$ が $p(\theta)$ と等しくなる理由はちゃんとわかっていないが、おそらくどの初期値を選んでもEP法の一回の更新で $\tilde{f}_0(\theta) = p(\theta)$ となるからだと考えている。（演習解答10.37）

ここからEP法の手続きを実施していく。p.223 の手続きを参照しながら、対応する操作を確認しながら進めていくこととする。

1. 近似因子 $\tilde{f}_n(\theta)$ をすべて初期化する。

近似 $\tilde{f}_n(\theta)$ は、n = 1,...,N については1に初期化される。$\tilde{f}_0(\theta)$ については $p(\theta)$ とする。従って、初期の近似事後分布は、 $q(\theta) = p(\theta)$ となり事前分布と一致する。

2. 事後分布の近似を $q(\theta) = \prod_n \tilde{f}_n(\theta)$ とする。

既に、 $q(\theta) = \tilde{f}_0(\theta) \prod_{n=1}^{N} \tilde{f}_n(\theta)$ としている。

3. 近似が収束するまで、以下を繰り返す

(a) 改良したい因子 $\tilde{f}_n(\theta)$ を選ぶ

ここで、$\tilde{f}_0(\theta)$ は更新の対象とならず、n = 1,...,N についての $\tilde{f}_n(\theta)$ のみ更新対象となる。なぜなら、 $\tilde{f}_0(\theta)$ はEP法の更新で変化がないためである。これについては後で演習解答10.37を参照する。

(b) $\tilde{f}_n(\theta)$ を事後分布から除算して取り除く

$$
\begin{aligned}
q^{\backslash n}(\theta) &= q(\theta) / \tilde{f}_n(\theta) \\
&= \underbrace{\tilde{f}_0(\theta)}_{\text{ガウス分布}} \prod_{i \neq n} \underbrace{\tilde{f}_i(\theta)}_{\text{ガウス分布}}
\end{aligned}
$$

これを指数部分の中身について平方完成すると、平均と分散が次のように求まる（演習10.38←やらない）

$$\text{平均:} \quad \mathbf{m}^{\backslash n} = \mathbf{m} + v^{\backslash n} v_n^{-1} (\mathbf{m} - \mathbf{m}_n) \tag{10.214}$$

$$\text{分散:} \quad (v^{\backslash n})^{-1} = v^{-1} - v_n^{-1} \tag{10.215}$$

（注意）

$q^{\backslash n}(\theta)$ は正規化されていないガウス分布として得られる。

(c) 新しい事後分布 $q^{new}(\theta)$ を、そのモーメントを $q^{\backslash n}(\theta) f_n(\theta)$ と一致させることで求め、正規化定数 Z_n を計算する。

$$q^{\backslash n}(\theta) f_n(\theta) = \underbrace{q^{\backslash n}(\theta)}_{\text{すでにモーメントが求まっている。}} \cdot \left\{ (1-w)\mathcal{N}(x_n \mid \theta, I) + w\mathcal{N}(x_n \mid 0, \alpha I) \right\}$$

これの平均と分散は教科書の式 (10.217), (10.218), (10.219) のように求まる。これらモーメントを $q^{new}(\theta)$ と一致させることで $q^{new}(\theta)$ が求まる。正規化定数 Z_n は次のように求まる。

$$\begin{aligned}
Z_n &= \int q^{\backslash n}(\theta) f_n(\theta)\, d\theta \\
&= (1-w)\int q^{\backslash n}(\theta)\mathcal{N}(x_n \mid \theta, I)\, d\theta + w\int q^{\backslash n}(\theta)\, d\theta\ \mathcal{N}(x_n \mid 0, \alpha I) \\
&= (1-w)\int \underbrace{\mathcal{N}(\theta \mid m^{\backslash n}, v^{\backslash n} I)}_{\text{正規化されたものとして良いのか不安}} \mathcal{N}(x_n \mid \theta, I)\, d\theta + w \cdot \underbrace{1}_{\text{ここも同じく不安}} \cdot \mathcal{N}(x_n \mid 0, \alpha I) \\
&= (1-w)\mathcal{N}(x_n \mid m^{\backslash n}, (v^{\backslash n}+1) I) + w\mathcal{N}(x_n \mid 0, \alpha I) \qquad (\because 2.115)
\end{aligned}$$

(d) 新しい因子を、 $\tilde{f}_n(\theta) = Z_n \dfrac{q^{new}(\theta)}{q^{\backslash n}(\theta)}$ より求める。

$$\tilde{f}_n(\theta) = Z_n \frac{q^{new}(\theta)}{q^{\backslash n}(\theta)}$$

両方とも指数二次関数の形なので、平方完成することができる。

$$\begin{aligned}
&= \langle \text{計算略} \rangle \\
&= s_n \mathcal{N}(\theta \mid m_n, v_n I)
\end{aligned}$$

ここで、

$$\begin{cases}
v_n^{-1} = (v^{new})^{-1} - (v^{\backslash n})^{-1} \\
m_n = m^{\backslash n} + (v_n + v^{\backslash n})(v^{\backslash n})^{-1}(m^{new} - m^{\backslash n}) \\
s_n = \dfrac{Z_n}{(2\pi v_n)^{D/2}\mathcal{N}(m_n \mid m^{\backslash n}, (v_n + v^{\backslash n}) I)}
\end{cases}$$

この更新をパラメータが収束する十分な回数実行すれば良い。

4. モデルエビデンスの近似を $p(\mathcal{D}) \simeq \int \tilde{f}_0(\boldsymbol{\theta}) \prod_{n=1}^{N} \tilde{f}_n(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ より求める。

$$
\begin{aligned}
p(\mathcal{D}) &\simeq \int \tilde{f}_0(\boldsymbol{\theta}) \prod_{n=1}^{N} \tilde{f}_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \int \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{0}, b\mathbf{I}) \cdot \prod_{n=1}^{N} s_n \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_n, v_n \mathbf{I})\, d\boldsymbol{\theta}
\end{aligned}
$$

$=$ (これを計算すれば良いと思ったが、答えに $\mathbf{m}^{new}$, v^{new} が表われるので、方針が違うかも。)

$$
= (2\pi v^{new})^{D/2} \exp\left(\frac{B}{2}\right) \prod_{n=1}^{N} \left\{ s_n (2\pi v_n)^{-\frac{D}{2}} \right\}
$$

ここで $B = \dfrac{(\mathbf{m}^{new})^{T} \mathbf{m}^{new}}{v} - \displaystyle\sum_{n=1}^{N} \dfrac{\mathbf{m}_n^{T} \mathbf{m}_n}{v_n}$

以上の手続きにより、事後分布 $q^{new}(\boldsymbol{\theta})$ とモデルエビデンスを求められた。

図10.16 と図10.17の解釈は当日読み合わせとする。演習 10.37 も確かめる。

10.7.2 グラフィカルモデルとEP法

前節までは各因子は全てのパラメータ θ に依存することを考えてきた。ここでは、各因子が一部のパラメータのみに依存するような、より一般的な場合にEP法を適用する例を学ぶ。このために8章で学んだグラフィカルモデルを導入する。前半は簡単のため4変数のみのグラフィカルモデルについてEP法を適用し、後半はより一般的なグラフに対して適用することを考える。

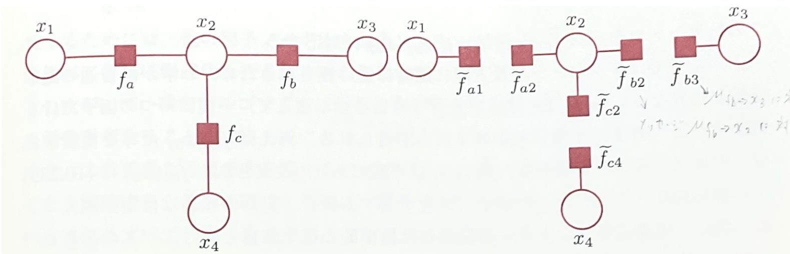
この節において強調すべき点は、全ての変数を独立とする近似分布において、EP法の手続きが積和アルゴリズムと同様のメッセージパッシングを行うことで達成できる点である。すなわち、隣接因子から送られるメッセージを集約することで近似因子を更新することができ、特に木構造においては2N回のメッセーパッシングで近似分布が得られる。

細かい前提は教科書を参照とし、式のみ下記に記す。

同時分布 : $p(x) = f_a(x_1, x_2)\, f_b(x_2, x_3)\, f_c(x_2, x_4)$

近似分布 : $q(x) \propto \tilde{f}_{a1}(x_1)\, \tilde{f}_{a2}(x_2)\, \tilde{f}_{b2}(x_2)\, \tilde{f}_{b3}(x_3)\, \tilde{f}_{c2}(x_2)\, \tilde{f}_{c4}(x_4)$

=ではない？　全ての変数の独立を仮定している。



順にEP法の手続きを踏んでいく。$f_b(x_2, x_3) = \tilde{f}_{b2}(x_2)\, \tilde{f}_{b3}(x_3)$ を改良していくとすると、$q^{\backslash b}(x)$ は次のようになる。

$$q^{\backslash b}(x) \propto \tilde{f}_{a1}(x_1)\, \tilde{f}_{a2}(x_2)\, \tilde{f}_{c2}(x_2)\, \tilde{f}_{c4}(x_4)$$

$q^{new}(x)$ は次の式の式のモーメントと一致させればよい。

$$\hat{p}(x) = q^{\backslash b}(x)\, f_b(x_2, x_3) = \tilde{f}_{a1}(x_1)\, \tilde{f}_{a2}(x_2)\, \tilde{f}_{c2}(x_2)\, \tilde{f}_{c4}(x_4)\, f_b(x_2, x_3)$$

ここのモーメントマッチングで少し工夫ができる。もともとモーメントマッチングはKLダイバージェンス　$KL(\hat{p} \mid q^{new})$ を最小化することが目的であった。このKLダイバージェンスを最小化する近似分布は、近づけたい分布の周辺分布で表現できる（ref. (10.17)式）。ただし、この操作は全ての変数が独立で、近似分布を完全に分解できる場合に限る。各変数における近似分布は次のようになる。

$$
\begin{aligned}
\hat{p}(x_1) &\propto \int \hat{p}(x)\, dx_2 dx_3 dx_4 \\
&= \int \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4) f_b(x_2, x_3)\, dx_2 dx_3 dx_4 \\
&\propto \tilde{f}_{a1}(x_1) \\
\hat{p}(x_2) &\propto \int \hat{p}(x)\, dx_1 dx_3 dx_4 \\
&= \int \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4) f_b(x_2, x_3)\, dx_1 dx_3 dx_4 \\
&\propto \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \int f_b(x_2, x_3)\, dx_3 \\
\hat{p}(x_3) &\propto \int \hat{p}(x)\, dx_1 dx_2 dx_4 \\
&= \int \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4) f_b(x_2, x_3)\, dx_1 dx_2 dx_4 \\
&\propto \int \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) f_b(x_2, x_3)\, dx_2 \\
\hat{p}(x_4) &\propto \int \hat{p}(x)\, dx_1 dx_2 dx_3 \\
&= \int \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4) f_b(x_2, x_3)\, dx_1 dx_2 dx_3 \\
&\propto \tilde{f}_{c4}(x_4)
\end{aligned}
$$

求めたい近似分布はこれらの掛け合わせとなる。近似分布について独立を仮定しているためこのようになる

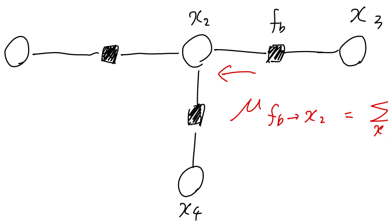
$$
q^{new}(x) \propto \hat{p}(x_1) \hat{p}(x_2) \hat{p}(x_3) \hat{p}(x_4)
$$

EP法の手続きに従うと因子の更新式は次のようになる。

$$
\begin{aligned}
\tilde{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2) \tilde{f}_{b3}(x_3) &\propto \frac{q^{new}(x)}{q^{\backslash b}(x)} \\
&\propto \frac{\hat{p}(x_1) \hat{p}(x_2) \hat{p}(x_3) \hat{p}(x_4)}{\tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4)} \\
&= \frac{\tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \int f_b(x_2, x_3)\, dx_3 \int \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) f_b(x_2, x_3)\, dx_2\, \tilde{f}_{c4}(x_4)}{\tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4)} \\
&= \int f_b(x_2, x_3)\, dx_3 \int \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) f_b(x_2, x_3)\, dx_2
\end{aligned}
$$

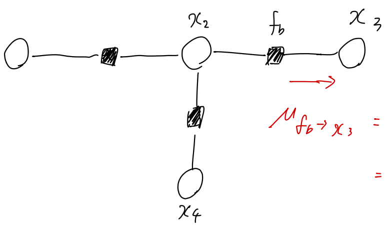
$$\therefore \begin{cases} \tilde{f}_{b2}(x_2) = \int f_b(x_2, x_3)\, dx_3 \\ \tilde{f}_{b3}(x_3) = \int \tilde{f}_{a2}(x_2)\, \tilde{f}_{c2}(x_2)\, f_b(x_2, x_3)\, dx_2 \end{cases}$$

これらの式はメッセージパッシングにおけるメッセージと酷似している。(8.81)式、(8.78),(8.79)式あたりを参照する。



$$\mu_{f_b \to x_2} = \sum_{x_3} f_b(x_2, x_3)$$

(8.81)式が $\tilde{f}_{b2}(x_2)$ と同じ形



$$\mu_{f_b \to x_3} = \sum_{x_2} f_b(x_2, x_3)\, \mu_{x_2 \to f_b}(x_2)$$

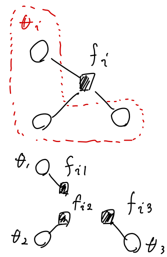
$$= \sum_{x_2} f_b(x_2, x_3)\, \mu_{f_a \to x_2}\, \mu_{f_c \to x_2}$$

(8.79)(8.78)式が $\tilde{f}_b(x_3)$ と同じ形

このことから、積和アルゴリズムと同じように局所的な計算とメッセージの送信によってEP法を適用できることがわかる。特に、木構造を持つグラフの場合には葉から根のメッセージパッシングと根から葉のメッセージパッシングの2回の操作で近似分布の厳密解を得られる。

ここまで4変数のグラフィカルモデルについて考えてきたが、より一般的なグラフィカルモデルにおける近似分布の導出を考える。基本的に4変数の場合と比べて変数が置き換わっただけで操作として同じである。

同時分布は次のような因子グラフで表現できるとする。



$$p(\theta) = \prod_i f_i(\theta_i)$$

完全に分解した近似分布は次のように表現される。

$$q(\theta) \propto \prod_i \prod_k \tilde{f}_{ik}(\theta_k)$$

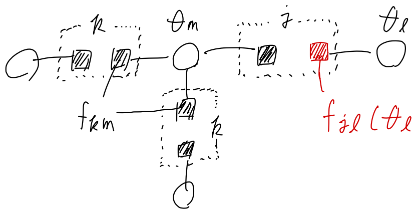
$\tilde{f}_{jl}(\theta)$ を改良したいとする。 $\tilde{f}_j(\theta)$ を取り除いた近似分布は次のようになる。

$$q^{\backslash j}(\theta) \propto \prod_{i \neq j} \prod_k f_{ik}(\theta_k)$$

次に $q^{new}(\theta)$ を求めるために $q^{\backslash j}(\theta)\, f_j(\theta_j)$ とモーメントマッチングを行うのが通常のEP法だが、完全分離された近似分布の元ではモーメントマッチングの必要がない。代わりに $q^{\backslash j}(\theta)\, f_j(\theta_j)$ の周辺化を実施し、全ての周辺分布を掛け合わせることで $q^{new}(\theta)$ を得る。

更新された因子分布を求める操作はレジュメ6頁下とほとんど同じである。注目すべき点として、着目している因子と隣接している変数以外の多くの因子は、$q^{new}(\theta) / q^{\backslash j}(\theta)$ の操作で約分される。最終的には次の式が得られる。（計算略）

$$\tilde{f}_{jl}(\theta_l) \propto \sum_{\theta_m \in \theta_j} f_j(\theta_j) \prod_k \prod_{m \neq l} \tilde{f}_{km}(\theta_m)$$



$f_{jl}(\theta_l)$: メッセージパッシングにおける f_j から θ_l へのメッセージ。

これは、(10.235) 式とも整合していることを確認できる。式の解釈はメッセージパッシングの時と同様で、他の因子ノードから来るメッセージの積をとり、自身の因子をかけて周辺化を行うことに相当する。

以上より、近似分布が完全分離できる場合のEP法の手続きが、メッセージパッシングと同じであることを確認した。