第10章　近似推論法

確率モデルを適用する際の中心的なタスクは、観測データが与えられた時の潜在変数の事後分布を求めること、さらにこの分布に従った期待値を求めることである。前章のEMアルゴリズムにおいても、潜在変数を導入して完全データ対数尤度の期待値を求めることで、真の分布を推定するケースをみてきた。

しかし、事後分布を求めることや、その事後分布に従った期待値を計算することは不可能なことが多い。これは、隠れ変数の空間全体を直接扱うには次元が高すぎることや、事後分布が複雑すぎて期待値を解析的に計算できないことなどが理由である。

このような場合には、**近似法**を用いる必要がある。近似法は、確率的か決定的かによって二つに分けられる。11章で議論されるマルコフ連鎖モンテカルロ法などは確率的な方法であり、無限の計算資源がある場合に厳密な結果を計算することができる。このような確率的な方法は計算量が多く、小さいスケールの問題にしか適用できないことが多い。

本章では大規模な問題にも適用できる決定的な近似法をいくつか紹介する。これらは比較的大事後分布を解析的に近似する方法に基づいており、例えば事後分布が特定の方法で分解されることや、ガウス分布のようなパラメトリックな分布となることを仮定する。特に、**変分推論法（variation inference）**や**変分ベイズ法（variation also Bayes）**と呼ばれる変分を用いた関数近似の手法や、**EP 法（Expectation Propagation）**と呼ばれる手法を扱う。

10.1 変分推論

（変分法についてはキョーマの方がよっぽど詳しいので略します。）変分法の目的は、汎関数を最大 or 最小にする関数を求めることであり、~~オイラー-ラグランジュ方程式を解くことでその解を得られる~~。（オイラーラグランジュ方程式は特別な仮定を置いた元での解の方程式でした。）

この節の主たる論点は、事後分布の近似に変分法をどのように活用できるか、という点である。ネタバレをすると、KLダイバージェンスを汎関数とみなして、これを最小にするパラメトリックな分布関数を算出するために変分法が使われる。

我々の目的は、事後分布 p(**Z**|**X**) およびモデルエビデンス p(**X**) を求めることである。潜在変数についての任意の分布 q(**Z**) を導入すると、EM 法の議論と同じように、周辺分布の対数は次のように分解できる。

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q \| p) \tag{10.2}$$

ここで、

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \tag{10.3}$$

$$\mathrm{KL}(q \| p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} | \mathbf{X})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \tag{10.4}$$

前章と同じように、下界 $\mathcal{L}(q)$ を分布 q(**Z**) について最大化するが、これは KL ダイバージェンスを最小化することと同値である。 q(**Z**) = p(**Z**|**X**) のとき KL ダイバージェンスは 0 となるが、今回は真の事後分布がわからないという点が前回と異なる。したがって代わりに、ある制限したクラスの分布 q(**Z**) を考え、この中で下界 $\mathcal{L}(q)$ を最大にするものを探す。このときの制限についての教科書の記述がよくわからなかった。「我々の目的は」から始まる部分。

このときの $\mathcal{L}(q)$ の最大化に変分法が適用される。すなわち、$\mathcal{L}(q)$ が q(**Z**) に関する汎関数であり、これを変分法で解くことで得られる $q^*(\mathbf{Z})$ が $\mathcal{L}(q)$ を最大化する。これにより $\mathrm{KL}(q\|p)$ が最小化され、真の事後分布をよく近似する分布が得られるのである。

制限したクラスの分布の例としては、ガウス分布などのパラメトリックな分布を使うことである。図10.1 は変分近似をガウス分布とし、元の分布を近似した例である。ラプラス近似と比べると、モードの位置が異なり、より元の分布との重なりが大きいように見える。

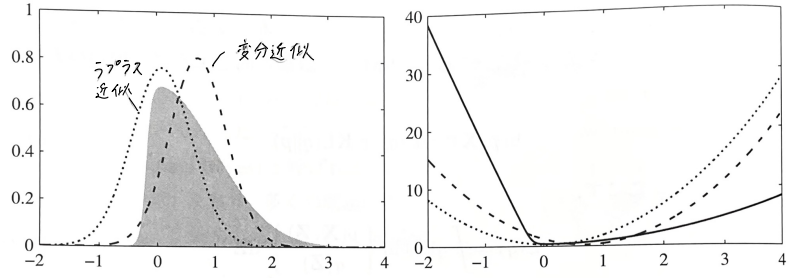


図 10.1 図 4.14 の例題の変分近似による結果．左図に，もともとの分布（灰色・実線）をラプラス近似（点線）および変分近似（鎖線）とともに示す．右図は，同じ曲線を負の対数をとって表したもの．

10.1.1 分布の分解

ここでは、 q(Z) に関して別の制限をもつ分布について変分推論を適用した結果を見てみる。特に、平均場近似と呼ばれる分解の性質を仮定する。このときに $\mathcal{L}(q)$ を最大にする q(Z) について、$q_i(Z_i)$ を逐次的に更新する式を得られることが利点となる。

q(Z) が次のように分解できると仮定する。

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i) \qquad (10.5)$$

この仮定は各 $\mathbf{Z}_i$ が独立であることを表しているが、物理学において**平均場近似（mean field approximation）**と呼ばれる近似法に対応している。ここでの分布はこれ以上の仮定を置いておらず、特にパラメトリックな分布なども仮定していないことに注意する。

(10.5) の形をもつ $q(\mathbf{Z})$ の中で、下界 $\mathcal{L}(q)$ を最大になるものを探したい。ここで、下界を分布 $q_i = q_i(\mathbf{Z}_i)$ で分解して、各 i について順番に変分最適化を実施する。

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z} && (\because 式(10.3)) \\
&= \int \prod_i q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \right\} d\mathbf{Z} && (\because 式(10.5)) \\
&= \int_{\mathbf{Z}_{i\neq j}} \int_{\mathbf{Z}_j} \prod_i q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \right\} d\mathbf{Z}_j \, d\mathbf{Z}_{i\neq j} \\
&= \iint q_j \prod_{i\neq j} q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \ln q_j - \sum_{i\neq j} \ln q_i \right\} d\mathbf{Z}_j \, d\mathbf{Z}_{i\neq j}
\end{aligned}$$

方針：q_j に依存しない項を const. に出していく。

$$\begin{aligned}
&= \iint q_j \prod_{i\neq j} q_i \left(\ln p(\mathbf{X},\mathbf{Z}) \right) d\mathbf{Z}_j \, d\mathbf{Z}_{i\neq j} \\
&\quad - \iint q_j \prod_{i\neq j} q_i \left(\ln q_j \right) d\mathbf{Z}_j \, d\mathbf{Z}_{i\neq j} \\
&\quad - \iint q_j \prod_{i\neq j} q_i \left(\sum_{i\neq j} \ln q_i \right) d\mathbf{Z}_j \, d\mathbf{Z}_{i\neq j} \\
&= \int_{\mathbf{Z}_j} q_j \left\{ \int_{\mathbf{Z}_{i\neq j}} \ln p(\mathbf{X},\mathbf{Z}) \prod_{i\neq j} q_i \, d\mathbf{Z}_{i\neq j} \right\} d\mathbf{Z}_j \\
&\quad - \int q_j \ln q_j \, d\mathbf{Z}_j \int \prod_{i\neq j} q_i \, d\mathbf{Z}_{i\neq j} \\
&\qquad \{\mathbf{Z}\}_{i\neq j} \text{ の同時確率} \quad (\because 10.5) \\
&\quad - \int q_j \, d\mathbf{Z}_j \cdot \int \prod_{i\neq j} q_i \left(\sum_{i\neq j} \ln q_i \right) d\mathbf{Z}_{i\neq j}
\end{aligned}$$

$$= \int q_j \left\{ \int \ln p(X, Z) \prod_{i \neq j} q_i \, dZ_{i \neq j} \right\} dZ_j$$
$$- \int q_j \ln q_j \, dZ_j \cdot 1$$
$$- 1 \cdot \int \prod_{i \neq j} q_i \left(\sum_{i \neq j} \ln q_i \right) dZ_{i \neq j} \quad (\because \text{全確率の公式})$$

$$= \int q_j \left\{ \int \ln p(X, Z) \prod_{i \neq j} q_i \, dZ_{i \neq j} \right\} dZ_j - \int q_j \ln q_j \, dZ_j + \text{const.}$$

$$= \int q_j \, E_{i \neq j} \left[\ln p(X, Z) \right] dZ_j - \int q_j \ln q_j \, dZ_j + \text{const.} \quad (\because \prod_{i \neq j} q_i \text{ は } \{Z\}_{i \neq j} \text{ の同時確率})$$

$$= \int q_j \ln \tilde{p}(X, Z_j) \, dZ_j - \int q_j \ln q_j \, dZ_j + \text{const.} \quad (10.6)$$

教科書はここまで。

$$= \int q_j \left\{ \ln \frac{\tilde{p}(X, Z_j)}{q_j} \right\} dZ_j + \text{const.}$$

$$= -KL(q \| \tilde{p}) + \text{const.} \quad (10.6^*)$$

途中で、以下の式を定義した。

$$\ln \tilde{p}(X, Z_j) = E_{i \neq j} \left[\ln p(X, Z) \right] + \text{const.}$$

下界 $\mathcal{L}(q)$ の分解により得られた式を $q_j(Z_j)$ について変分することで最大値を満たす $q_j(Z_j)$ を得る。ただし、今回のケースは~~オイラーラグランジュ方程式を解くまでもなく~~（そもそも使うことはできない）、KL ダイバージェンスの性質から次の式が成り立つときに (10.6*) が最大化される。

$$\ln q_j^*(Z_j) = \ln \tilde{p}(X, Z_j)$$
$$= E_{i \neq j} \left[\ln p(X, Z) \right] + \text{const.}$$

この式が示すことについて考察してみる。これは、因子 q_j の最適解の対数は、観測データと隠れ変数の同時分布の対数を考え、$i \neq j$ である他の因子 q_i すべてについての期待値をとったものに等しい、という意味になっている。

全ての因子について最適解を求めるためには、この更新を j をずらしながら順に実行する必要がある。というのも、(10.9) で得られた式は他の因子 q_i に依存しており、他の因子が更新されるとその期待値も変化するためである。

10.1.2 分解による近似のもつ性質

前節では下界を最大化することで事後分布を近似する手法を確かめたが、これは事後分布を分解によって近似することに基づいている。ここで、一般に分布を分解して近似することによって、どんな不正確さが生じるかについて考察する。

まず、事後分布が二変量ガウス分布な場合に、独立な単変量ガウス分布に分解する例を見てみる。この例の結論としては、近似分布によって平均は正確に捉えられるが、共分散が大きくはずれる結果となる。（独立性を仮定したので、当然といえば当然だが。）

（ここで疑問。事後分布をガウス分布で仮定するという話の流れのようだが、この後に式では同時分布 p(**X**,**Z**) をガウス分布として 10.9 を用いている。これは同時分布を **Z** について着目したときにガウス分布になっている、という解釈でいいのだろうか。）

真の事後分布　$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

近似させる分布　$q(\mathbf{z}) = q_1(z_1)\, q_2(z_2)$

ここで、$q_1(z_1)$, $q_2(z_2)$ はガウス分布を仮定していないことに注意する。このとき、(10.9) を適用して、$q_1(z_1)$, $q_2(z_2)$の変分最適解を求めると、それぞれガウス分布が導出される。

$$
\begin{aligned}
\ln q_1^*(z_1) &= E_{z_2}[\ln p(\mathbf{x}, \mathbf{z})] + \text{const.} \qquad \because \text{式(10.9)} \\
&= E_{z_2}[\ln p(\mathbf{z})] + \text{const.} \qquad \text{（上の「ここで疑問」にも書いたけど、$\mathbf{z}$ のみに注目するので $\mathbf{x}$ は省略した解釈する）} \\
&= E_{z_2}\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^{\top}\boldsymbol{\Lambda}(\mathbf{z}-\boldsymbol{\mu})\right] + \text{const.} \\
&= E_{z_2}\left[-\frac{1}{2}(z_1-\mu_1)^2\Lambda_{11} - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + \text{const.} \\
&= -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_1\Lambda_{11} - z_1\Lambda_{12}(E[z_2]-\mu_2) + \text{const.}
\end{aligned}
$$

これを平方完成すると次のガウス分布が得られる。

$$
\begin{aligned}
&q_1^*(z_1) = \mathcal{N}(z_1 \mid m_1, \Lambda_{11}^{-1}) \\
&\text{ここで}\quad m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(E[z_2]-\mu_2)
\end{aligned}
$$

同様に q_2^* についてもガウス分布が得られる。

$$
\begin{aligned}
&q_2^*(z_2) = \mathcal{N}(z_2 \mid m_2, \Lambda_{22}^{-1}) \\
&\text{ここで}\quad m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(E[z_1]-\mu_1)
\end{aligned}
$$

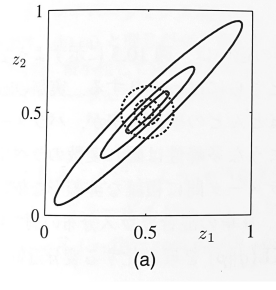
この更新式は一方の因子の期待値に依存していることから、一般的な解法としては繰り返し更新が必要となる。ただし、今回の例は十分簡単なため、期待値に関して次のような解が得られることがわかる。

$$\begin{cases} \mathbb{E}[z_1] = \mu_1 \\ \mathbb{E}[z_2] = \mu_2 \end{cases}$$

これは平均パラメータの更新式を方程式として見たときに、同時に方程式を満たしていることから確かめられる。これより分布は次のようになる。

$$q_1^*(z_1) = \mathcal{N}(z_1 \mid \mu_1, \Lambda_{11}^{-1})$$
$$q_2^*(z_2) = \mathcal{N}(z_2 \mid \mu_2, \Lambda_{22}^{-1})$$

この近似分布と本来の事後分布を比べると、平均は一致しているが、分散が過小評価されていることがわかる（図10.2 (a)）。一般に、分解による近似は分散が過小評価される傾向にある（らしい）。



(a)

10.1.1 節、 10.1.2 節ともに、下界 $\mathcal{L}(q)$ を最大にすることを考えていたが、(10.4) 式の KL ダイバージェンス KL(q || p) の**逆の** KL ダイバージェンス KL(p || q) を最小化する場合にどうなるかを考察する。実はこの場合繰り返し法を必要としない閉形式の解が求まる。
（なぜ逆の KL ダイバージェンスを考えて良いのかは後々の宿題とさせてください。おそらく EP 法あたりで出てきそう）

(10.4) に関する逆の KL ダイバージェンスを、分解の仮定を用いて展開する。

$$
\begin{aligned}
KL(p\|q) &= -\int p(\mathbf{Z}|\mathbf{X}) \ln \left\{ \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \right\} d\mathbf{Z} \\
&= -\int p(\mathbf{Z}) \ln \left\{ \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \right\} d\mathbf{Z} \qquad (\because \mathbf{X} \text{を省略}) \\
&= -\int p(\mathbf{Z}) \ln \left\{ \frac{\prod_{i=1}^{M} q_i(\mathbf{Z}_i)}{p(\mathbf{Z})} \right\} d\mathbf{Z} \qquad (\because \text{式}(10.5)) \\
&= -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] d\mathbf{Z} + \int p(\mathbf{Z}) \ln p(\mathbf{Z}) d\mathbf{Z} \\
&= -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] d\mathbf{Z} - \underbrace{H(p)}_{\text{エントロピー}} \qquad (\because \text{式}(1.98)) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (10.16)
\end{aligned}
$$

これを因子 q_j について最適化するためにラグランジュ未定乗数法を適用すると、次の解が得られる。（演習10.3 余裕あれば後で追記します）これは閉形式であり、繰り返しが必要でない点に注意する。

（周辺化を解析的に解けるとは限らないので、こちらが万能というわけではなさそう）

$$
q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j)
$$

この近似により得られる分布は下図のように分散が過大評価される結果となる。この辺りの定性的な性質の違いは、次回以降の担当範囲にお任せします。