変分ベイズ推論 (1)

正田 備也

masada@rikkyo.ac.jp

Contents

前回のEMアルゴリズムでの議論のパターン

- 潜在変数 $\mathcal{Z} = \{z_1, \ldots, z_N\}$ を含むモデリングを行いたい
- 確率モデルを指定することで同時分布
  $p(\mathcal{X}, \mathcal{Z}) = p(\mathcal{Z})p(\mathcal{X}|\mathcal{Z}) = \prod_{i=1}^{N} p(z_i)p(x_i|z_i)$ が得られる
- 潜在変数 $\mathcal{Z}$ の周辺化 $\sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z})$ により観測データの尤度
  $p(\mathcal{X})$ は得られるのだが、大抵この尤度は計算できない
- Jensen の不等式を使い、対数尤度 $\ln p(\mathcal{X})$ の下界を得る

$$\ln p(\mathcal{X}) \geq \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}}$$

- この下界を最大化することで、様々な未知量を推定する

ベイズ的なモデリング

- 観測データを表す確率変数を $\mathcal{X} \equiv \{x_1, \ldots, x_N\}$ とする
- 確率モデルのパラメータを表す確率変数を $\mathbf{\Theta}$ とする
- ベイズ的なモデリングで知りたいのは、事後分布 $p(\mathbf{\Theta}|\mathcal{X})$

$$p(\mathbf{\Theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathcal{X})} \tag{1}$$

- 変分ベイズ推論は $p(\mathbf{\Theta}|\mathcal{X})$ を近似する分布 $q(\mathbf{\Theta})$ を求める
  - $q(\mathbf{\Theta})$ を変分事後分布 (variational posterior distribution) と呼ぶ
  - $q(\mathbf{\Theta})$ には比較的扱いやすい分布を選ぶ

EMアルゴリズムでの議論のパターンを適用

- 潜在変数 $\mathbf{\Theta}$ を含むモデリングを行いたい
- 確率モデルを指定することで観測データと潜在変数の同時分布 $p(\mathcal{X}, \mathbf{\Theta}) = p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta}) = p(\mathbf{\Theta}) \prod_{i=1}^{N} p(x_i|\mathbf{\Theta})$ が得られる
- 潜在変数 $\mathbf{\Theta}$ の周辺化 $\int p(\mathcal{X}, \mathbf{\Theta}) d\mathbf{\Theta}$ により観測データの周辺尤度 $p(\mathcal{X})$ は得られるのだが、大抵この尤度は計算できない
- Jensen の不等式を使い、対数周辺尤度 $\ln p(\mathcal{X})$ の下界を得る

$$\ln p(\mathcal{X}) \geq \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}$$

- この下界を最大化することで、様々な未知量を推定する
  - この下界を ELBO(Evidence Lower BOund; 変分下限) と呼ぶ

変分ベイズ推論 (variational inference) とは

- Jensen の不等式を適用することで、ELBO を次のように得た

$$\ln p(\mathcal{X}) \geq \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}$$

- 実は、ELBO を大きくすればするほど、$\mathbf{\Theta}$ が従う確率分布である $q(\mathbf{\Theta})$ が、事後分布 $p(\mathbf{\Theta}|\mathcal{X})$ に近くなっていく
- つまり、この $q(\mathbf{\Theta})$ は、事後分布を近似する分布とみなせるような分布になっている
- $q(\mathbf{\Theta})$ を、変分事後分布 (variational posterior) と呼ぶ

対数周辺尤度と ELBO の差

- Jensen の不等式の左辺から右辺を引いたものを求めてみる

$$
\begin{aligned}
&\ln p(\mathcal{X}) - \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta}|\mathcal{X})p(\mathcal{X})}{q(\mathbf{\Theta})} d\mathbf{\Theta} \\
&= \ln p(\mathcal{X}) - \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta}|\mathcal{X})}{q(\mathbf{\Theta})} d\mathbf{\Theta} - \int q(\mathbf{\Theta}) \ln p(\mathcal{X}) d\mathbf{\Theta} \\
&= \ln p(\mathcal{X}) - \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta}|\mathcal{X})}{q(\mathbf{\Theta})} d\mathbf{\Theta} - \ln p(\mathcal{X}) \int q(\mathbf{\Theta}) d\mathbf{\Theta} \\
&= \int q(\mathbf{\Theta}) \ln \frac{q(\mathbf{\Theta})}{p(\mathbf{\Theta}|\mathcal{X})} d\mathbf{\Theta} = D_{\mathsf{KL}}(q(\mathbf{\Theta}) \parallel p(\mathbf{\Theta}|\mathcal{X})) \qquad (2)
\end{aligned}
$$

$\therefore$ ELBO を $\ln p(\mathcal{X})$ に近づける $\Leftrightarrow$ $q(\mathbf{\Theta})$ を $p(\mathbf{\Theta}|\mathcal{X})$ に近づける

「変分 (variational)」の意味

- ELBO の最大化は、$q(\mathbf{\Theta})$ を変化させることでおこなう
- このとき、$q(\mathbf{\Theta})$ の密度関数そのものを変化させる
- 逆に言うと、$q(\mathbf{\Theta})$ の密度関数が特定のかたちを持つと仮定した上で、その関数のパラメータを動かすのではない
  - パラメータについて微分することで最大化問題を解くのではなく、いわば “関数について微分する” ことで最大化問題を解いている
- 関数のかたちを決めてそのパラメータを動かすのではなく、関数のかたち自体を動かすことで問題を解く方法を、変分法と呼ぶ（cf. 汎関数微分）

実際の変分推論

- 実際には、$q(\mathbf{\Theta})$ の密度関数が特定のかたちを持つと仮定した上で、その関数のパラメータを動かすことによって、ELBO を最大化することも多い

$$\ln p(\mathcal{X}) \geq \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}$$

  - 例えば、$q(\mathbf{\Theta})$ が多変量正規分布だと仮定して、ELBO を最大化するような平均パラメータと共分散行列パラメータを求める、など
  - 変分オートエンコーダでは、$q(\mathbf{\Theta})$ が多変量正規分布だと仮定し、さらにその共分散行列が対角行列だと仮定する

Contents

例：メッセージ受信数の変化点の検知

- この授業の最初に採り上げた例

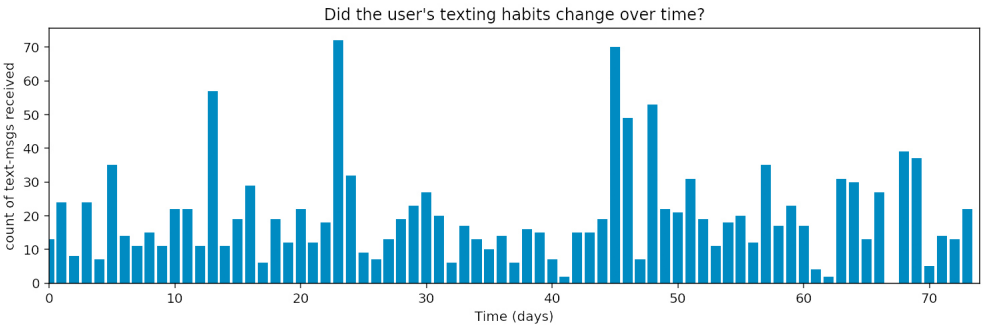

Figure: メッセージの受信数

モデルを指定する

- c_n が n 日目の受信数、τ が受信数の変化点、λ_1, λ_2 がそれぞれ $n < \tau, n \geq \tau$ の場合のポアソン分布のパラメータとする

$$
\begin{aligned}
\tau &\sim \texttt{Uniform}(1, N) \\
\lambda_1 &\sim \texttt{Gam}(a, b) \\
\lambda_2 &\sim \texttt{Gam}(a, b) \\
c_n &\sim \texttt{Poi}(\lambda_1) \quad \text{for } n < \tau \\
c_n &\sim \texttt{Poi}(\lambda_2) \quad \text{for } n \geq \tau
\end{aligned}
$$

- 参考: http://machine-learning.hatenablog.com/entry/2017/08/19/200841

ELBO を求める

同時分布は、$\boldsymbol{c} = \{c_1, \ldots, c_N\}$ として

$$
\begin{aligned}
p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau) &= p(\boldsymbol{c}|\lambda 1, \lambda_2, \tau) p(\lambda_1; a, b) p(\lambda_2; a, b) p(\tau) \\
&= p(\lambda_1; a, b) p(\lambda_2; a, b) p(\tau) \prod_{n=1}^{N} p(c_n|\lambda_1)^{\delta(n<\tau)} p(c_n|\lambda_2)^{\delta(n\geq\tau)} \qquad (3)
\end{aligned}
$$

ELBO は

$$
\begin{aligned}
\ln p(\boldsymbol{c}) &= \ln \int \sum_{\tau} p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau) d\lambda_1 d\lambda_2 \\
&\geq \int \sum_{\tau} q(\lambda_1, \lambda_2, \tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1, \lambda_2, \tau)} d\lambda_1 d\lambda_2 \qquad (4)
\end{aligned}
$$

平均場近似

- 変分事後分布 $q(\lambda_1, \lambda_2, \tau)$ について、
  $q(\lambda_1, \lambda_2, \tau) = q(\lambda_1)q(\lambda_2)q(\tau)$ と分解できることを仮定する
  - このような仮定を平均場近似 (mean field approximation) と呼ぶ

$$\begin{aligned}\ln p(\boldsymbol{c}) &\geq \int \sum_{\tau} q(\lambda_1, \lambda_2, \tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1, \lambda_2, \tau)} d\lambda_1 d\lambda_2 \\ &= \int \sum_{\tau} q(\lambda_1)q(\lambda_2)q(\tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1)q(\lambda_2)q(\tau)} d\lambda_1 d\lambda_2 \qquad (5)\end{aligned}$$

- ELBO の最大化 $\Leftrightarrow D_{\mathrm{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}))$ の最小化（cf. 式 (2)）

$q(\lambda_1)$ のかたちを求める

$q(\lambda_2)$ と $q(\tau)$ を固定し、$D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}))$ を最小にする $q(\lambda_1)$ を求める。

$$
\begin{aligned}
& D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})) \\
&= \int \sum_{\tau} q(\lambda_1)q(\lambda_2)q(\tau) \ln \frac{q(\lambda_1)q(\lambda_2)q(\tau)}{p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})} d\lambda_1 d\lambda_2 \\
&= \int q(\lambda_1)\big\{ \ln q(\lambda_1) - \int \sum_{\tau} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2 \big\} d\lambda_1 + const. \\
&= \int q(\lambda_1) \ln \frac{q(\lambda_1)}{\exp \int \sum_{\tau} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2} d\lambda_1 + const. \\
&= D_{\mathsf{KL}}(q(\lambda_1) \parallel \frac{1}{Z} \exp \int q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2 d\tau) + const. \qquad (6)
\end{aligned}
$$

$q(\lambda_1) = \frac{1}{Z} \exp \int \sum_{\tau} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2$ のとき、上の KL 情報量は最小。つまり、
$\ln q(\lambda_1) = \int \sum_{\tau} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2 - \ln Z$ のとき、上の KL 情報量は最小。

$$\ln q(\lambda_1) = \int \sum_{\tau} q(\lambda_2) q(\tau) \ln \Big\{ p(\lambda_1; a, b) p(\lambda_2; a, b) p(\tau) \prod_{n=1}^{N} p(c_n|\lambda_1)^{\delta(n<\tau)} p(c_n|\lambda_2)^{\delta(n\geq\tau)} \Big\} d\lambda_2 - \ln Z$$

$$= \ln p(\lambda_1; a, b) + \int q(\lambda_2) \ln p(\lambda_2; a, b) d\lambda_2 + \sum_{\tau} q(\tau) \ln p(\tau)$$

$$+ \sum_{n=1}^{N} \sum_{\tau} q(\tau) \delta(n < \tau) \ln p(c_n|\lambda_1) + \sum_{n=1}^{N} \int \sum_{\tau} q(\lambda_2) q(\tau) \delta(n \geq \tau) \ln p(c_n|\lambda_2) d\lambda_2 - \ln Z$$

$$= \ln \frac{b^a}{\Gamma(a)} \lambda_1^{a-1} e^{-b\lambda_1} + \sum_{n=1}^{N} \Big(\sum_{\tau} q(\tau) \delta(n < \tau) \Big) \ln \frac{\lambda_1^{c_n} e^{-\lambda_1}}{c_n!} + const.$$

$$= \Big(a - 1 + \sum_{n=1}^{N} \Big(\sum_{\tau} q(\tau) \delta(n < \tau) \Big) c_n \Big) \ln \lambda_1 - \Big(b + \sum_{n=1}^{N} \Big(\sum_{\tau} q(\tau) \delta(n < \tau) \Big) \Big) \lambda_1 + const.$$

よって、$q(\lambda_1)$ は、shape パラメータが $a + \sum_{n=1}^{N} \Big(\sum_{\tau} q(\tau) \delta(n < \tau) \Big) c_n$ で、rate パラメータが $b + \sum_{n=1}^{N} \Big(\sum_{\tau} q(\tau) \delta(n < \tau) \Big)$ のガンマ分布となる。
$q(\lambda_2)$ についても同様に計算すると、やはりガンマ分布であることが分かる。

$q(\tau)$ のかたちを求める

$q(\lambda_1)$ と $q(\lambda_2)$ を固定し、$D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}))$ を最小にする $q(\tau)$ を求める。

$$
\begin{aligned}
&D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})) \\
&= \int \sum_{\tau} q(\lambda_1)q(\lambda_2)q(\tau) \ln \frac{q(\lambda_1)q(\lambda_2)q(\tau)}{p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})} d\lambda_1 d\lambda_2 \\
&= \sum_{\tau} q(\tau)\big\{ \ln q(\tau) - \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2 \big\} + const. \\
&= \sum_{\tau} q(\tau) \ln \frac{q(\tau)}{\exp \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2} + const. \\
&= D_{\mathsf{KL}}(q(\tau) \parallel \frac{1}{Z} \exp \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2) + const. \qquad (7)
\end{aligned}
$$

$q(\tau) = \frac{1}{Z} \exp \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2$ のとき、上の KL 情報量は最小。つまり、
$\ln q(\tau) = \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2 - \ln Z$ のとき、上の KL 情報量は最小。

$$\ln q(\tau) = \int q(\lambda_1)q(\lambda_2) \ln \left\{ p(\lambda_1; a, b)p(\lambda_2; a, b)p(\tau) \prod_{n=1}^{N} p(c_n|\lambda_1)^{\delta(n<\tau)} p(c_n|\lambda_2)^{\delta(n\geq\tau)} \right\} d\lambda_1 d\lambda_2 - lnZ$$

$$= \int q(\lambda_1) \ln p(\lambda_1; a, b) d\lambda_1 + \int q(\lambda_2) \ln p(\lambda_2; a, b) d\lambda_2 + \ln p(\tau)$$

$$+ \sum_{n=1}^{N} \delta(n < \tau) \int q(\lambda_1) \ln p(c_n|\lambda_1) d\lambda_1 + \sum_{n=1}^{N} \delta(n \geq \tau) \int q(\lambda_2) \ln p(c_n|\lambda_2) d\lambda_2 + const. \tag{8}$$

よって、$q(\tau)$ はカテゴリカル分布であり、

$$q(\tau) \propto \exp \left[\sum_{n=1}^{N} \delta(n < \tau) \int q(\lambda_1) \ln p(c_n|\lambda_1) d\lambda_1 + \sum_{n=1}^{N} \delta(n \geq \tau) \int q(\lambda_2) \ln p(c_n|\lambda_2) d\lambda_2 \right]$$

となる。ただし、$\sum_{\tau=1}^{T} q(\tau) = 1$ を満たす。

まとめ

- メッセージ受信数の変化点を検知するため、ベイズ的なモデルを立てた
- 事後分布を近似するために、変分ベイズ推論を行った
- その際、変分事後分布 $q(\lambda_1, \lambda_2, \tau)$ について、$q(\lambda_1, \lambda_2, \tau) = q(\lambda_1)q(\lambda_2)q(\tau)$ と分解できることを仮定した
- このように仮定すると、$q(\lambda_1)$ と $q(\lambda_2)$ はガンマ分布となり、$q(\tau)$ はカテゴリカル分布となった

課題9

- メッセージ受信数の変化点検知の例を考える。
- λ_1 の値が従う変分事後分布 $q(\lambda_1)$ は、ガンマ分布であることが分かった。
- そこで、$q(\lambda_1)$ の shape パラメータを α_1 とし、rate パラメータを β_1 とする。
- このとき、$\int q(\lambda_1) \ln p(\lambda_1; a, b) d\lambda_1$ を計算せよ。
  - これは ELBO の算出に必要な計算。