Pattern Recognition and Machine Learning

第10章　近似推論法

修士2年
松村草也

BehaviorInNetworks
都市生活学・ネットワーク行動学研究室

目次

変分について

- 微分は，入力値として受け取る変数を少し変化させた時に関数の値がどう変化するか
- 同様に汎関数とは入力として関数を受取り，出力値として汎関数の値を返すもの．
- 汎関数微分を考えることができ，これは入力関数が微小に変わった時の汎関数の値の変化を表している．これを変分という．
- 変分法を用いて，近似解を求めることができる．

変分推論

- 9章で行われた議論を，変文推論を用いて行ってゆく.
- 潜在変数をZとおく．同様に観測変数すべてをXと書く．$X=\{x_1,x_2,\cdots x_N\}$, $Z=\{z_1,z_2,\cdots,z_N\}$
- 同時分布p(X,Z)が与えられた状態で，事後分布p(Z|X)とモデルエビデンスp(X)の近似を求めることを目的とする.
- まず抽象的なイメージでモデルについて考え，あとで具体的に混合ガウス分布に対して適用する.

変分推論

すべてパラメータの事前分布が与えられたモデルがあるとする.
モデルにはパラメータの他に潜在変数がある可能性があり，それをZとおく.

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

尤度関数　　下限　　カルバックライブラー
ダイバージェンス

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

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

KLダイバージェンスが0になるとき，q(Z)は真の事後分布p(Z|X)になる.

変分推論

- EM法の議論とは，パラメータベクトルが現れず，確率変数としてZの中に含まれていることである.
- 積分記号は和に変えることができ，離散変数についても同様の議論が可能.
- 分布q(Z)を変化させて下限を最大化させることを考えるが，これはKLダイバージェンスを最小化=0になるとき.
- 真の事後分布を発見することは難しいので，ある制限をしたクラスq(Z)の中でKLを最小にするものを探す.

10.1.1 分布の分解

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

q(Z)を分解していく．この分布の中で各因子$q_i(Z_i)$の中で，下限L(q)が最大になるものを探したい．10.5を10.3に代入すると，下記の式が得られる．

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} d\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \right\} d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z_j} + \text{const} \qquad (10.6) \\
&= \int q_i \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z_j} + \text{const}
\end{aligned}$$

ただし，

$$\ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const} \qquad (10.7)$$

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \qquad (10.8)$$

10.1.1 分布の分解

最適解q*は一般的に下記の式で与えられる.

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const} \qquad (10.9)$$

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,
$i \neq j$であるほかの因子全てについて期待値をとったものに等しいということになる.

定数項については

$$q_j^*(\mathbf{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]) d\mathbf{Z}_j}$$

10.9の右辺は, 他の因子による期待値に依存していて, 完全な解析解ではない.
求めるためには, 適当にすべての因子に初期値を与えて, 他の因子について
期待値を計算していく必要がある. これは収束することが知られている.
(Boyd and Vandenberghe, 2004)

10.1.2 分解による近似の持つ性質

相関のある2つの潜在変数z=(z_1,z_2)について，下記のガウス分布を考える．

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu, \Lambda^{-1})$$

平均と精度について要素を展開する．精度行列は対称行列である．

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

q(z)=q1(z1)q2(z2)に分解して近似をしたいとする．10.9より，

$$\begin{aligned} \ln q_1^*(z_1) &= \mathbb{E}_{z_2}[\ln p(\mathbf{z})] + \text{const} \\ &= \mathbb{E}_{z_2}\left[-\frac{1}{2}(z_1-\mu_1)^2\Lambda_{11} - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + \text{const} \\ &= -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_1\Lambda_{11} - z_1\Lambda_{12}(\mathbb{E}[z_2]-\mu_2) + \text{const} \end{aligned}$$

q1(z1)がガウス分布に従う，という仮定は特においていないが，
この式の右辺はz1の2次関数なので，q1はガウス分布になる！

10.1.2 分解による近似の持つ性質

結果として下記の式が得られる．q2も同様．

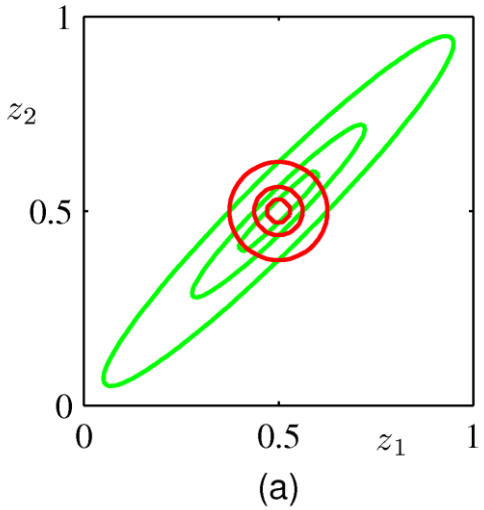
$$\begin{cases} q_1^*(z_1) = \mathcal{N}(z_1 | m_1, \Lambda_{11}^{-1}) \\ q_2^*(z_2) = \mathcal{N}(z_2 | m_2, \Lambda_{22}^{-1}) \end{cases}$$

$$\begin{cases} m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) \\ m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) \end{cases}$$

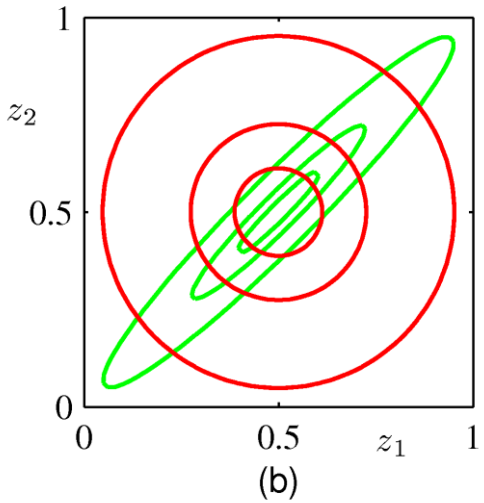
- これらの解には相互関係があり，それぞれを使って得られる期待値に依存している．
- 一般に変分ベイズ法の解を求めるには，これを収束条件を満たすまで順番に更新していく．
- $\mu_1 = m_1$，$\mu_2 = m_2$が収束解になることが予想される．

KLダイバージェンスの形の比較

- 相関のあるz_1z_2を軸にとり，標準偏差の1,2,3倍の等高線を引いたグラフ.
- 緑線はp(z)，赤線はq(z)
- 上はKLダイバージェンスKL(q||p)の，下は負のKLダイバージェンスKL(p||q)の最小化による近似を行ったものである.
- 一般に分解による変分近似は事後分布をコンパクトに近似しすぎる傾向がある.
- 逆に，KL(p||q)の最小化による近似は分布を広くカバーする傾向にある.



(a)



(b)

負のKLダイバージェンスについて

負の場合の表記.

$$KL(p||q) = -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] d\mathbf{Z} + \text{const}$$

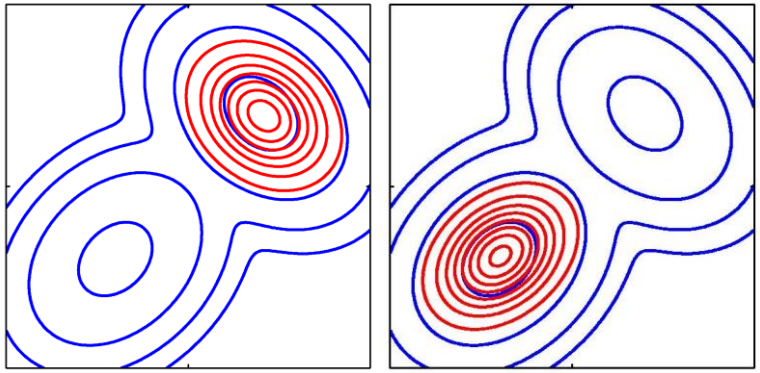
$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j)$$

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

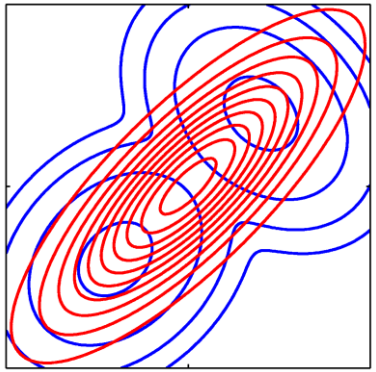
について，大きな正の寄与がp(Z)がほとんどゼロ，q(z)がそうでない領域から来ると考えられる. したがって，q(z)はp(z)が小さい領域を避けるようにする.

負のKLダイバージェンスについて

- 多峰性の分布の場合に変分近似を行った場合を考える
- KL(q||p)を最小化する変分近似はこれらの峰のいずれかを近似することになる.
- 混合モデルを考える場合は，KL(p||q)を用いることは悪い結果をもたらすといえる.



KL(q||p)最小化の意味で，
もっともよく近似するガウス分布q(z)



KL(p||q)最小化の意味で，
もっともよく近似するガウス分布q(z)

KLダイバージェンスまとめ

KLダイバージェンスの2つの形は次式で定義される.

$$\mathrm{D}_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2} dx\right)$$

αは(−∞,∞)の値をとるパラメータ. $\alpha \to 1$の極限がKL(p||q), $\alpha \to -1$の極限がKL(q||p)に当たる.
どんなαの値に対してもDは0以上. 統合はp=qのときに成り立つ.

$$\mathrm{D}_\mathrm{H}(p||q) = \int \left(p(x)^{1/2} - q(x)^{1/2}\right)^2 dx$$

α＝0のときは対称なダイバージェンスになり, ヘリンガー距離と呼ばれる.

10.1.3 例：一変数ガウス分布

- 一変数xについてガウス分布を用いて，分解による変分近似の例を示す．(MacKay,2003)
- ガウス分布から独立に発生した観測値について，もともとのガウス分布の平均 μ と精度 τ の事後分布を求める．

尤度関数：
$$p(\mathcal{D}|\mu,\tau)=\left(\frac{\tau}{2\pi}\right)^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \qquad (10.21)$$

共役事前分布の導入

$$p(\mu|\tau)=\mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1}) \qquad :\text{ガウス分布}$$

$$p(\tau)=\Gamma(\tau|a_0,b_0) \qquad :\text{ガンマ分布}$$

10.1.3 例：一変数ガウス分布

「分解」を行い，変分近似する．

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau) \quad (10.24)$$

$$\begin{aligned}
\ln q_\mu^*(\mu) &= \mathbb{E}\tau\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \text{const} \\
&= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right\} + \text{const} \qquad (10.25)
\end{aligned}$$

$$\begin{aligned}
\ln q_\tau^*(\tau) &= \mathbb{E}_\mu[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)] + \ln p(\tau) + \text{const} \\
&= (a_0-1)\ln\tau - b_0\tau + \frac{N+1}{2}\ln\tau \\
&\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right] + \text{const} \qquad (10.28)
\end{aligned}$$

10.1.3 例：一変数ガウス分布

「分解」を行い，変分近似する．

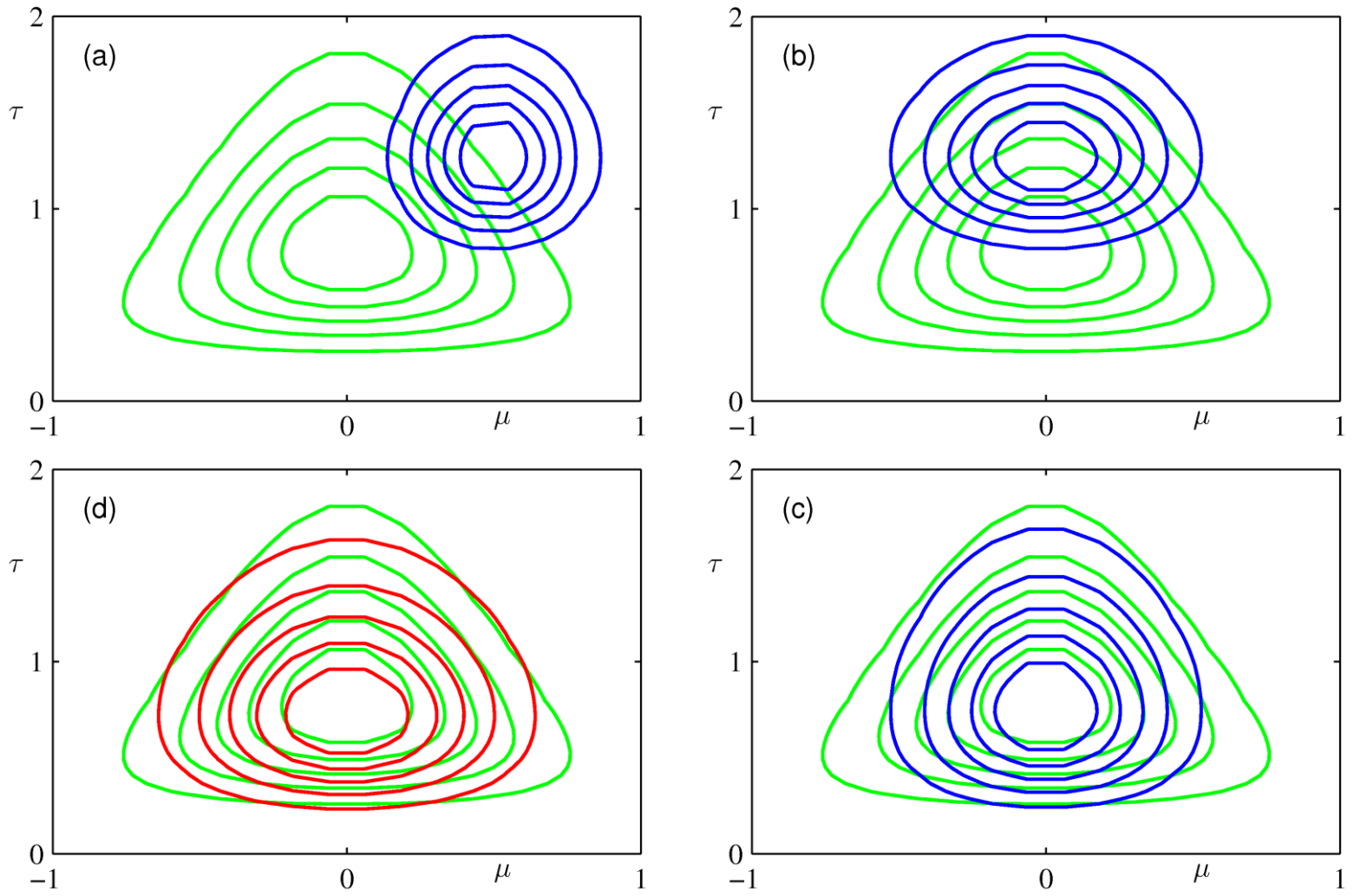
$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau) \quad (10.24)$$

$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau]$$

$$\mu_N = \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N} \quad (10.28)$$

$$a_N = a_0 + \frac{N+1}{2}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_\mu \left[\sum_{n=1}^{N} (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right]$$


(a)
(b)
(d)
(c)
τ
μ
2
1
0
−1
0
1