混合分布と教師なし学習
(復習)

正田 備也

masada@rikkyo.ac.jp

Contents

なぜこの復習をするか

- 混合分布モデルで、各データがどのコンポーネントに属するかを観測できない場合、つまり、教師なし学習の設定の場合、EM アルゴリズムを使っていた
- EM アルゴリズムの場合でモデルパラメータの更新式に至るまでの計算の流れが、変分ベイズ法の場合で事後分布のパラメータの更新式に至るまでの計算の流れに、そっくり！
  - 潜在変数の周辺化により得られる尤度の対数を取り、これに対して Jensen の不等式を使うことで、より最大化しやすい式を下界として得る、という計算の流れが、そっくり！

混合分布

- 観測データの集まりを、いくつかのまとまりに分けられそうな場合、混合分布をデータのモデリングに用いる
  - そのまとまりのことを、以下、「コンポーネント」と呼ぶ

- 各々の観測データ $\boldsymbol{x}_i$ が、どのコンポーネントに属するか、すでに分かっている場合は、教師あり学習をおこなう
  - これは、分類 (classification)
- 各々の観測データ $\boldsymbol{x}_i$ が、どのコンポーネントに属するか、不明な場合は、教師なし学習をおこなう
  - これは、クラスタリング (clustering) $\leftarrow$ 今回の設定はこちら

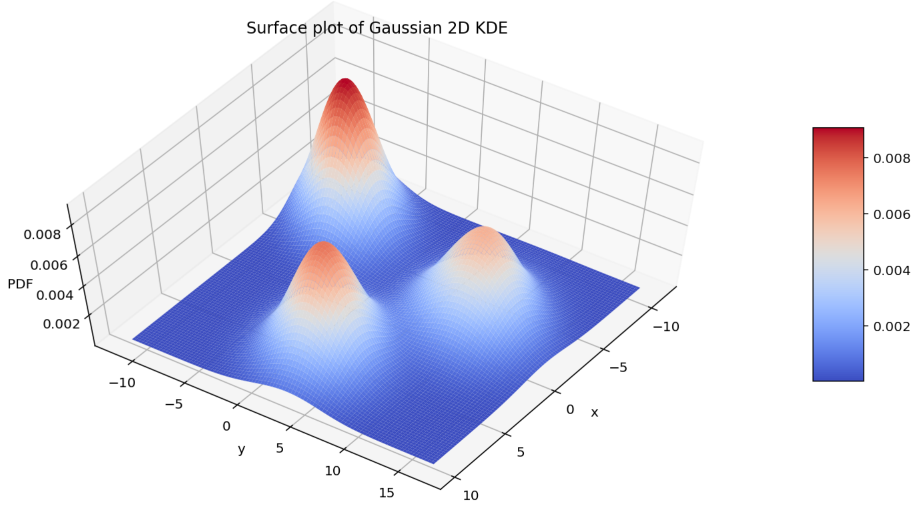

Figure: 2 次元ベクトルの集合をモデリングする混合分布の例

混合分布モデルを指定する

- 以下の2点を決めることにより混合分布モデルを指定できる

1. コンポーネントの個数 K を決める
   - k 番目のコンポーネントが選ばれる確率を θ_k と書くことにする
   - $(\theta_1, \ldots, \theta_K)$ を $\boldsymbol{\theta}$ と書くことにする（$\sum_k \theta_k = 1$ を満たす）
2. 各コンポーネントを表す分布を決める
   - 全てのコンポーネントは同じ確率分布（例：正規分布）で表され、パラメータの値が違うだけだとすることが多い
   - k 番目のコンポーネントを表す確率分布のパラメータを ϕ_k と書くことにする（例：単変量正規分布の場合は $\phi_k = (\mu_k, \sigma_k)$）
   - $\phi_1, \ldots, \phi_K$ をまとめて $\boldsymbol{\Phi}$ と書くことにする

混合分布によるデータの生成

- 混合分布を使ったモデリングでは、N 個の観測データ $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ が以下のように生成されると仮定する

1. カテゴリカル分布 $p(z; \boldsymbol{\theta})$ から、確率変数 z_i の値を draw
2. z_i 番目のコンポーネントを表す確率分布 $p(\boldsymbol{x}|z_i; \boldsymbol{\phi}_{z_i})$ から、$\boldsymbol{x}_i$ の値を draw

注： 確率モデルでデータをモデリングすることとは、データがどのように生成されるかを様々な確率分布を組み合わせて記述することである

混合分布モデルの同時分布

- 前のスライドで書いたデータの生成を、そのまま式にすると、同時分布の式になる
- i 番目の観測データ $\boldsymbol{x}_i$ がどのコンポーネントに属するかを表す確率変数を z_i とする
  - $z_i = k$ は $\boldsymbol{x}_i$ が k 番目のコンポーネントに属するという意味
- すると同時分布 $p(\mathcal{X}, \mathcal{Z})$ は

$$p(\mathcal{X}, \mathcal{Z}) = \prod_{i=1}^{N} p(z_i; \boldsymbol{\theta}) p(\boldsymbol{x}_i | z_i; \boldsymbol{\phi}_{z_i}) \tag{1}$$

  - ただし、$\mathcal{X} \equiv \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\mathcal{Z} \equiv \{z_1, \ldots, z_N\}$ と定義した

混合分布モデルの教師なし学習

- 各 z_i がその値の分からない確率変数、つまり潜在変数(latent variable) である場合、教師なし学習をおこなう
- 潜在変数 $\mathcal{Z} = \{z_1, \ldots, z_N\}$ を、周辺化 $\sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z})$ によって消去し、観測データの尤度 $p(\mathcal{X})$ を得る
- そしてデータの尤度 $p(\mathcal{X})$ を最大化する、という問題を解く
  - 通常は対数尤度 $\ln p(\mathcal{X})$ を最大化する
- この最大化問題を解くことで、$\boldsymbol{\theta}$ と $\boldsymbol{\Phi}$ を推定する

Contents

混合正規分布によるデータの生成

- 混合正規分布を使ったモデリングでは、N 個の観測データ $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ が以下のように生成されると仮定する

1. カテゴリカル分布 $p(z; \boldsymbol{\theta})$ から、確率変数 z_i の値を draw する
2. その z_i の値に対応する正規分布から、$\boldsymbol{x}_i$ の値を draw する
   - 第 k コンポーネントを表す正規分布のパラメータを $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ とする

$$
\begin{aligned}
z_i &\sim \mathsf{Cat}(\boldsymbol{\theta}) \\
\boldsymbol{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \qquad (2)
\end{aligned}
$$

単変量正規分布の混合分布の場合のパラメータ

- K 個のコンポーネントから一つを選ぶ際に使われるカテゴリカル分布のパラメータは $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$
  - θ_k は k 番目のコンポーネントが選ばれる確率（$\sum_{k=1}^{K} \theta_k = 1$）

$$p(z; \boldsymbol{\theta}) = \theta_z \tag{3}$$

- k 番目のコンポーネントを表す正規分布の平均パラメータを μ_k とし、標準偏差パラメータを σ_k とする

$$p(x|z; \mu_z, \sigma_z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{(x-\mu_z)^2}{2\sigma_z^2} \right) \tag{4}$$

単変量正規分布の混合分布の場合の同時分布

- $\mathcal{X} = \{x_1, \ldots, x_N\}$ と $\mathcal{Z} = \{z_1, \ldots, z_N\}$ との同時分布は

$$
\begin{aligned}
p(\mathcal{X}, \mathcal{Z}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) &= \prod_{i=1}^{N} p(z_i; \boldsymbol{\theta}) p(x_i | z_i; \mu_{z_i}, \sigma_{z_i}) \\
&= \prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(- \frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2} \right) \right]
\end{aligned}
\tag{5}
$$

- $p(z_i; \boldsymbol{\theta})$ はカテゴリカル分布 $\mathrm{Cat}(\boldsymbol{\theta})$ の pmf から求まる z_i の尤度
- $p(x_i | z_i; \mu_{z_i}, \sigma_{z_i})$ は正規分布 $\mathcal{N}(\mu_{z_i}, \sigma_{z_i})$ の pdf から求まる x_i の尤度

混合正規分布モデルの教師なし学習

- 各データ x_i がどのコンポーネントから生成されたか分からない場合、z_i は潜在変数 (latent variable) となる
- このとき、教師なし学習をおこなう
- 観測変数と潜在変数の同時分布 $p(\mathcal{X}, \mathcal{Z})$ は、式 (5) のとおり

$$
\begin{aligned}
p(\mathcal{X}, \mathcal{Z}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) &= \prod_{i=1}^{N} p(z_i; \boldsymbol{\theta}) p(x_i | z_i; \mu_{z_i}, \sigma_{z_i}) \\
&= \prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(- \frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2} \right) \right]
\end{aligned}
$$

観測データの尤度

- 潜在変数 $\mathcal{Z}$ を周辺化することによって、観測データの尤度 $p(\mathcal{X}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\})$ を得る
  - 周辺化＝潜在変数の値の全ての場合（K^N 通り）について和をとる

$$
\begin{aligned}
p(\mathcal{X}) &= \sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z}) \\
&= \sum_{z_1=1}^{K} \sum_{z_2=1}^{K} \cdots \sum_{z_{N-1}=1}^{K} \sum_{z_N=1}^{K} \prod_{i=1}^{N} p(x_i, z_i) \\
&= \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} p(x_i, z_i) \right) = \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} p(z_i) p(x_i | z_i) \right)
\end{aligned} \quad (6)
$$

観測データの対数尤度

- よって、観測データ $\mathcal{X}$ の対数尤度は、式 (5) と式 (6) より

$$
\begin{aligned}
\ln p(\mathcal{X}) &= \ln \sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z}) = \ln \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} p(x_i, z_i) \right) \\
&= \ln \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} p(z_i) p(x_i | z_i) \right) = \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} p(z_i) p(x_i | z_i) \right) \\
&= \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp \left(- \frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2} \right) \right] \right) \qquad (7)
\end{aligned}
$$

対数尤度の最大化によるパラメータ推定

- あとは、対数尤度 $\ln p(\mathcal{X}) = \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} p(z_i)p(x_i|z_i) \right)$ を最大にする $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ や $\mu_1, \ldots, \mu_K$ や $\sigma_1, \ldots, \sigma_K$ を求めれば良い・・・？？？

- 通常、式 (7) をそのまま最大化することはしない
- EM アルゴリズムを使う

積の対数と和の対数

- 何かを掛け算したものの対数は、何かの対数の和に書き直せるので、扱いやすい

$$\log(a \times b) = \log(a) + \log(b) \tag{8}$$

- 何かを足し算したものの対数は、それ以上変形のしようがないので、扱いにくい

$$\log(a + b) = \ldots \tag{9}$$

イェンセン Jensen の不等式（対数関数の場合）

- $p_1, \ldots, p_K$ を、$\sum_{k=1}^{K} p_k = 1$ を満たす正の実数とする
- $a_1, \ldots, a_K$ を任意の正の実数とする
- このとき、以下の不等式が成り立つ

$$\ln\left(\sum_{k=1}^{K} p_k a_k\right) \geq \sum_{k=1}^{K} p_k \ln(a_k) \tag{10}$$

- 和の対数（扱いにくい！）の下界 (lower bound) を、対数の和（扱いやすい！）として得るため、イェンセンの不等式をよく使う
- なお、対数関数に限らず、上に凸な関数なら、上の不等式は成立

対数尤度の下界

- イェンセンの不等式を利用して $\ln p(\mathcal{X})$ の下界を得たい
- そこで、各観測データ x_i について $\boldsymbol{q}_i \equiv (q_{i,1}, \ldots, q_{i,K})$ という $\sum_{k=1}^{K} q_{i,k} = 1$ を満たす潜在変数を用意すると、式 (7) より

$$\ln p(\mathcal{X}) = \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} p(z_i) p(x_i|z_i) \right)$$
$$= \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} q_{i,z_i} \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}} \right) \geq \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}} \tag{11}$$

  - この下界を $\mathcal{L}(\{\boldsymbol{q}_i\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\})$ と書くことにする

$$\ln p(\mathcal{X}) \geq \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}} \equiv \mathcal{L}(\{\boldsymbol{q}_i\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) \tag{12}$$

- この下界は以下のようにも書ける

$$\begin{aligned} &\mathcal{L}(\{\boldsymbol{q}_i\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) \\ &= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln p(z_i)p(x_i|z_i) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln q_{i,z_i} \end{aligned} \tag{13}$$

対数尤度の下界の最大化

- $\ln p(\mathcal{X})$ の代わりに $\mathcal{L}(\{\boldsymbol{q}_i\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\})$ を最大化することによって、次の未知量を推定する
  - 新たに導入した $\{\boldsymbol{q}_i\} \equiv \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N\}$ where $\boldsymbol{q}_i = (q_{i,1}, \ldots, q_{i,K})$
  - モデルパラメータ $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}\}$
- 単変量正規分布の混合分布の場合（cf. 式 (3)、式 (4)）

$$p(z_i = k) = \theta_k \tag{14}$$

$$p(x_i | z_i = k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(- \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right) \tag{15}$$

- これらを式 (13) に当てはめることで、以下、推定計算を行う

$$
\begin{aligned}
&L(\{\boldsymbol{q}_i\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) \\
&= \mathcal{L}(\{\boldsymbol{q}_i\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} q_{i,k}\right) + \lambda_0 \left(1 - \sum_{k=1}^{K} \theta_k\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \ln \frac{\theta_k p(x_i | z_i = k)}{q_{i,k}} + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} q_{i,k}\right) + \lambda_0 \left(1 - \sum_{k=1}^{K} \theta_k\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \ln \bigl(\theta_k p(x_i | z_i = k)\bigr) - \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \ln q_{i,k} + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} q_{i,k}\right) + \lambda_0 \left(1 - \sum_{k=1}^{K} \theta_k\right)
\end{aligned}
\tag{16}
$$

$$
\frac{\partial L}{\partial q_{i,k}} = \ln \bigl(\theta_k p(x_i | z_i = k)\bigr) - \ln q_{i,k} - 1 - \lambda_i \tag{17}
$$

$\frac{\partial L}{\partial q_{i,k}} = 0$ と $\sum_k q_{i,k} = 1$ より $q_{i,k} = \frac{\theta_k p(x_i|z_i=k)}{\sum_{k'} \theta_{k'} p(x_i|z_i=k')}$ を得る。

$q_{i,k} \ln \big(\theta_k p(x_i|z_i = k)\big) = q_{i,k} \ln \theta_k + q_{i,k} \ln p(x_i|z_i = k)$ より、

$$\frac{\partial L}{\partial \theta_k} = \frac{\sum_{i=1}^N q_{i,k}}{\theta_k} - \lambda_0 \tag{18}$$

$\frac{\partial L}{\partial \theta_k} = 0$ と $\sum_k \theta_k = 1$ より $\theta_k = \frac{\sum_{i=1}^N q_{i,k}}{\sum_{k=1}^K \sum_{i=1}^N q_{i,k}} = \frac{\sum_{i=1}^N q_{i,k}}{N}$ を得る。

$\frac{\partial}{\partial \mu_k} \ln p(x_i|z_i = k) = \frac{x_i - \mu_k}{\sigma_k^2}$ と $\frac{\partial}{\partial \sigma_k} \ln p(x_i|z_i = k) = -\frac{1}{\sigma_k} + \frac{(x_i - \mu_k)^2}{\sigma_k^3}$ より

$$\frac{\partial L}{\partial \mu_k} = \frac{\sum_{i=1}^N q_{i,k}(x_i - \mu_k)}{\sigma_k^2} \tag{19}$$

$$\frac{\partial L}{\partial \sigma_k} = \frac{\sum_{i=1}^N q_{i,k}(-\sigma_k^2 + (x_i - \mu_k)^2)}{\sigma_k^3} \tag{20}$$

$\frac{\partial L}{\partial \mu_k} = 0$ より $\mu_k = \frac{\sum_{i=1}^N q_{i,k} x_i}{\sum_{i=1}^N q_{i,k}}$ を得る。また、$\frac{\partial L}{\partial \sigma_k} = 0$ より $\sigma_k^2 = \frac{\sum_{i=1}^N q_{i,k}(x_i - \mu_k)^2}{\sum_{i=1}^N q_{i,k}}$ を得る。

混合正規分布のEMアルゴリズム

- E step

$$q_{i,k} \leftarrow \frac{\theta_k p(x_i|z_i = k)}{\sum_{k'} \theta_{k'} p(x_i|z_i = k')} \tag{21}$$

ただし $p(x_i|z_i = k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right)$

- M step

$$\theta_k \leftarrow \frac{\sum_{i=1}^N q_{i,k}}{N} \tag{22}$$

$$\mu_k \leftarrow \frac{\sum_{i=1}^N q_{i,k} x_i}{\sum_{i=1}^N q_{i,k}} \tag{23}$$

$$\sigma_k^2 \leftarrow \frac{\sum_{i=1}^N q_{i,k}(x_i - \mu_k)^2}{\sum_{i=1}^N q_{i,k}} \tag{24}$$

$q_{i,k}$ とは何なのか (1/4)

- イェンセンの不等式を使って、次の下界を得たのだった

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

- 左辺から右辺を引くと

$$\sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} p(z_i)p(x_i|z_i) \right) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}}$$

$$= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \left(\sum_{z_i=1}^{K} p(z_i)p(x_i|z_i) \right) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}}$$

$q_{i,k}$ とは何なのか (2/4)

（続き）

$$
\begin{aligned}
&= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \left(\sum_{z_i=1}^{K} p(x_i, z_i) \right) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}} \\
&= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln p(x_i) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(x_i, z_i)}{q_{i,z_i}} \\
&= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(x_i)q_{i,k}}{p(x_i, z_i)} = \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{q_{i,z_i}}{p(z_i|x_i)} \qquad (25)
\end{aligned}
$$

- $q_{i,k} = p(z_i = k|x_i)$ のとき、等号が成立する

カルバック・ライブラー情報量

- p, q を離散確率分布とすると、q から p への（p の q に対する）カルバック・ライブラー情報量 Kullback–Leibler divergence とは

$$D_{\mathrm{KL}}(p \parallel q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} \tag{26}$$

- p, q が連続確率分布の場合は

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \ln \left(\frac{p(x)}{q(x)} \right) dx \tag{27}$$

- $p = q$ ならば、またそのときに限り $D_{\mathrm{KL}}(p \parallel q) = 0$

注. $q(x) = 0$ なのに $p(x) \neq 0$ となる x があってはいけない！

$q_{i,k}$ とは何なのか (3/4)

- $q_i(z_i)$ を、K 個のアイテム $\{1,\dots,K\}$ 上に定義されたカテゴリカル分布とし、$q_i(z_i=k)\equiv q_{i,k}$ と設定する
- 式 (25) は $q_i(z_i)$ の $p(z_i|x_i)$ に対するカルバック・ライブラー情報量になっている
- ところで、式 (21) より、

$$
\begin{aligned}
q_{i,k} &= \frac{\theta_k p(x_i|z_i=k)}{\sum_{k'}\theta_{k'}p(x_i|z_i=k')} = \frac{p(z_i=k)p(x_i|z_i=k)}{\sum_{k'}p(z_i=k')p(x_i|z_i=k')} \\
&= \frac{p(x_i,z_i=k)}{\sum_{k'}p(x_i,z_i=k')} = \frac{p(x_i,z_i=k)}{p(x_i)} = p(z_i=k|x_i) \qquad (28)
\end{aligned}
$$

$q_{i,k}$ とは何なのか (4/4)

- つまり、式 (21) は $q_{i,k} = p(z_i = k|x_i)$ を意味している
- このとき、式 (25) のカルバック・ライブラー情報量はゼロ！
- ということは、E ステップで得られる $q_{i,k}$ は、最善の答え
- ただし、この $q_{i,k}$ は、パラメータ $\boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}$ の値を特定の値に固定した上で、$\ln p(x_i)$ の下界を最大化して求めたもの
- M ステップでは、逆に $\{\boldsymbol{q}_i\}$ のほうを固定し、$\ln p(x_i)$ の下界を最大化している

E step で何をしているか

- モデルのパラメータ $\mathbf{\Theta} = \{\boldsymbol{\theta}, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K\}$ の値を固定した状態で、対数尤度の下界を最大化する $\{\boldsymbol{q}_i\}$ を求めているのが、E step
- パラメータ $\mathbf{\Theta}$ の、固定された値を、$\mathbf{\Theta}_{\mathsf{old}}$ と書くことにする
- その最大化によって得られる答えは

$$q_{i,k} = p(z_i = k | x_i; \mathbf{\Theta}_{\mathsf{old}}) \tag{29}$$

- つまり、モデルパラメータ $\mathbf{\Theta}$ の値を固定したうえで、観測データを所与とする潜在変数の条件付き分布を求めている

ここまでの議論のパターン

- 確率モデルが潜在変数 $\mathcal{Z}$ を含む
- 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 $p(\mathcal{X}, \mathcal{Z})$ の式を得る
- 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 $p(\mathcal{X})$ が得られるが、実際には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
- この下界を最大化することで、様々な未知量を推定する

Contents

ベイズ的モデリングにおける変分法

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

$$p(\mathbf{\Theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathcal{X})} \tag{30}$$

- 変分ベイズ法は $p(\mathbf{\Theta}|\mathcal{X})$ を近似する分布 $q(\mathbf{\Theta})$ を求める
  - $q(\mathbf{\Theta})$ を変分法 (variational methods) で求める（後述）
  - $q(\mathbf{\Theta})$ を変分事後分布 (variational posterior distribution) と呼ぶ

先ほどまでの議論のパターンを適用

- 確率モデルが $\boldsymbol{\Theta}$ という潜在変数を含む
  - ベイズの枠組みの中では、モデルパラメータ $\boldsymbol{\Theta}$ は確率変数だから、$\boldsymbol{\Theta}$ はその値が見えていない確率変数、つまり、潜在変数になる
- ベイズ的な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 $p(\mathcal{X}, \boldsymbol{\Theta})$ の式を得る
- 潜在変数 $\boldsymbol{\Theta}$ の周辺化 $\int p(\mathcal{X}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}$ により観測データの尤度 $p(\mathcal{X})$ が得られるが、実際には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
- この下界を最大化することで、様々な未知量を推定する

変分ベイズ法(variational Bayesian methods)

- 潜在変数を含むモデルの対数尤度最大化の話と似ている
- その値が見えている確率変数 $\mathcal{X}$ と、その値が見えていない確率変数 $\boldsymbol{\Theta}$ とがある
- $p(\mathcal{X}, \boldsymbol{\Theta})$ は式で書けるが、$p(\mathcal{X}) = \int p(\mathcal{X}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}$ は書けない
- $\ln p(\mathcal{X})$ の代わりに、$\ln p(\mathcal{X})$ の下界を最大化する
- $\ln p(\mathcal{X})$ の下界は Jensen の不等式を使って求める

$$\ln p(\mathcal{X}) = \ln \int q(\boldsymbol{\Theta}) \frac{p(\mathcal{X}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} \geq \int q(\boldsymbol{\Theta}) \ln \frac{p(\mathcal{X}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta}$$

- この $q(\boldsymbol{\Theta})$ が変分事後分布

「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
- $q(\mathbf{\Theta})$ の密度関数がどんなかたちを持つかに制約を設けない
- 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だけ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
- とても直感的に言うと、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）