混合分布

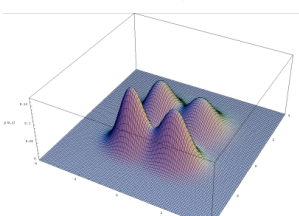
正田 備也

masada@rikkyo.ac.jp

Contents

これまでのモデリングの問題点

- これまでは、データ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ 全体に対して、一つの確率分布を使うモデリングだけ議論していた
- しかし、多くのデータ集合は、たった一つの分布ではモデリングし切れない多様性を含んでいる
- 例えば、数値データの集合であれば、それに近い数値が頻繁に出現するという数値が、複数あったりする
  - 例：多峰性をもつデータ集合



混合分布

- これまでは、全てのデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ が同じ分布から生成されると仮定していた
- 一方、混合分布によるモデリングでは、同じ種類の分布だがパラメータの値が違う分布を K 個用意する
  - これらの分布をコンポーネントと呼ぶことがある
- そして、各データ $\boldsymbol{x}_i$ について…

1. カテゴリカル分布 $\mathsf{Cat}(\boldsymbol{\theta})$ に従って K 個の分布から一つ選ぶ
   - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ はパラメータで、θ_k は k 番目の分布が選ばれる確率。もちろん $\sum_k \theta_k = 1$ が成り立つ
2. そして、$\boldsymbol{x}_i$ がその選ばれた分布に従うと考える。

Contents

混合正規分布

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

1. i 番目のデータ $\boldsymbol{x}_i$ を生成するため、まず、カテゴリカル分布 $\mathrm{Cat}(\boldsymbol{\theta})$ から、確率変数 z_i の値を draw する
   - 「$z_i = k$」は、i 番目のデータについては k 番目のコンポーネントが選ばれたことを意味する
2. その z_i の値に対応する確率分布から、$\boldsymbol{x}_i$ を draw する

$$
\begin{aligned}
z_i &\sim \mathrm{Cat}(\boldsymbol{\theta}) \\
\boldsymbol{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})
\end{aligned}
\tag{1}
$$

単変量正規分布の混合分布の場合

- K 個のコンポーネントのなか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
  - $\sum_{k=1}^{K} \theta_k = 1$ が成り立つ
- K 個の単変量正規分布をコンポーネントとして用意する
- k 番目の分布のパラメータは、平均 μ_k と標準偏差 σ_k
  - k 番目のコンポーネントの確率密度関数は

$$p(x; \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2} \right) \tag{2}$$

単変量正規分布の混合分布における同時分布

- 単変量正規分布の混合分布を使うと、観測されたデータを表す確率変数 $\mathcal{X} = \{x_1, \ldots, x_N\}$ とコンポーネントへの所属を表す確率変数 $\mathcal{Z} = \{z_1, \ldots, z_N\}$ との同時分布は

$$p(\mathcal{X}, \mathcal{Z}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) = \prod_{i=1}^{N} p(z_i; \theta_{z_i}) p(x_i | z_i; \mu_{z_i}, \sigma_{z_i})$$

$$= \prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(- \frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \tag{3}$$

  - $p(z_i)$ はカテゴリカル分布 $\mathrm{Cat}(\boldsymbol{\theta})$ の pmf
  - $p(x_i|z_i)$ は z_i が表すコンポーネントの正規分布 $\mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$ の pdf

Contents

教師ありの設定の場合 (1/2)

- 教師ありの設定の場合、各データ x_i について、それがどのコンポーネントから生成されたかは、すでに分かっている
- 言い換えれば、z_i の値も観測データに含まれる
  - つまり、観測データを $\mathcal{D}$ で表すとすると、
    $\mathcal{D} = \{(x_1, z_1), \ldots, (x_N, z_N)\}$ （x_i も z_i も、両方見えている）
- このとき、式 (3) の同時分布が、そのまま、観測データ $\mathcal{D}$ の尤度を表すことになる

教師ありの設定の場合 (2/2)

- そして、観測データ $\mathcal{D} = \{(x_1, z_1), \ldots, (x_N, z_N)\}$ の尤度は、下のように書き直すことができる

$$
\begin{aligned}
p(\mathcal{D}; \boldsymbol{\theta}, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K) &= p(\mathcal{D}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) \\
&= \prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(- \frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \\
&= \prod_{k=1}^{K} \left[\theta_k^{c_k} \times \frac{1}{(\sqrt{2\pi\sigma_k^2})^{c_k}} \exp\left(- \frac{\sum_{\{i:z_i=k\}} (x_i - \mu_k)^2}{2\sigma_k^2} \right) \right] \quad (4)
\end{aligned}
$$

  - c_k は、k 番目のコンポーネントから生成されたデータの個数

教師ありの場合の混合正規分布の最尤推定

- 混合正規分布のパラメータを、式 (4) の観測データの尤度を最大化することによって推定する方法を、以下に示す

目的関数は

$$L(\boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) = \ln p(\mathcal{D}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) + \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right)$$

$$= \sum_{k=1}^{K} c_k \ln \theta_k - \sum_{k=1}^{K} c_k \ln \sigma_k - \sum_{k=1}^{K} \frac{\sum_{\{i:z_i=k\}} (x_i - \mu_k)^2}{2\sigma_k^2} + \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right) + const. \quad (5)$$

目的関数 L を、各パラメータで偏微分する。

$$\frac{\partial L}{\partial \mu_k} = \frac{\sum_{\{i:z_i=k\}} (x_i - \mu_k)}{\sigma_k^2} = \frac{\sum_{\{i:z_i=k\}} x_i - c_k \mu_k}{\sigma_k^2} \quad (6)$$

$\frac{\partial L}{\partial \mu_k} = 0$ より、$\mu_k = \frac{\sum_{\{i:z_i=k\}} x_i}{c_k} = \bar{x}_k$ を得る。

$$\frac{\partial L}{\partial \sigma_k} = -\frac{c_k}{\sigma_k} + \frac{\sum_{\{i:z_i=k\}} (x_i - \mu_k)^2}{\sigma_k^3} \quad (7)$$

$\frac{\partial L}{\partial \sigma_k} = 0$ より、$\sigma_k^2 = \frac{\sum_{\{i:z_i=k\}} (x_i - \bar{x}_k)^2}{\sigma_k^2}$ を得る。

$$\frac{\partial L}{\partial \theta_k} = \frac{c_k}{\theta_k} - \lambda \ , \ \frac{\partial L}{\partial \lambda} = 1 - \sum_{k=1}^{K} \theta_k \tag{8}$$

$\frac{\partial L}{\partial \theta_k} = 0$ より、$\theta_k = \frac{c_k}{\lambda}$ を得る。
$\frac{\partial L}{\partial \lambda} = 0$ より、$1 - \sum_{k=1}^{K} \frac{c_k}{\lambda} = 0$ を得る。
つまり、$\lambda = \sum_k c_k$ が言えるので、$\theta_k = \frac{c_k}{\sum_k c_k}$ を得る。

まとめると、

- θ_k は、k 番目のコンポーネントから生成されたデータの割合となる。
- μ_k と σ_k は、k 番目のコンポーネントから生成されたデータだけの尤度をもとに最尤推定した値となる。

Contents

教師なしの設定の場合

- 教師なしの設定の場合、各データ x_i について、それがどのコンポーネントから生成されたかは、分からない！
- z_i は、値が観測されない確率変数、すなわち潜在変数
  - つまり、$\mathcal{X} = \{x_1, \ldots, x_N\}$ だけが観測変数
  - 一方、潜在変数 latent variables の集合を $\mathcal{Z} = \{z_1, \ldots, z_N\}$ とする
- 観測変数と潜在変数の同時分布 $p(\mathcal{X}, \mathcal{Z})$ は

$$p(\mathcal{X}, \mathcal{Z}; \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) = \prod_{i=1}^{N} p(x_i, z_i; \theta_{z_i}, \mu_{z_i}, \sigma_{z_i})$$

$$= \prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \quad (9)$$

観測データの尤度

- 潜在変数を含むモデリングの場合、観測データの尤度 $p(\mathcal{X})$ は、潜在変数 $\mathcal{Z}$ を周辺化して得られる
  - $\mathcal{X}$ を不完全データ incomplete data と呼ぶことがある
    - 潜在変数を周辺化する＝潜在変数がとりうる値の全てを考慮する

$$
\begin{aligned}
p(\mathcal{X}) &= \sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z}) = \sum_{z_1=1}^{K} \sum_{z_2=1}^{K} \cdots \sum_{z_{N-1}=1}^{K} \sum_{z_N=1}^{K} p(\mathcal{X}, \mathcal{Z}) \\
&= \sum_{z_1=1}^{K} \sum_{z_2=1}^{K} \cdots \sum_{z_{N-1}=1}^{K} \sum_{z_N=1}^{K} \prod_{i=1}^{N} p(x_i, z_i) \\
&= \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} p(x_i, z_i) \right)
\end{aligned}
\tag{10}
$$

$N = 5, K = 3$ とすると

$$
\begin{aligned}
&\prod_{i=1}^{N} \Big(\sum_{z_i=1}^{K} p(x_i, z_i) \Big) \\
&= \big(p(x_1, z_1 = 1) + p(x_1, z_1 = 2) + p(x_1, z_1 = 3) \big) \\
&\quad \times \big(p(x_2, z_1 = 1) + p(x_2, z_1 = 2) + p(x_2, z_1 = 3) \big) \\
&\quad \times \big(p(x_3, z_1 = 1) + p(x_3, z_1 = 2) + p(x_3, z_1 = 3) \big) \\
&\quad \times \big(p(x_4, z_1 = 1) + p(x_4, z_1 = 2) + p(x_4, z_1 = 3) \big) \\
&\quad \times \big(p(x_5, z_1 = 1) + p(x_5, z_1 = 2) + p(x_5, z_1 = 3) \big)
\end{aligned} \tag{11}
$$

括弧を外して展開すると、足し合わされる項の数は K^N 個。
K がそれほど大きな値でないとしても、データのサイズ N は、通常、大きな値なので…

観測データの対数尤度

- 混合正規分布の場合の観測データの対数尤度は、式 (3) より

$$
\begin{aligned}
\ln p(\mathcal{X}) &= \ln \sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z}) \\
&= \ln \sum_{z_1=1}^{K} \cdots \sum_{z_N=1}^{K} \left(\prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \right) \\
&= \ln \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \right) \\
&= \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \right) \qquad (12)
\end{aligned}
$$

観測データの対数尤度の最大化？

- 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・・・？？？

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

$$\log(a \times b) = \log(a) + \log(b) \tag{13}$$

- 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

$$\log(a + b) = \ldots \tag{14}$$

イェンセンの不等式（対数関数の場合）

- $p_1, \ldots, p_K$ を、$\sum_k p_k = 1$ を満たす正の実数とする
- また、$x_1, \ldots, x_K$ を正の実数とする
- このとき、以下の不等式が成り立つ

$$\ln\left(\sum_{k=1}^{K} p_k x_k\right) \geq \sum_{k=1}^{K} p_k \ln(x_k) \tag{15}$$

  - 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

観測データの対数尤度の下界

- 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 x_i について、$q_{i,1}, \ldots, q_{i,K}$ という $\sum_k q_{i,k} = 1$ を満たす変数を用意する
  - $q_{i,k}$ はデータ x_i が k 番目のコンポーネントに属する確率を意味する

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

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

- この下界を $\ell(\{q_{i,k}\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\})$ と書くことにする (16)

観測データの対数尤度の下界の最大化

- そして、$\ln p(\mathcal{X})$ の代わりに $\ell(\{q_{i,k}\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\})$ を最大化することで、パラメータを推定する
- 単変量正規分布の混合分布の場合

$$p(z_i = k) = \theta_k \tag{17}$$

$$p(x_i | z_i = k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(- \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right) \tag{18}$$

- 以下、推定計算を行う

$$
\begin{aligned}
&L(\{q_{i,k}\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) \\
&= \ell(\{q_{i,k}\}, \boldsymbol{\theta}, \{\mu_k\}, \{\sigma_k\}) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} q_{i,k}\right) + \lambda_0 \left(1 - \sum_{k=1}^{K} \theta_k\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \ln \frac{\theta_k p(x_i | z_i = k)}{q_{i,k}} + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} q_{i,k}\right) + \lambda_0 \left(1 - \sum_{k=1}^{K} \theta_k\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \ln \big(\theta_k p(x_i | z_i = k)\big) - \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \ln q_{i,k} + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} q_{i,k}\right) + \lambda_0 \left(1 - \sum_{k=1}^{K} \theta_k\right)
\end{aligned}
\tag{19}
$$

$$
\frac{\partial L}{\partial q_{i,k}} = \ln \big(\theta_k p(x_i | z_i = k)\big) - \ln q_{i,k} - 1 - \lambda_i \tag{20}
$$

$\frac{\partial L}{\partial q_{i,k}} = 0$ 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。

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

$$\frac{\partial L}{\partial \theta_k} = \frac{\sum_{i=1}^{N} q_{i,k}}{\theta_k} - \lambda_0 \tag{21}$$

$\frac{\partial L}{\partial \theta_k} = 0$ 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

$$\frac{\partial L}{\partial \mu_k} = \frac{\sum_{i=1}^{N} q_{i,k}(x_i - \mu_k)}{\sigma_k^2} \tag{22}$$

$$\frac{\partial L}{\partial \sigma_k} = \frac{\sum_{i=1}^{N} q_{i,k}(-\sigma_k^2 + (x_i - \mu_k)^2)}{\sigma_k^3} \tag{23}$$

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

$q_{i,k}$ とは何なのか

- イェンセンの不等式を使って、次の下界を得たのだった

$$\sum_{i=1}^{N} \ln \Big(\sum_{z_i=1}^{K} p(z_i) p(x_i|z_i) \Big) \geq \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,k} \ln \frac{p(z_i) p(x_i|z_i)}{q_{i,k}} \tag{24}$$

- 左辺から右辺を引いた差をとる

$$\begin{aligned}
&\sum_{i=1}^{N} \ln \Big(\sum_{z_i=1}^{K} p(z_i) p(x_i|z_i) \Big) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,k} \ln \frac{p(z_i) p(x_i|z_i)}{q_{i,k}} \\
&= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,k} \ln \Big(\sum_{z_i=1}^{K} p(z_i) p(x_i|z_i) \Big) - \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,k} \ln \frac{p(z_i) p(x_i|z_i)}{q_{i,k}} \\
&= \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,k} \ln \frac{p(x_i) q_{i,k}}{p(x_i, z_i)} = \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,k} \ln \frac{q_{i,k}}{p(z_i|x_i)} = D_{\mathsf{KL}}(q_{i,k} \parallel p(z_i|x_i))
\end{aligned} \tag{25}$$

- 差 $D_{\mathsf{KL}}(q_{i,k} \parallel p(z_i|x_i))$ は、$p(z_i|x_i)$ から $q_{i,k}$ への KL ダイバージェンス
- $q_{i,k} = p(z_i|x_i)$ のとき等号成立（$q_{i,k}$ は $p(z_i|x_i)$ を近似しているとも言える）