混合分布

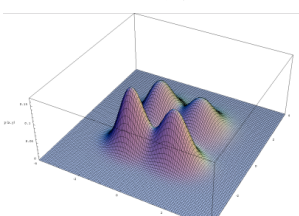
正田 備也

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、その周辺の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 $\boldsymbol{x}_i$ for $i = 1, \ldots, N$ を、同じ一つの分布から draw していた
  - 全ての確率変数 $\boldsymbol{x}_i$ for $i = 1, \ldots, N$ が同じ分布に従うと考えていた
- 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だけの分布を、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、まず、カテゴリカル分布 $\mathrm{Cat}(\boldsymbol{\theta})$ に従って K 個のコンポーネントから一つ選ぶ
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ 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
- 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によるモデリングでは、データ集合 $\{x_1, \ldots, x_N\}$ と、コンポーネントへの所属を表す情報 $\{z_1, \ldots, z_N\}$ との、同時分布が、以下のように得られる

$$p(\{x_1, \ldots, x_N\}, \{z_1, \ldots, z_N\}; \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{i=1}^{N} p(x_i, z_i; \theta_{z_i}, \sigma_{z_i})$$

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

  - 個々のデータは、（同じ分布からではないにせよ）独立に生成されると仮定している

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

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

  - 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

$$L(\boldsymbol{\theta}, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K) = \ln p(\mathcal{D}; \boldsymbol{\theta}, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K) + \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right)$$

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

目的関数 L を、各パラメータで偏微分する。

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

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

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

$\frac{\partial L}{\partial \sigma_k} = 0$ 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、$\mathcal{D} = \{x_1, \ldots, x_N\}$
  - 一方、潜在変数の集合を $\mathcal{Z} = \{z_1, \ldots, z_N\}$ とする
- このとき、下の $p(\mathcal{D}, \mathcal{Z})$ は、観測データの尤度 $p(\mathcal{D})$ ではない

$$p(\mathcal{D}, \mathcal{Z}; \boldsymbol{\theta}, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K) = \prod_{i=1}^{N} p(x_i, z_i; \theta_{z_i}, \sigma_{z_i})$$

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

周辺尤度 marginal likelihood

- 潜在変数を含むモデリングの場合、観測データの尤度 $p(\mathcal{D})$ は、潜在変数を周辺化 marginalize してはじめて得られる
- 周辺化が与える尤度を周辺尤度 marginal likelihood と呼ぶ

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

対数周辺尤度 log marginal likelihood

- 周辺尤度の対数は、式 (3) より、以下のように書ける

$$
\begin{aligned}
\ln p(\mathcal{D}) &= \ln \sum_{\mathcal{Z}} p(\mathcal{D}, \mathcal{Z}) \\
&= \ln \sum_{z_1=1}^{K} \cdots \sum_{z_N=1}^{K} \left(\prod_{i=1}^{N} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \right) \\
&= \ln \prod_{i=1}^{N} \left(\sum_{z_i=1}^{K} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \right) \\
&= \sum_{i=1}^{N} \ln \left(\sum_{z_i=1}^{K} \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp\left(-\frac{(x_i - \mu_{z_i})}{2\sigma_{z_i}^2} \right) \right] \right) \qquad (11)
\end{aligned}
$$

積の対数と和の対数（天国と地獄？）

- 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

$$\log(a \times b) = \log(a) + \log(b) \tag{12}$$

- しかし、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

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