

混合分布

正田 備也

masada@rikkyo.ac.jp

Contents

なぜ混合分布を使うのか

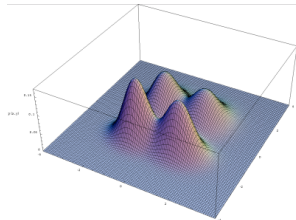
混合正規分布

混合正規分布：教師ありの場合

混合正規分布：教師なしの場合

これまでのモデリングの問題点

- ▶ これまでは、データ集合 $D = \{x_1, \dots, x_N\}$ 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 x_i for $i = 1, \dots, N$ を、同じ一つの分布から draw していた
 - ▶ 全ての確率変数 x_i for $i = 1, \dots, N$ 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 x_i について、まず、カテゴリカル分布 $\text{Cat}(\theta)$ に従って K 個のコンポーネントから一つ選ぶ
 - ▶ $\theta = (\theta_1, \dots, \theta_K)$ 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、 x_i がその選ばれた分布に従うと考える。

Contents

なぜ混合分布を使うのか

混合正規分布

混合正規分布：教師ありの場合

混合正規分布：教師なしの場合

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 $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 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 \mathbf{x}_i を生成するため、まず、カテゴリカル分布 $\text{Cat}(\boldsymbol{\theta})$ 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、 \mathbf{x}_i を draw する

$$z_i \sim \text{Cat}(\boldsymbol{\theta})$$

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \quad (1)$$

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 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$
 - ▶ 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

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

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 $\{x_1, \dots, x_N\}$ と、コンポーネントへの所属を表す情報 $\{z_1, \dots, z_N\}$ との、同時分布が、以下のように得られる

$$\begin{aligned} p(\{x_1, \dots, x_N\}, \{z_1, \dots, z_N\}; \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) &= \prod_{i=1}^N p(x_i, z_i; \theta_{z_i}, \sigma_{z_i}) \\ &= \prod_{i=1}^N \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp \left(-\frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2} \right) \right] \end{aligned} \quad (3)$$

- ▶ 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

なぜ混合分布を使うのか

混合正規分布

混合正規分布：教師ありの場合

混合正規分布：教師なしの場合

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、
$$\mathcal{D} = \{(x_1, z_1), \dots, (x_N, z_N)\} \quad (x_i \text{ も } z_i \text{ も、両方見えている})$$
- ▶ 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 $\mathcal{D} = \{(x_1, z_1), \dots, (x_N, z_N)\}$ の尤度は、下のように書き直すことができる

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

- ▶ 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

$$\begin{aligned} L(\boldsymbol{\theta}, \mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K) &= \ln p(\mathcal{D}; \boldsymbol{\theta}, \mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K) + \lambda \left(1 - \sum_{k=1}^K \theta_k \right) \\ &= \sum_{k=1}^K c_k \ln \theta_k - \sum_{k=1}^K c_k \ln \sigma_k - \sum_{k=1}^K \frac{\sum_{\{i: z_i=k\}} (x_i - \mu_k)^2}{2\sigma_k^2} + \lambda \left(1 - \sum_{k=1}^K \theta_k \right) + \text{const.} \end{aligned} \quad (5)$$

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

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

$\frac{\partial L}{\partial \theta_k} = 0$ 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、 $\lambda = \sum_k c_k$ と言えるので、 $\theta_k = \frac{c_k}{\sum_k c_k}$ 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

なぜ混合分布を使うのか

混合正規分布

混合正規分布：教師ありの場合

混合正規分布：教師なしの場合

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、 $\mathcal{D} = \{x_1, \dots, x_N\}$
 - ▶ 一方、潜在変数 latent variables の集合を $\mathcal{Z} = \{z_1, \dots, z_N\}$ 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

$$\begin{aligned} p(\mathcal{D}, \mathcal{Z}; \boldsymbol{\theta}, \mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K) &= \prod_{i=1}^N p(x_i, z_i; \theta_{z_i}, \sigma_{z_i}) \\ &= \prod_{i=1}^N \left[\theta_{z_i} \times \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \exp \left(-\frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2} \right) \right] \end{aligned} \tag{9}$$

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 $p(\mathcal{D})$ は、潜在変数を周辺化して得られる
 - ▶ \mathcal{D} を不完全データ incomplete data と呼ぶことがある

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

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

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

観測データの対数尤度の最大化

- ▶ あとは、 $\ln p(\mathcal{D})$ を最大にする $\theta = (\theta_1, \dots, \theta_K)$ や μ_1, \dots, μ_K や $\sigma_1, \dots, \sigma_K$ を求めれば良い・・・ ???
- ▶ （偏微分の計算をしようとしてみると、問題がそう簡単ではなさそうであることに気づかれると思います。）

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

$$\log(a \times b) = \log(a) + \log(b) \quad (12)$$

- ▶ 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

$$\log(a + b) = \dots \quad (13)$$

イェンセンの不等式（対数関数の場合）

- ▶ p_1, \dots, p_K を、 $\sum_k p_k = 1$ を満たす正の実数とする
- ▶ また、 x_1, \dots, 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

$$\ln \left(\sum_{k=1}^K p_k x_k \right) \geq \sum_{k=1}^K p_k \ln(x_k) \quad (14)$$

- ▶ 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