

離散-連続混合型分布・条件付き分布・事前分布・事後分布

樋口さぶろお <https://hig3.net>

龍谷大学大学院 先端理工学部研究科 数理・情報科学コース

計算科学特論 L02(2024-10-02 Wed)

最終更新: Time-stamp: "2024-10-02 Wed 12:05 JST hig"

今日の目標

- 離散-連続型の多次元確率分布を説明できる
- 条件付き確率密度を説明できる
- ベイズ推定を説明できる



ここまで来たよ

① 多次元の確率分布

② 離散-連続混合型分布・条件付き分布・事前分布・事後分布

- 離散-連続混合型分布
- 条件付き分布

離散-連続混合型

Hoff バイズ p.28

確率変数の組 (X, Y) を考える.

X : 連続型確率変数,

Y : 離散型確率変数, $Y = y_1, \dots, y_n$.

次を満たす確率密度関数 $f: \mathbb{R} \times \mathbb{Z} \rightarrow \mathbb{R}_{\geq 0}$ で指定される ($\mathbb{Z}$ は「例えば」).

定義 (離散-連続混合型分布)

$$P(a < X \leq b, y_i \leq Y \leq y_j) = \int_a^b \sum_{y_i \leq y' \leq y_j} f(x', y') \, dx'$$

周辺分布 (連続型 $f_X(x)$, 離散型 $f_Y(y)$) .

累積分布関数.

$$F(x, y) := \int_{-\infty}^x \sum_{y'=-\infty}^y f(x', y') dx'.$$

気分を出すために, 表や場合分けで $f(x, y)$.

ただし, $Y = 0, 1$.

x についての確率密度関数 $h_0(x), h_1(x)$.

$$f(x, y) = \begin{cases} h_0(x) & (y = 0) \\ h_1(x) & (y = 1) \\ 0 & (y \neq 0, 1) \end{cases}$$

$y \backslash x$	$-\infty < x < +\infty$
0	$h_0(x)$
1	$h_1(x)$

$$\text{全} = \int_{-\infty}^{+\infty} h_0(x) dx + \int_{-\infty}^{+\infty} h_1(x) dx = 1.$$

混合ガウス分布 GMM=Gaussian mixture model

正規分布 normal distribution = ガウス分布 Gaussian distribution

岩薩林 確率・統計 §4.5

混合する = mix

混合ガウス分布を定義する準備として、次の離散-連続混合型分布 $f(x, y)$ を考える。

$\pi_0 + \pi_1 = 1, \pi_y \geq 0, \pi_{\text{他}} = 0$ とする。

$$f(x, y) = \pi_y \frac{1}{(2\pi\sigma_y^2)^{1/2}} e^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} = \begin{cases} \pi_0 \cdot \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} & (y = 0) \\ \pi_1 \cdot \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} & (y = 1) \\ 0 & (y \neq 0, 1) \end{cases}$$

$$= \delta_{y0} \pi_0 \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} + \delta_{y1} \pi_1 \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}.$$

Y の周辺分布は離散型でベルヌイ分布 $B(1, \pi_1)$

$$f_Y(y) = \int f(x, y) dx = \int_{-\infty}^{+\infty} \pi_y \cdot \frac{1}{(2\pi\sigma_y^2)^{1/2}} e^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} dx = \pi_y.$$

X の周辺分布は連続型で GMM

定義 (GMM=混合ガウス分布)

連続型確率変数 X がパラメタ $(\pi_0, \pi_1, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$ の GMM に従うとは、次の確率密度関数にしたがうことをいう。

$$f_X(x) = \sum_y f(x, y) = \pi_0 \cdot \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} + \pi_1 \cdot \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}.$$

L02-Q1

Quiz(混合ガウス分布の確率密度関数)

X は混合ガウス分布 ($\pi_0 = \frac{3}{10}, \pi_1 = \frac{7}{10}, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = \frac{1}{2}$) にしたがう. 確率密度関数 $f(x)$ のグラフの概形を描こう.
手で, または, `scipy.stats.norm` の浮動小数点計算で.

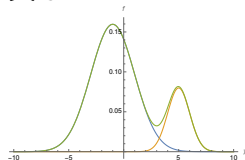
一般の混合ガウス分布: y の値が 2 個と限らず有限個

`scipy.stats` には GMM はないので, `scipy.stats.norm` を使って自分で計算する.

`sklearn.mixture.GaussianMixture`

一般の混合分布: ガウス分布とは限らない分布を重み π_y で「重ね合わせ」

注: 混合 $\pi_0 h_0(x) + \pi_1 h_1(x)$ と, 確率変数の和 $S = X_1 + X_2$ はもちろん異なる.



L02-Q2

Quiz(混合ガウス分布の確率)

X は混合ガウス分布 ($\pi_0 = \frac{3}{10}, \pi_1 = \frac{7}{10}, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = \frac{1}{2}$) にしたがう.

- ① X の確率密度関数の $x = 4$ における値を求めよう.
- ② 確率 $P(X \leq 4)$ を, 標準正規分布の累積分布関数 $\Phi(x)$ で表そう.

さらに `scipy.stats.norm` で浮動小数点数で求めてみると安心できるかも.

ここまで来たよ

① 多次元の確率分布

② 離散-連続混合型分布・条件付き分布・事前分布・事後分布

- 離散-連続混合型分布
- 条件付き分布

条件付き確率

岩薩林 確率・統計 §2.3

定義 (条件付き確率)

岩薩林 確率・統計 (2.12)

条件 (事象) A のもとでの, 事象 B の条件付き確率

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

縦棒「|」=given」の前後は不平等: P (確率を考える事象|条件の事象)
 $P(B|A) \neq P(A|B), \neq P(A \cup B)$.

意味 A が起きる場合に限定した, B が起きる確率

$P(B|A)$ を P (of) B given A と発音する人がいる. given 条件=条件が与えられたとき, という言い方から.

縦棒の前だけ見る (縦棒の後を固定する) と, ただの確率. 全事象は 1.

離散型確率変数に対する条件付き確率

離散型確率変数に対する条件付き確率 岩薩林 確率・統計 (3.15)

条件 (事象) $Y = y$ のもとでの, 事象 $X = x$ の条件付き確率の記号
 $P(X = x|Y = y) = p_{X|Y}(x \text{ の値} | y \text{ の値})$

例 $p_{X|Y}(150|45), p_{Y|X}(45|150)$. 縦棒の前だけ見る (縦棒の後ろを固定する) と, 前の変数についての, ただの確率分布.

命題 (条件付き確率を同時確率分布と周辺分布で表す式 岩薩林 確率・統計 (2.12))

$$P(X = x|Y = y) = p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)},$$
$$P(Y = y|X = x) = p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}.$$

条件付き確率と周辺分布から、同時確率分布、周辺分布

命題 (同時確率分布, 周辺分布, 条件付き確率の関係)

乗法公式 (同時確率分布との関係)

岩薩林 確率・統計 (2.13)

$$\begin{aligned} p(x, y) &= p_{X|Y}(x|y) \times p_Y(y) \\ &= p_{Y|X}(y|x) \times p_X(x). \end{aligned}$$

全確率の法則 (周辺分布との関係)

岩薩林 確率・統計 (2.17),(2.20)

$$\begin{aligned} p_X(x) &= \sum_y p_{X|Y}(x|y) \times p_Y(y), \\ p_Y(y) &= \sum_x p_{Y|X}(y|x) \times p_X(x) \end{aligned}$$

ベイズの定理 I

定理 (ベイズの定理 岩薩林 確率・統計 (2.18),(2.19))

$$p_{X|Y}(x|y) \stackrel{\text{定義}}{=} \frac{p(x, y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')},$$

$$p_{Y|X}(y|x) \stackrel{\text{定義}}{=} \frac{p(x, y)}{p_X(x)} = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y'} p_{X|Y}(x|y')p_Y(y')}.$$

$p_{X|Y}(x|y)$ を $p_{Y|X}(y|x')$ (と $p_X(x')$) で書き表す式,

$p_{Y|X}(y|x)$ を $p_{X|Y}(x|y')$ (と $p_Y(y')$) で書き表す式.

定義の分子分母を, 乗法公式と全確率の公式で書き直しただけ (なので余分の記憶は不要).

条件付き確率密度関数

Hoff ベイズ p.28

$f(x, y)$ の一方が連続型の時も, p を f で置き換えて, 同様の式で $f_{X|Y}(x|y)$, $f_{Y|X}(y|x)$ が定義できる.

全確率の法則 (周辺分布との関係)

X が連続型のとき,

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{Y|X}(y|x) f_X(x) dx$$

連続型の記号で, f について書く.

定義 (条件付き母期待値)

条件 $Y = y$ のもとでの $g(x, y)$ の条件付き母期待値とは, y の関数

$$E[g(X, Y)|Y = y] = \int_{-\infty}^{+\infty} g(x, y) f_{X|Y}(x, y) \, dx.$$

命題 (独立性)

確率変数 X, Y が独立とは,

$$f_X(x) = f_{X|Y}(x|y), \quad f_Y(y) = f_{Y|X}(y|x).$$

「 x の条件付き確率は, 条件 y によらない」

確率密度で書いたベイズの定理 I

定理 (ベイズの定理)

岩薩林 確率・統計 (2.18),(2.19)

Hoff ベイズ p.29

$$f_{X|Y}(x|y) \stackrel{\text{定義}}{=} \frac{f(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x')f_X(x')dx'},$$

$$f_{Y|X}(y|x) \stackrel{\text{定義}}{=} \frac{f(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{+\infty} f_{X|Y}(x|y')f_Y(y')dy'}.$$

$f_{X|Y}(x|y)$ を $f_{Y|X}(y|x')$ (と $f_X(x')$) で書き表す式,

$f_{Y|X}(y|x)$ を $f_{X|Y}(x|y')$ (と $f_Y(y')$) で書き表す式.

定義の分子分母を, 乗法公式と全確率の公式で書き直しただけ (なので余分の記憶は不要).

$\frac{1}{f_Y(y)}, \frac{1}{f_X(x)}$ を比例定数と思って, 比だけ考えることもできる.

L02-Q3

Quiz(連続型の条件付き確率)

連続型確率変数 X 離散型確率変数 Y は, 次の確率分布を持つ.

$$f(x, y) = \pi_y \frac{1}{(2\pi\sigma_y^2)^{1/2}} e^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}}$$

ただし, $\pi_0 = \frac{3}{10}, \pi_1 = \frac{7}{10}, \pi_{\text{他}} = 0, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = \frac{1}{2}$ である.

- ① $f_{X|Y}(x|y)$ を求めよう. 特に, $y = 1$ のとき.
- ② $f_{Y|X}(y|x)$ を求めよう. 特に, $x = 4$ のとき.

L02-Q4

Quiz(連続型の条件付き確率)

連続型確率変数 X , 離散型確率変数 Y は, 次の確率分布を持つ.

$$\begin{aligned} f(x, y) &= \pi_y h_y(x), \\ \pi_0 &= \frac{1}{3}, \pi_1 = \frac{2}{3}, \pi_{\text{他}} = 0, \\ h_y(x) &= \begin{cases} \lambda_y e^{-\lambda(x-b_y)} (x \geq b_y) \\ 0 \end{cases} \quad (\text{他}), \\ \lambda_0 &= 2, \lambda_1 = 1/2, b_0 = 0, b_1 = 3 \end{aligned}$$

- ① $f_Y(y)$ を求めよう.
- ② $f_{X|Y}(x|y)$ を求めよう. 特に, $y = 1$ のとき.
- ③ $f_X(x)$ のグラフを描こう (「混合指数分布」).
- ④ $f_{Y|X}(y|x)$ を求めよう. 特に, $x = 4$ のとき.

ベイズ推定の考え方

Hoff ベイズ p.29

X, Y が連続型のとき,

事後確率密度 $f_{Y|X}(y|x)$

←

事前確率密度 $f_Y(y)$

↑

情報 $X = x$

追加の情報が得られるたびに, 事後確率密度は正確になっていく.

主観確率=信念 **berief** 事前確率密度に主観を許す考え方. 事後確率密度にも主観が含まれる.

L02-Q5

Quiz(連続型の条件付き確率)

ある畑でとれる柿の実の $\frac{3}{10}$ は品種 A, $\frac{7}{10}$ は品種 B である.
品種 A の実の質量は, 母平均値 200 g, 母分散 20^2 g^2 ,
品種 B の実の質量は, 母平均値 180 g, 母分散 10^2 g^2 の, それぞれ正規分布にしたがう.

- ① 畑でとれる柿の実の質量の分布を求めよう (確率密度関数を書こう).
- ② 畑でとれたひとつの柿の実の質量が 190 g であるとき, 品種 A である確率を求めよう.