L06 混合ガウス分布

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論 L06(2021-11-02 Tue)

最終更新: Time-stamp: "2021-11-02 Tue 12:32 JST hig"

今日の目標

- 混合ガウス分布の確率密度関数を描ける
- 混合ガウス分布の母期待値を求められる
- 混合ガウス分布の (条件付き) 確率を求められる
- 混合ガウス分布の標本抽出ができる

ここまで来たよ

現実の汚い複雑なデータの pandas での扱い

前処理, データクレンジング

- 正規化
- 欠測値 NaN

データベース

DataFrame の縦分割

- Series df['pclass']
- boolean list df['pclass']==1
- df[df['pclass']==1] 条件を満たす行だけを残した DataFrame
- コラムの値によってグループ化した処理
  df.groupby('pclass').dosomething()

DataFrame の縦連結

- DataFrame の縦方向連結 pd.concat([df1,df2]), df1.append(df2)

ここまで来たよ

離散, 連続型確率変数の同時分布

X:連続型, Y:離散型確率変数 の多次元分布 確率統計☆演習 I(2021)L04 岩薩林 確率・統計 §3.3
同時分布 $f(x,y)$. x については確率密度, y について確率.

$$\mathrm{E}[g(X,Y)] = \int_{-\infty}^{+\infty} \sum_{y} g(x,y) f(x,y)\ \mathrm{d}x.$$

$$P(c \leq X < d, Y = y_0) = \int_{-\infty}^{+\infty} \sum_{y} \mathrm{I}_{[c \leq X < d, Y = y_0]}(x,y) f(x,y) \mathrm{d}x = \int_{c}^{d} f(x,y_0) \mathrm{d}x$$

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	$h_1(x)$

$\int_{-\infty}^{+\infty} h_0(x)\ \mathrm{d}x + \int_{-\infty}^{+\infty} h_1(x)\ \mathrm{d}x = 1.$

混合ガウス分布

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 Gaussian mixture, 他の混合分布もある.
準備として, $\pi_0 + \pi_1 = 1, \pi_y \geq 0, \pi_{他} = 0$ として, 次の同時分布 $f(x, y)$ を考える.

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

X, Y は独立ではない ($\sigma_0 = \sigma_1, \mu_0 = \mu_1$ でない限り).

Y の周辺分布は離散型

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

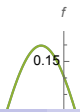
$Y \sim B(1,\pi_1)$ じゃん (パラメタ $p=\pi_1$ のベルヌイ分布, パラメタ $n=1, p=\pi_1$ の二項分布 確率統計☆演習 I(2021)L07 | 岩薩林 確率・統計 §3.4).

X の周辺分布は連続型

これが, パラメタ $(\pi_0,\pi_1,\mu_0,\mu_1,\sigma_0^2,\sigma_1^2)$ の GMM=Gaussian Mixture Model=混合ガウス分布

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

y の値が 2 個と限らず有限個であるものが一般の混合ガウス分布.
ガウス分布以外も, 重み π_y で重ね合わせることができる. 一般の混合分布.



L06-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
$(\pi_0 = 1/3, \pi_1 = 2/3, \mu_0 = -1, \mu_1 = 3, \sigma_0 = 1, \sigma_1 = 4)$ にしたがう. 確率密度関数 $f(x)$ のグラフの概形を描こう.
教科書の表で $I(z)$ で, または, scipy.norm で浮動小数点数で.

th-d06-2.ipynb

L06-Q2

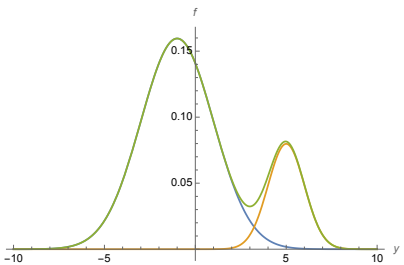
Quiz(混合ガウス分布の確率)

X は混合ガウス分布
$(\pi_0 = 1/3, \pi_1 = 2/3, \mu_0 = -1, \mu_1 = 3, \sigma_0 = 1, \sigma_1 = 4)$ にしたがう. X は混合ガウス分布 $(\pi_0 = 1/3, \pi_1 = 2/3, \mu_0 = -2, \mu_1 = 3, \sigma_0 = 1, \sigma_1 = 4)$ にしたがう. 確率 $P(X \leq -3)$ を求めよう.
教科書の表で $I(z)$ で, または, scipy.norm で浮動小数点数で.

th-d06-2.ipynb

混合ガウス分布のモーメント，母期待値

$$
\begin{aligned}
\mathrm{E}[X^k] &= \int \sum_y x^k \pi_y \cdot \frac{1}{(2\pi\sigma_y^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} \,\mathrm{d}x \\
&= \pi_0 \int x^k \frac{1}{(2\pi\sigma_0^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \,\mathrm{d}x + \pi_1 \int x^k \frac{1}{(2\pi\sigma_1^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \,\mathrm{d}x. \\
\mathrm{E}[X^0] &= \pi_0 1 + \pi_1 1 = 1. \\
\mathrm{E}[X^1] &= \pi_0\mu_0 + \pi_1\mu_1. \\
\mathrm{V}[X] &= \text{着実な計算} = \pi_0\sigma_0^2 + \pi_1\sigma_1^2 + \pi_0\pi_1(\mu_0 - \mu_1)^2. \\
\mathrm{E}[Y^k] &= \text{周辺分布はただのベルヌイ分布} = \pi_1 \quad (k = 1, 2, 3, \ldots).
\end{aligned}
$$



L06-Q3

Quiz(混合ガウス分布の共分散)

周辺分布 $f_{\mathrm{X}}(x)$ が混合ガウス分布 (π_y, μ_y, σ_y) になる同時分布 $f(x, y)$ $(y = 0, 1)$ を考える.

1. 母分散 $\mathrm{V}[X], \mathrm{V}[Y]$ を求めよう.
2. 母共分散 $\mathrm{Cov}[X, Y]$ を求めよう.

th-d06-2.ipynb

ここまで来たよ

6 ロジスティック回帰

6 混合ガウス分布
- DataFrame 操作
- 混合ガウス分布
- 混合ガウス分布と関係する条件付き分布と推定
- ガウス混合分布の標本抽出

どんな現実のシーン?

X: 体温 (or 何かの測定値)
Y: 人の感染の有 (1) 無 (0)
1 個の x のデータをとったとき, y を知りたい $\to$ $X = x_0$ であるという条件のもとでの y の条件付き確率を求めたい
しかし, 現実には, 母ナントカ π_y, μ_y, σ_y を知っているのは仏だけ. データサイエンティストは知らないが, (x, y) のデータ群を持っているので推定しようとする.
線形回帰やロジスティック回帰と同じ予測 (教師あり) の問題. 判別分析.

来週

混合ガウス分布と関係する条件付き分布

一般に, $X = x$ という条件のもとでの $Y = y$ の条件付き確率 確率統計☆演習 I(2021)L05

岩薩林 確率・統計 p.59

$$P(Y = y|X = x) = f_{\mathrm{Y|X}}(y|x) = \frac{f(x, y)}{\sum_{y'} f(x, y')}$$

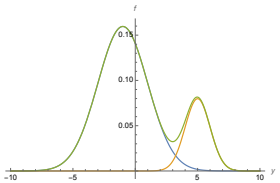
以下, 略記 $f(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

$X = x_0$ であるという条件のもとでの y の条件付き確率

$$\begin{aligned} P(Y = 1|X = x_0) =& f_{\mathrm{Y|X}}(1|x_0) = \frac{f(1, y)}{\sum_{y'} f(x_0, y')} \\ =& \frac{\pi_1 f(x_0; \mu_1, \sigma_1^2)}{\pi_0 f(x_0; \mu_0, \sigma_0^2) + \pi_1 f(x_0; \mu_1, \sigma_1^2)} \end{aligned}$$

$Y = y_0$ であるという条件のもとでの x の条件付き確率密度

$$p(X = x|Y = y) = \frac{\pi_y f(x; \mu_y, \sigma_y^2)}{\pi_y \int_{-\infty}^{+\infty} f(x'; \mu_y, \sigma_y^2)\, r dx'} = f(x; \mu_y, \sigma_y^2)$$



th-d06-2.ipynb

ここまで来たよ

6 ロジスティック回帰

6 混合ガウス分布
- DataFrame 操作
- 混合ガウス分布
- 混合ガウス分布と関係する条件付き分布と推定
- ガウス混合分布の標本抽出

標本抽出

(π_y, μ_y, σ_y):既知とする (仏の立場).

道具 1　コイン (二項分布) scipy.stats.binom(n=1,p=π_1).rvs(size=1)
道具 2　正規分布連続サイコロ
scipy.stats.norm(loc=μ_y,scale=σ_y).rvs(size=1)

アルゴリズム

- n 回繰り返す.
  - コインを投げて $y = 0, 1$ を決定
  - 次に μ_y, σ_y に調節したサイコロを投げて x を得る
  - 組み合わせた (x, y) がひとつのデータ.

これで, Y の周辺分布 $f_{\mathrm{Y}}(y)$, 条件付き分布 $f_{\mathrm{X|Y}}(x|y)$ が正しいでしょ.
th-d06-1.ipynb

標本抽出して散布図を描こう.