L06 混合ガウス分布

樋口さぶろお　https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習　L06(2021-11-04 Thu)

最終更新: Time-stamp: "2021-11-03 Wed 10:31 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

4. 線形回帰モデル - 重回帰

6. 混合ガウス分布
   - DataFrame 操作
   - 混合ガウス分布
   - 混合ガウス分布を用いた分類
   - 混合ガウス分布の標本抽出

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

データベース

- 正規化
- 欠測値 NaN

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

4 線形回帰モデル - 重回帰

6 混合ガウス分布
- DataFrame 操作
- 混合ガウス分布
- 混合ガウス分布を用いた分類
- 混合ガウス分布の標本抽出

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 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	$h_1(x)$

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

混合ガウス分布 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, $\pi_0 + \pi_1 = 1, \pi_y \geq 0, \pi_{他} = 0$ 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でベルヌイ分布 $B(1,\pi_1)$

確率統計 I(2021)L07 | 岩薩林 確率・統計 §3.4

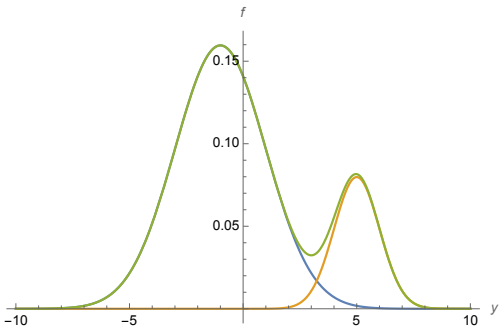
$$f_{\mathrm{Y}}(y)=\int f(x,y)\mathrm{d}x=\int_{-\infty}^{+\infty}\pi_y\cdot\frac{1}{(2\pi\sigma_y^2)^{1/2}}\mathrm{e}^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}}\,\mathrm{d}x=\pi_y.$$

X の周辺分布は連続型

GMM=混合ガウス分布

パラメタ $(\pi_0,\pi_1,\mu_0,\mu_1,\sigma_0^2,\sigma_1^2)$ の GMM とは

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



mva-d06-1.ipynb

一般の混合ガウス分布: y の値が 2 個と限らず有限個

一般の混合分布: ガウス分布以外を重み π_y で重ね合わせ

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
$(\pi_0 = 3/10, \pi_1 = 7/10, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = 1/2)$ 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.norm で浮動小数点数で.

mva-d06-1.ipynb

L06-Q2

Quiz(混合ガウス分布の確率)

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

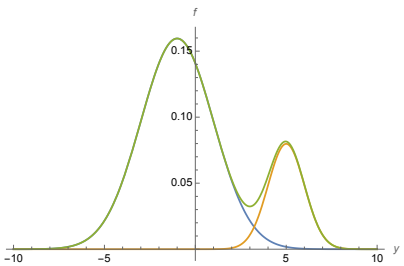
1. 確率密度 $f(4)$ を求めよう.
2. 確率 $P(X \leq 4)$ を求めよう.

教科書の表で $I(z)$ で, または, scipy.norm で浮動小数点数で.

mva-d06-1.ipynb

混合ガウス分布のモーメント，母期待値を周辺分布から

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



ここまで来たよ

4 線形回帰モデル - 重回帰

6 混合ガウス分布
- DataFrame 操作
- 混合ガウス分布
- 混合ガウス分布を用いた分類
- 混合ガウス分布の標本抽出

どんな現実のシーン?

確率変数 X: 体温 (or 何かの測定値)
確率変数 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 $\rightarrow$ $X = x_0$ 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.
ロジスティック回帰と似た分類問題 (教師あり)

1. 線形回帰, ロジスティック回帰では説明変数 x は確率変数でなく, 人間が指定する.
2. 判別分析では, 独立でない確率変数 X,Y が特定の確率分布 (例:混合ガウス分布のもとの同時分布)=モデルにしたがうことがわかっている.

来週

混合ガウス分布と関係する条件付き分布

一般に, $Y = y_0$ という条件のもとでの $X = x$ の条件付き確率 確率統計 I(2021)L05

岩薙林 確率・統計 p.59

$$P(X = x|Y = y_0) = f_{\mathrm{X|Y}}(x|y_0) = \frac{f(x, y_0)}{\sum_{x'} f(x', y_0)}$$

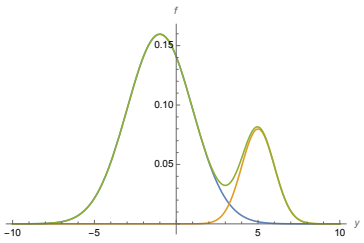
以下, 略記 $f(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

$X = x_0$ であるという条件のもとでの y の条件付き確率

次週以降

$Y = y$ であるという条件のもとでの x の条件付き確率密度

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



mva-d06-1.ipynb

ここまで来たよ

4. 線形回帰モデル - 重回帰

6. 混合ガウス分布
   - DataFrame 操作
   - 混合ガウス分布
   - 混合ガウス分布を用いた分類
   - 混合ガウス分布の標本抽出

混合ガウス分布の標本抽出のアルゴリズム

(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)

アイデア
$f_{\mathrm{X}}(x)$ は難しい.
$f(x,y) = f_{\mathrm{X|Y}}(x|y) \cdots f_{\mathrm{Y}}(y)$ で (x,y) を得た後, 周辺分布を作る (y を無視する).
アルゴリズム `mva-d06-2.ipynb`

- 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.