ベイズ推定

樋口さぶろお https://hig3.net

龍谷大学大学院 理工学研究科 数理情報学専攻

理論物理学特論　L04 (2022-10-12 Wed)

最終更新: Time-stamp: "2022-10-12 Wed 07:54 JST hig"

今日の目標

- 混合ガウス分布の標本を抽出できる
- 混合ガウス分布の標本から判別できる
- ナイーブベイズ推定を説明できる

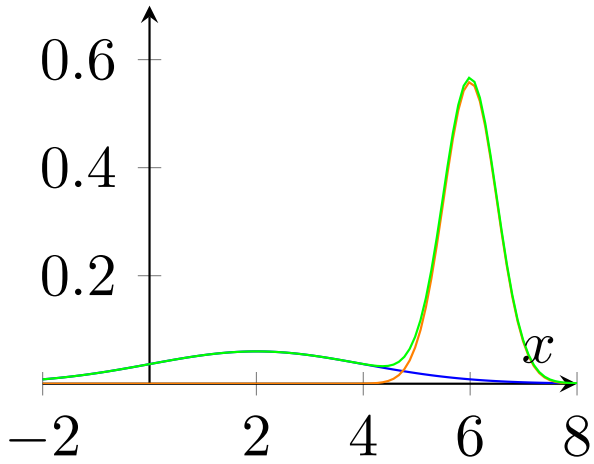
L03-Q1

Quiz 解答:2 次元正規分布の条件付き

1. 条件付き確率密度関数は $f_{\mathrm{Y|X}}(y|2) = C'\mathrm{e}^{-4^2+8y-7y^2} = C''\mathrm{e}^{-\frac{1}{2(1/14)}(y-\frac{4}{7})^2} = \frac{1}{\sqrt{2\pi(1/14)}}\mathrm{e}^{-\frac{1}{2(1/14)}(y-\frac{4}{7})^2}$. $X=2$ という条件のもとで, $Y \sim \mathrm{N}(\frac{4}{7}, \frac{1}{14})$.
2. $\mathrm{E}[Y|X=2] = \frac{4}{7}$.
3. 条件付き確率密度関数は $f_{\mathrm{X|Y}}(x|-1) = C'\mathrm{e}^{-x^2-4x-7} = C''\mathrm{e}^{-\frac{1}{2(1/2)}(x-2)^2} = \frac{1}{\sqrt{2\pi(1/2)}}\mathrm{e}^{-\frac{1}{2(1/2)}(x-2)^2}$. $Y=-1$ という条件のもとで, $Y \sim \mathrm{N}(2, \frac{1}{2})$.
4. $\mathrm{E}[X|Y=-1] = 2$. $\mathrm{V}[X|Y=-1] = \frac{1}{2}$.
   $\mathrm{V}[X^2|Y=-1] = \frac{1}{2} + 2^2 = \frac{9}{2}$.

L03-Q2

Quiz 解答: 混合ガウス分布の確率密度関数



L03-Q3

Quiz 解答: 混合ガウス分布の確率

1. $f(4) = \frac{3}{10}\frac{1}{\sqrt{2\pi\cdot 2^2}}\mathrm{e}^{-\frac{(4-2)^2}{2\cdot 2^2}} + \frac{7}{10}\frac{1}{\sqrt{2\pi\cdot(1/2)^2}}\mathrm{e}^{-\frac{(4-6)^2}{2\cdot(1/2)^2}}$.
2. $P(X \leq 4) = \frac{3}{10}F(\frac{4-2}{2}) + \frac{7}{10}F(\frac{4-6}{1/2})$.

ここまで来たよ

3 条件付き確率, 条件付き母期待値

4 ベイズ推定
- 混合ガウス分布と関係する条件付き分布
- ナイーブベイズによる分類
- ベイズ推定

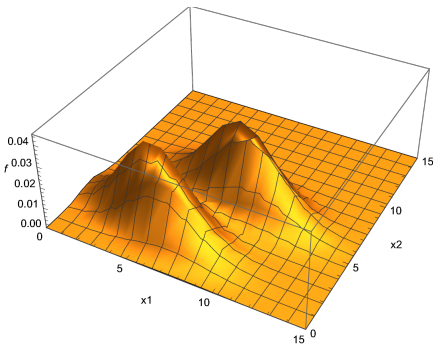
復習: 周辺分布が混合ガウス分布になる 2 次元分布

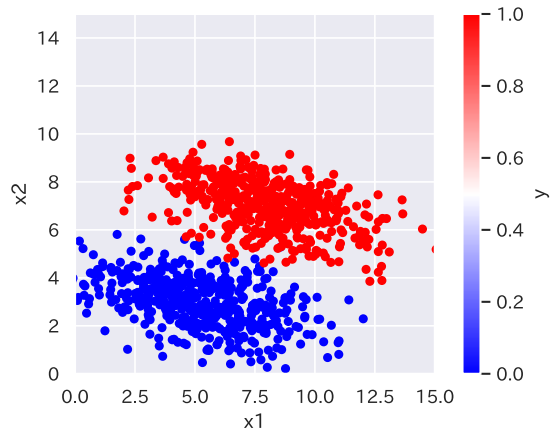
2 次元の混合ガウス分布

定義 (2 次元の混合ガウス分布)

1 次元混合ウス分布で, x が 2 次元ベクトル $\boldsymbol{x} = {}^{\mathrm{t}}(x_1, x_2)$ になっただけ.

$$f(\boldsymbol{x}, y) = \begin{cases} \dfrac{1}{(2\pi)^{2/2}(\det \Sigma_0)^{1/2}} \mathrm{e}^{-\frac{1}{2}\,{}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{\mu}_0)\Sigma_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)} \cdot \pi_0 & (y=0) \\ \dfrac{1}{(2\pi)^{2/2}(\det \Sigma_1)^{1/2}} \mathrm{e}^{-\frac{1}{2}\,{}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{\mu}_1)\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)} \cdot \pi_1 & (y=1) \\ 0 & (y\text{が他}) \end{cases}$$





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

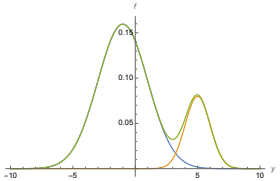
$X = x_0$ であるという条件のもとでの y の条件付き確率

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

$Y = y$ であるという条件のもとでの x の条件付き確率密度

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

$X|y \sim \mathrm{N}(\mu_y, \sigma_y^2)$ とかく.



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.
$f(x,y) = f_{\mathrm{X|Y}}(x|y) \cdot f_{\mathrm{Y}}(y)$ 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

- 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) が同時分布のひとつのデータ
  - x が混合ガウス分布のひとつのデータ

ここまで来たよ

3. 条件付き確率, 条件付き母期待値

4. ベイズ推定
   - 混合ガウス分布と関係する条件付き分布
   - ナイーブベイズによる分類
   - ベイズ推定

分類問題 classification

(x, y) は, ある同時分布 (パラメタ未知) から生成される

- 訓練データ: 既知の大きい標本 (x_i, y_i) $(i = 1, \ldots, n)$. $(x_{\text{train}}, y_{\text{train}})$ とも書かれる.
  - $y = 0, 1$: ラベル, カテゴリ, 分類結果. 今の場合, 教師シグナル.
- テストデータ: x_{test} と未知の正解 y_{test}. 1 個または多数.

ップ 1 訓練データから予測器を作っておき (=母分布のパラメタを推定しておき)

ップ 2 テストデータに対して予測 (分類) する

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 $\to$ $X = x_0$ 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母ナントカを知らないが, (x, y) のデータ群を持っているので推定しようとする.

混合ガウス分布のナイーブベイズ法 (ステップ 1)

パラメタ (π_y, μ_y, σ_y) 未知, 訓練データ (x_i, y_i) $(i = 1, \ldots, n)$ 既知のとき
$y_i = 1$ を $y_j^{(1)}$ $(j = 1, \ldots, k)$, $y_i = 0$ を $y_j^{(0)}$ $(j = 1, \ldots, n-k)$ と命名.
確率統計 岩薙林 確率・統計 §7 のりで, パラメタを推定する.

確率 π_1

Y の周辺分布 $\mathrm{B}(1, \pi_1)$ で, π_1 は確率または母比率 $P(Y = 1)$.
母比率の推定 岩薙林 確率・統計 §7.2. 標本比率 $\hat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{k}{n}$.

母平均値 μ_1 と母分散 σ_1^2

X の条件付き分布 $P(X = x|Y = 1)$ は, $\mathrm{N}(\mu_1, \sigma_1^2)$. 訓練データのうち
$y_j^{(1)}$ はこれの標本.
母平均値の推定値 岩薙林 確率・統計 §7.1 標本平均値 $\overline{x}^{(1)} = \frac{1}{k}[x_1^{(1)} + \cdots + x_k^{(1)}]$.
母分散の推定 岩薙林 確率・統計 §7.3 不偏標本分散
$(S^{(1)})^2 = \frac{1}{k-1}[(x_1^{(1)} - \overline{x}^{(1)})^2 + \cdots + (x_k^{(1)} - \overline{x}^{(1)})^2]$.
π_0, μ_0, σ_0 も同様.

混合ガウス分布のナイーブベイズ法 (ステップ 2)

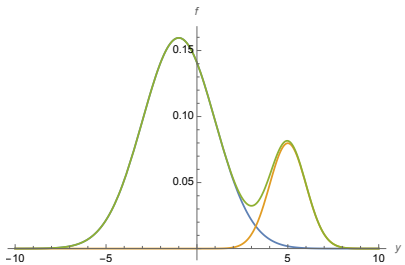
パラメタ (π_y, μ_y, σ_y) 既知のとき, 条件付き確率 (ベイズの定理) で

条件付確率

$$P(Y=y|X=x_{\text{test}}) = \frac{f(x_{\text{test}}, y)}{\sum_{y'} f(x_{\text{test}}, y')} = \frac{f_{\text{Y|X}}(x_{\text{test}}|y)\cdot f_{\text{Y}}(y)}{\sum_{y'} f_{\text{Y|X}}(x_{\text{test}}|y')\cdot f_{\text{Y}}(y')}$$
$$= \frac{f(x_{\text{test}}; \mu_y, \sigma_y^2)\pi_y}{f(x_{\text{test}}; \mu_0, \sigma_0^2)\pi_0 + f(x_{\text{test}}; \mu_1, \sigma_1^2)\pi_1}$$

正規分布の確率密度関数 $f(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

積み重なる確率密度関数のグラフの下側の棒の長さの比.



混合ガウス分布の分類問題のナイーブベイズ法

ステップ 1,2 をまとめる

前提

- 2 カテゴリの混合ガウス分布

手順 (ナイーブベイズ法)

入力　　出力

訓練データ $(x_i, y_i) \overset{\text{推定}}{\rightsquigarrow} (\pi_y, \mu_y, \sigma_y)$

テストデータ x_{test}　　$\overset{\text{条件付き確率}}{\rightsquigarrow} P(Y = y|X = x_{\text{test}})$

確率でなく, Y の値そのものを答えろと言われたら, 条件付き確率 $P(Y = y|X = x_{\text{test}})$ が大きい方の y を答える.
正答率 $P(Y = y|X = x_{\text{test}})$, 誤答率 $1 - P(Y = y|X = x_{\text{test}})$.

L04-Q1

Quiz(混合ガウス分布のナイーブベイズ)

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.
パラメタを $(\pi_0 = 3/10, \pi_1 = 7/10, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = 1/2)$ とする.
$X = 1$ という条件のもとで $Y = 0$ である条件付き確率 $P(Y = 0|X = 1)$ を求めよう.

L04-Q2

Quiz(混合ガウス分布のナイーブベイズ推定 (訓練, 検証))

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.
次を学習データとして, $x = 0$ に対して, y の値を推定しよう (確率とともに).

x	y
-3	0
-1	0
1	1
2	1
3	1

scikit-learn を利用したナイーブベイズ

2 次元以上でも 1 次元と同様に実行できる. パラメタ $(\boldsymbol{\mu}_y, \Sigma_y)$.
$f(x; \mu, \sigma^2) \rightsquigarrow f(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$
Python では, scikit-learn ライブラリ `https://scikit-learn.org/` に含まれるものがある. $n = 1, 2, \ldots$ 次元で使える.

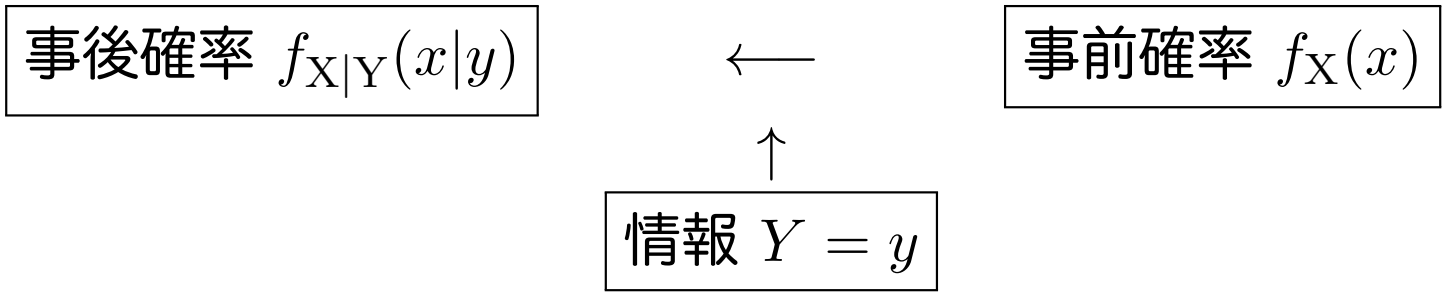
```
from sklearn import naive_bayes
nb=GaussianNB() # 訓練結果を保持するオブジェクト
nb.fit(x_train,x_train) # DataFrame を与える

nb.predict(x_test) # 0か1か答えてくれる
nb.いろいろ # パラメタの推定結果など
```

ここまで来たよ

3 条件付き確率, 条件付き母期待値

4 ベイズ推定
- 混合ガウス分布と関係する条件付き分布
- ナイーブベイズによる分類
- ベイズ推定

ベイズ推定の考え方



主観確率

事前確率に主観を許す考え方．結果として，事後確率にも主観が含まれる．

L04-Q3

Quiz(ベイズ推定)

ある病気の人の割合は全体の 0.005 と思われている.
検査では, 病気の人の 0.99 は陽性となり (真陽性), 0.01 の人は陰性になる (偽陰性). また, 病気でない人の 0.02 は (誤って) 陽性となり (偽陽性), 0.98 の人は陰性になる (真陽性).

1. 1 回の検査で陽性となった場合, その人が病気である確率を求めよう.
2. 2 回の検査で 2 回とも陽性となった場合, その人が病気である確率を求めよう (2 回の検査は互いに独立であるとする).

推定すべき母分布のパラメタを確率変数と考える

未知の正規分布 $Y \sim \mathrm{N}(\mu, \sigma^2)$ があり, μ, σ を Y の標本から推定することを考える.

ベイズ推定

μ, σ を確率変数 X であるかのように考え, (主観的な) 事前分布 $f(\mu, \sigma)$ が, Y の標本により, 事後分布 $f(\mu, \sigma|y)$ に修正される, と見る.

(病気の例では, 被験者が病気であるないを確率変数 $X = 0, 100$ とみなしたようなもの. X はベルヌーイ分布 $\mathrm{B}(1, p)$ にしたがう. $p \leftrightarrow f(\mu, \sigma)$.)

説明の簡単のため, なぜか $\sigma^2 = 2^2$ はわかっていて固定されている (確率変数でない) かのように扱う.

$$f_{\mathrm{Y|X}}(y|\mu) = \frac{1}{\sqrt{2\pi \cdot 2^2}} \mathrm{e}^{-\frac{1}{2\cdot 2^2}(y-\mu)^2}.$$

$X = \mu$ の事前分布は主観的に, または領域知識から選んでいいのだが, いま, これも正規分布にしたがうとする. $\mu \sim \mathrm{N}(\mu_0, \sigma_0)$ すなわち,

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \mathrm{e}^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

とする. さらに気分を出すため, $\mu_0 = -2, \sigma_0^2 = 3^2$ と固定しよう.
Y の標本サイズ 1 の標本 $\{-3\}$ が得られたときの, 事後分布

$$f_{\mathrm{X|Y}}(\mu|-3) = \frac{\frac{1}{\sqrt{2\pi\cdot 2^2}} \mathrm{e}^{-\frac{1}{2\cdot 2^2}((-3)-\mu)^2} \frac{1}{\sqrt{2\pi\cdot 3^2}} \mathrm{e}^{-\frac{1}{2\cdot 3^2}(\mu-(-2))^2}}{(\text{分子の } \mu \text{ 積分を 1 にする定数})} = C\mathrm{e}^A$$

$$A = -\frac{1}{2\cdot(36/13)}(\mu - (-\tfrac{35}{13}))^2$$

すなわち, $\mu|Y = -3 \sim \mathrm{N}(-\frac{35}{13}, \frac{36}{13})$.
$\mathrm{N}(-2, 3^2) \rightsquigarrow \mathrm{N}(-\frac{35}{13}, \frac{36}{13})$ と更新された.
この場合はたまたま事後分布が (事前分布と同じ) 正規分布にもどった!
「共役事前分布である」

一般に, サイズ n の標本 $\{y_i\}_{i=1,\dots,n}$ が得られたときの事後分布を考える.

$$f_{\mathrm{X|Y}}(\mu|\{y_i\}_{i=1}^n) = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\cdot\sigma^2}}\mathrm{e}^{-\frac{1}{2\cdot\sigma^2}(y_i-\mu)^2}\frac{1}{\sqrt{2\pi\cdot 3^2}}\mathrm{e}^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}}{(\text{分子の }\mu\text{ 積分を }1\text{ にする定数})} = C\mathrm{e}^A$$

$$A = -\frac{1}{2\frac{\sigma^2}{n}\cdot(1+\frac{\sigma^2}{n\sigma_0^2})^{-1}}\left(\mu - \frac{\frac{1}{n}\sum y_i + \frac{1}{n}\frac{\sigma^2}{\sigma_0^2}\mu_0}{1+\frac{1}{n}\frac{\sigma^2}{\sigma_0^2}}\right)^2$$

事後分布も正規分布で,
標本平均値と, 事前分布の母平均値を内分した点を母平均値,
σ^2/n を, 事前分布の母分散で調整したものを母分散, としたもの.