L07 ナイーブベイズ・線形判別分析

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　L07(2021-11-11 Thu)

最終更新: Time-stamp: "2021-11-11 Thu 06:51 JST hig"

今日の目標

- 2 次元混合ガウス分布を描ける
- 混合ガウス分布でナイーブベイズ法で分類できる
- 混合ガウス分布で線形判別分析で分類できる

ここまで来たよ

5 ロジスティック回帰

7 ナイーブベイズ・線形判別分析
- 1,2 次元の混合ガウス分布
- ナイーブベイズによる分類
- 1 次元の判別分析
- 2 次元の (線形) 判別分析

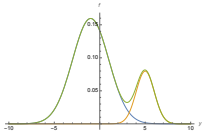
混合ガウス分布

GMM=混合ガウス分布

X:連続型確率変数, Y:離散型確率変数 $Y = 0, 1$.
パラメタ $(\pi_0, \pi_1, \mu_0, \mu_1, \sigma_0, \sigma_1)$.

$$
\begin{aligned}
f(x, y) = & f_{\mathrm{X|Y}}(x|y) && \text{(2) } y \text{ に応じた } \mu_y, \sigma_y \text{ で } x \text{ を決める} \\
& \cdot f_{\mathrm{Y}}(y) && \text{(1)} \pi_y \text{ で } y \text{ を決める} \\
= & \frac{1}{(2\pi\sigma_y^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} \\
& \cdot \pi_y
\end{aligned}
$$

この同時分布の周辺分布 $f_{\mathrm{X}}(x) = \sum_y f(x, y)$ が混合ガウス分布 (GMM)



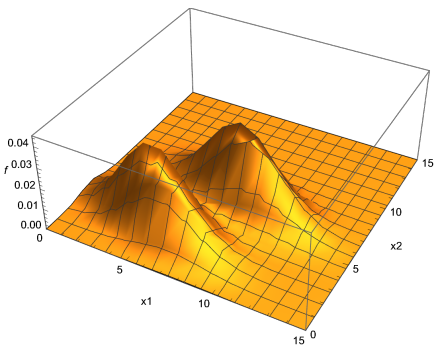
2 次元の混合ガウス分布

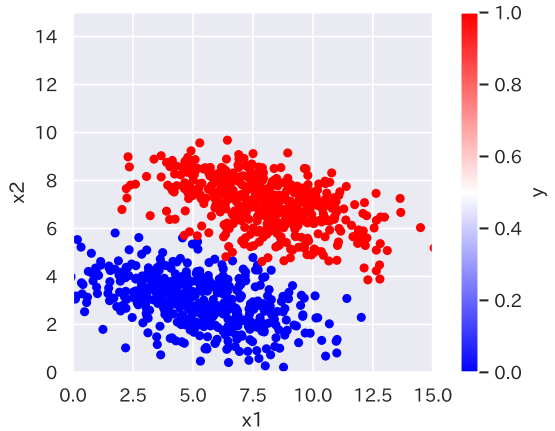
永田棟方 多変量解析法入門 §2.2(5) | 多変量解析☆演習 (2021)L02

2 次元の混合ガウス分布

x が 2 次元ベクトル $\boldsymbol{x} = {}^{\mathrm{t}}(x_1, x_2)$ になっただけ.

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





ここまで来たよ

5 ロジスティック回帰

7 ナイーブベイズ・線形判別分析
- 1,2 次元の混合ガウス分布
- ナイーブベイズによる分類
- 1 次元の判別分析
- 2 次元の (線形) 判別分析

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

y がラベルじゃなくて連続値なら回帰, 両方合わせて予測問題

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 岩薩林 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 岩薩林 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 岩薩林 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 岩薩林 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)

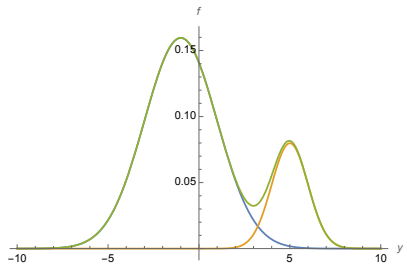
パラメタ (π_y, μ_y, σ_y) 既知のとき, 条件付き確率 (ベイズの定理) で

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

ナイーブベイズ法

入力 出力

訓練データ $(x_i, y_i) \overset{\text{推定}}{\rightsquigarrow} (\pi_y, \mu_y, \sigma_y) \overset{\text{条件付き確率}}{\rightsquigarrow} P(Y = y|X = x_{\text{test}})$

テストデータ x_{test} $\rightsquigarrow$

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 $1 - P(Y = y|X = x_{\text{test}})$.

mva-d07-0.ipynb

L07-Q1

Quiz(混合ガウス分布のナイーブベイズ)

周辺分布 $f_X(x)$ 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.

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.

```
1 from sklearn import naive_bayes
2 nb=GaussianNB() # 訓練結果を保持するオブジェクト
3 nb.fit(x_train,x_train) # DataFrame を与える
4
5 nb.predict(x_test) # かか答えてくれる01
6 nbいろいろ. # パラメタの推定結果など
```

ここまで来たよ

5 ロジスティック回帰

7 ナイーブベイズ・線形判別分析
- 1,2 次元の混合ガウス分布
- ナイーブベイズによる分類
- 1 次元の判別分析
- 2 次元の (線形) 判別分析

判別分析 Discriminant Analysis

永田棟方 多変量解析法入門 §7.2

前提

- 2 カテゴリの混合ガウス分布
- (追加)$\pi_0 = \pi_1 = \frac{1}{2}, \sigma_0 = \sigma_1 = \sigma$ を前提.

$y = 1$ と結論するのは次のとき.

$$P(Y = 1|X = x_{\text{test}}) = \frac{f(x_{\text{test}}; \mu_1, \sigma_1^2) \cdot \pi_1}{f(x_{\text{test}}; \mu_0, \sigma_0^2) \cdot \pi_0 + f(x_{\text{test}}; \mu_1, \sigma_1^2) \cdot \pi_1}$$

$$\geq \frac{1}{2} \geq \frac{f(x_{\text{test}}; \mu_0, \sigma_0^2) \cdot \pi_0}{f(x_{\text{test}}; \mu_0, \sigma_0^2) \cdot \pi_0 + f(x_{\text{test}}; \mu_1, \sigma_1^2) \cdot \pi_1} = P(Y = 0|X = x_{\text{test}})$$

正規分布の確率密度関数 $f(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
前提のせいで, 次の $z < 0$ と同値. x は中間点 $\frac{\mu_0+\mu_1}{2}$ のどちら側か判定してる.

Fisher の線形判別関数 ($z < 0$ なら $y = 1$ と答える)

$$z = \frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_0}{\sigma}\right)^2\right] \overset{\text{整理}}{=} \frac{\mu_0 - \mu_1}{\sigma} \cdot \frac{x - \frac{\mu_0+\mu_1}{2}}{\sigma}$$

マハラノビスの距離

永田棟方 多変量解析法入門 §7.2

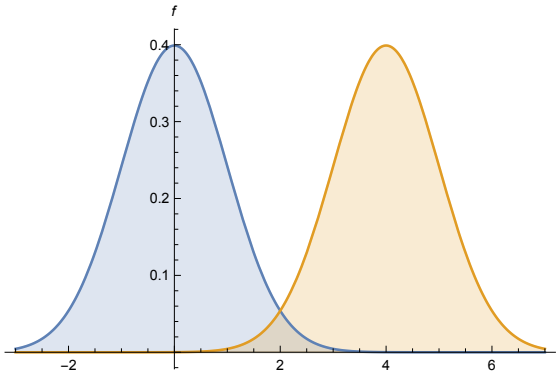
$$y_{\text{test}} = 1 \text{ と答える} \Leftrightarrow z \leq 0 \Leftrightarrow \left|\frac{x_{\text{test}}-\mu_1}{\sigma}\right| \leq \left|\frac{x_{\text{test}}-\mu_0}{\sigma}\right|.$$

x_{test} と 母平均値 μ_y とのマハラノビス (Maharanobis) 距離 $D = \left|\frac{x-\mu_y}{\sigma}\right|$ が短い方の y を答える, と言ってもいい.

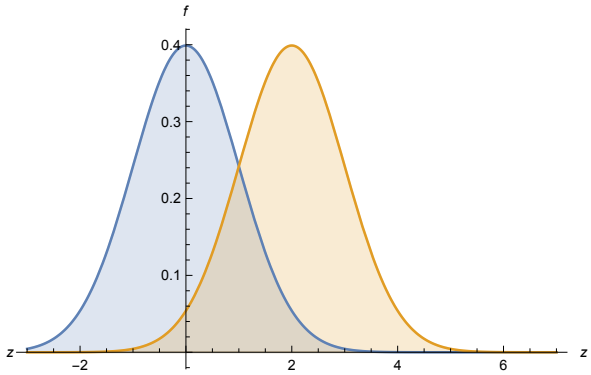
距離空間, 位相入門

分布から x_{test} を取ってくるときの誤答確率は, 2 つの母平均値の間のマハラノビス距離 $D = \left|\frac{\mu_0-\mu_1}{\sigma}\right|$.

誤答確率$= P(Z \geq \frac{D}{2}) \times \frac{1}{2} \times 2 = \frac{1}{2} - I(\frac{D}{2})$.



$D = 4$



$D = 2$

L07-Q2

Quiz(混合ガウス分布の線形判別関数)

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 $(\pi_0 = \pi_1 = 1/2, \mu_0 = 2, \mu_1 = 6, \sigma_0 = \sigma_1 = 2)$ とする.
Fisher の線形判別関数を作ろう.

ここまで来たよ

5 ロジスティック回帰

7 ナイーブベイズ・線形判別分析
- 1,2 次元の混合ガウス分布
- ナイーブベイズによる分類
- 1 次元の判別分析
- 2 次元の (線形) 判別分析

2 次元の (線形) 判別分析

永田棟方 多変量解析法入門 §7.3

Fisher の LDA=Linear Discriminant Analysis

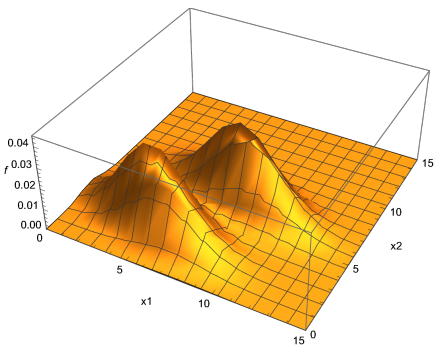
最近は, LDA で検索すると Latent Dirichlet Allocation 潜在的ディリクレ配分法という自然言語のトピックモデルが上位に出てくるが別物

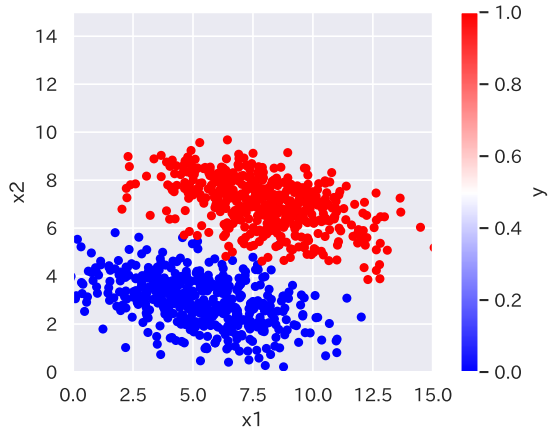
また強い前提 $\pi_0 = \pi_1 = \frac{1}{2}, \Sigma_0 = \Sigma_1 = \Sigma$.

(x_1, x_2, y) のうち, x_1 だけ, x_2 だけつかうものぐさが考えられる. x_1, x_2 どっちがいいの?

$z = w_1 x_1 + w_2 x_2 = {}^t\boldsymbol{w}\boldsymbol{x}$ としては? 最適の (w_1, w_2) に調節しよう

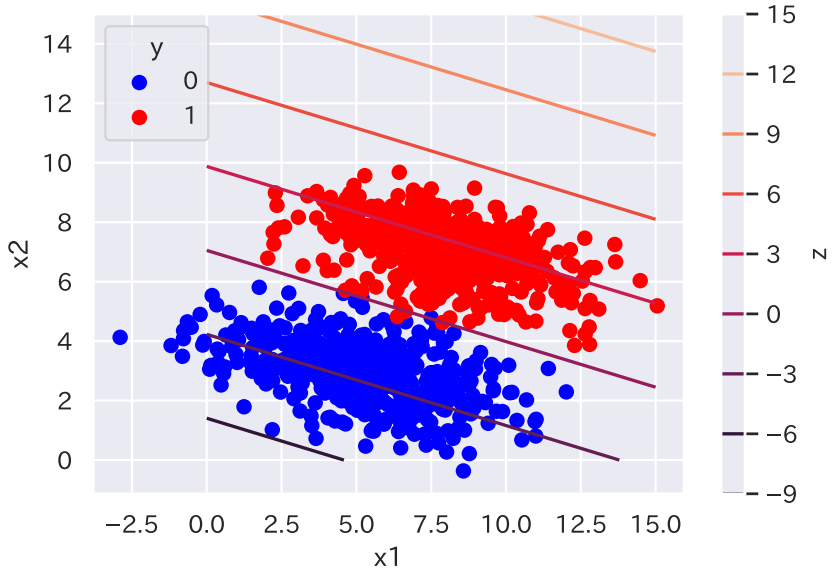
パラメタ $\boldsymbol{\mu}, \Sigma$ が既知の場合, $\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ とするとよい. 2 次元正規分布の等高線の, 横長効果+傾き効果を考慮.





Fisher の線形判別関数 (2 次元)

$$
\begin{aligned}
z &= {}^{\mathrm{t}}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\Sigma^{-1}\left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_0+\boldsymbol{\mu}_1}{2}\right) \\
&= \frac{1}{2}\left[{}^{\mathrm{t}}(\boldsymbol{x} - \boldsymbol{\mu}_1)\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - {}^{\mathrm{t}}(\boldsymbol{x} - \boldsymbol{\mu}_0)\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_0)\right]
\end{aligned}
$$



L07-Q3

Quiz(2 次元混合ガウス分布の線形判別関数)

周辺分布 $f_{\mathrm{X}}(\boldsymbol{x})$ が 2 次元混合ガウス分布 $(\pi_y, \boldsymbol{\mu}_y, \Sigma_y)$ になる同時分布 $f(\boldsymbol{x}, y)$ $(y = 0, 1)$ を考える.
パラメタを $(\pi_0 = \pi_1 = 1/2, \mu_0 = {}^{\mathrm{t}}(2, 3), \mu_1 = {}^{\mathrm{t}}(4, 9), \Sigma_0 = \Sigma_1 = \left(\begin{smallmatrix} 4 & 0 \\ 0 & 9 \end{smallmatrix}\right))$ とする.
Fisher の線形判別関数を作ろう.

傾いた直線 $z = C$ を境い目に $y = 0, 1$ を判別.

d 次元の, $\boldsymbol{x}$ と $\boldsymbol{x}'$ の間のマハラノビスの距離 (正規分布の e の肩に乗ってる形)

$$D = ({}^{\mathrm{t}}(\boldsymbol{x} - \boldsymbol{x}')\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{x}'))^{1/2}$$

距離空間, 位相

$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ のとき, 'x の成分ごとに標準化'

$$D = \left[\left(\frac{x_1 - x_1'}{\sigma_1} \right)^2 + \left(\frac{x_2 - x_2'}{\sigma_2} \right)^2 \right]^{1/2}$$

$\Sigma_1 \neq \Sigma_2$ の場合を想像すると, 判別の境界線は直線ではなく 2 次曲線になるはず $\rightsquigarrow$ 2 次判別分析 Quadratic Discriminant Analysis

scikit-learn の classifier comparison で, もっとハードな分類問題と, もっとハイテクな分類器. 右端の 2 個がナイーブベイズと 2 次判別分析.

`https://scikit-learn.org/stable/auto_examples/`

scikit-learn の LinearDiscriminantAnalysis

有名ライブラリ scikit-learn には, 訓練データから $\boldsymbol{w}$ を推定して, 判定までしてくれる LinearDiscriminantAnalysis がある.

```
1 from sklearn.discriminant_analysis import LinearDiscriminant
2 lda=LinearDiscriminantAnalysis() # 訓練結果を保持するオブジェクト
3 lda.fit(x_train,x_train) # DataFrame を与える
4
5 lda.predict(x_test) # かか答えてくれる01
6 lda.transform(x_test) # 線形判別関数の値
7 lda.coef_ # 係数 w
```

mva-d07-1.ipynb