

スパース回帰分析とパターン認識

第2章 統計手法によるパターン認識

担当: 長倉大輔

統計手法によるパターン認識

■ 判別分析(パターン認識)

判別分析 (パターン認識) とは、複数の母集団が与えられているとき、ある個体の観測値(ベクトル)より、その個体が属する群を推測する分析手法

例) 銀行のATMでの紙幣認識

走行中の車両ナンバーの自動読み取り

出入国の顔認証

ベイズ判別法

■ 判別ルールの評価基準

データを**誤って**判別する確率を**誤判別確率**という。

判別ルールの一つとして、誤判別確率が**小さいほどよい**というものは自然であろう。

以下では誤判別確率を最小化することによって判別する**ベイズ判別ルール**を紹介する。

ベイズ判別法

■ ベイズ判別ルール(2群の場合)

まずはデータが1次元(スカラー)の場合を考える。データ x について、母集団 Π_k , $k=1, 2$ に属するデータの確率密度関数は $f_k(x)$ で既知とする。

分類したいデータ x が Π_k から得られる確率を**事前確率**という。
以下では事前確率を π_k とする。

x が Π_k から得られた時 (x が Π_k に属するとき)、 x の**クラスラベル**は k であるということにする。

ベイズ判別法

観測値 x の取りうる値の集合を $\mathbb{R} = (-\infty, \infty)$ とする。

$\mathbb{R}$ を $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbb{R}$ かつ $\mathcal{R}_1 \cap \mathcal{R}_2 = \Phi$ (空集合)となるような領域 $\mathcal{R}_1$ と $\mathcal{R}_2$ に分割し、 $x \in \mathcal{R}_1$ なら x を Π_1 に $x \in \mathcal{R}_2$ なら x を Π_2 と判別することにする。

Π_1 の確率分布に従う確率変数 X の観測値を Π_2 の確率分布から得られた観測値であると誤判別する確率を

$$P(X \in \mathcal{R}_2 | \Pi_1)$$

のように表記する。逆も同様に表記する。

ベイズ判別法

この時、判別ルール $\{\mathcal{R}_1, \mathcal{R}_2\}$ による誤判別確率は

$$\begin{aligned} & P(\Pi_1)P(X \in \mathcal{R}_2 | \Pi_1) + P(\Pi_2)P(X \in \mathcal{R}_1 | \Pi_2) \\ &= \pi_1 P(X \in \mathcal{R}_2 | \Pi_1) + \pi_2 P(X \in \mathcal{R}_1 | \Pi_2) \quad (\text{事前確率 } P(\Pi_k) = \pi_k \text{ より}) \\ &= \pi_1 [1 - P(X \in \mathcal{R}_1 | \Pi_1)] + \pi_2 P(X \in \mathcal{R}_1 | \Pi_2) \\ & \quad (P(X \in \mathcal{R}_2 | \Pi_1) = 1 - P(X \in \mathcal{R}_1 | \Pi_1) \text{ より}) \\ &= \pi_1 - \pi_1 P(X \in \mathcal{R}_1 | \Pi_1) + \pi_2 P(X \in \mathcal{R}_1 | \Pi_2) \\ &= \pi_1 - \pi_1 \int_{\mathcal{R}_1} f_1(x) dx + \pi_2 \int_{\mathcal{R}_1} f_2(x) dx \quad (P(X \in \mathcal{R}_1 | \Pi_k) = \int_{\mathcal{R}_1} f_k(x) dx \text{ より}) \\ &= \pi_1 + \int_{\mathcal{R}_1} [\pi_2 f_2(x) - \pi_1 f_1(x)] dx \end{aligned}$$

と表せる。

ベイズ判別法

この誤判別確率を最小にするように領域 $\mathcal{R}_1$ を選ぶ。

第1項の π_1 は $\mathcal{R}_1$ の取り方に依存しない。第2項を最小にするには積分の中がマイナスになる領域で積分すればよいので

$$\mathcal{R}_1 = \{ x \in \mathbb{R} \mid \pi_2 f_2(x) - \pi_1 f_1(x) \leq 0 \} ,$$

すなわち $\mathcal{R}_1 = \{ x \in \mathbb{R} \mid \pi_2 f_2(x) \leq \pi_1 f_1(x) \}$ とすればよいことがわかる。
この領域を $\mathcal{R}_1^* = \{ x \in \mathbb{R} \mid \pi_2 f_2(x) \leq \pi_1 f_1(x) \}$ とする。

ベイズ判別法

$\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbb{R}$ かつ $\mathcal{R}_1 \cap \mathcal{R}_2 = \Phi$ であるので、 $\mathcal{R}_1$ が定まれば $\mathcal{R}_2$ も自動的に定まる。

よって、誤判別確率を最小にする判別ルール $\{\mathcal{R}_1^*, \mathcal{R}_2^*\}$ は

$$\mathcal{R}_1^* = \{ x \in \mathbb{R} \mid \pi_2 f_2(x) \leq \pi_1 f_1(x) \}$$

$$\mathcal{R}_2^* = \{ x \in \mathbb{R} \mid \pi_2 f_2(x) > \pi_1 f_1(x) \}$$

となる。 $\mathcal{R}_1^*$ と $\mathcal{R}_2^*$ は **ベイズ判別ルール** と呼ばれる。

ベイズ判別法

■ 正規母集団

ある確率変数 X が平均 μ , 分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うとは、 X が密度関数

$$\phi(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

を持つこと。この時 $P(a \leq X \leq b)$ は

$$P(a \leq X \leq b) = \int_a^b \phi(x | \mu, \sigma^2) dx$$

で与えられる。

ベイズ判別法

■ 正規母集団の2群判別

ある個体 i は2つの群のどちらかに属するとする。この個体について、1次元のデータ x_i が与えられた時に、この個体が2つの群のどちらに属するかを判別する問題を考える。

この個体が、群1に従っている時は $x_i \sim \Pi_1 = N(\mu_1, \sigma^2)$ であり、群2に従っている時は $x_i \sim \Pi_2 = N(\mu_2, \sigma^2)$ であるとする(ここではひとまず分散は同じであるとする。また $\mu_1 \neq \mu_2$ と仮定する)。

ベイズ判別法

先ほどの2群の場合のベイズ判別ルールでは、領域 $\mathcal{R}_1$ は $\pi_2 f_2(x) - \pi_1 f_1(x)$ が 0 以下になる領域。 $f_k(x) = \phi(x | \mu_k, \sigma^2)$ なので、これは

$$\pi_2 f_2(x) - \pi_1 f_1(x) \leq 0$$

$$\Leftrightarrow \pi_2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu_2)^2}{2\sigma^2}\right] - \pi_1 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu_1)^2}{2\sigma^2}\right] \leq 0$$

$$\Leftrightarrow \frac{1}{\sqrt{2\pi\sigma^2}} \left\{ \pi_2 \exp\left[-\frac{(x - \mu_2)^2}{2\sigma^2}\right] - \pi_1 \exp\left[-\frac{(x - \mu_1)^2}{2\sigma^2}\right] \right\} \leq 0$$

となる x の領域である。

ベイズ判別法

これはさらに

$$\pi_2 \exp\left[-\frac{(x - \mu_2)^2}{2\sigma^2}\right] - \pi_1 \exp\left[-\frac{(x - \mu_1)^2}{2\sigma^2}\right] \leq 0$$

$$\Leftrightarrow \pi_2 \exp\left[-\frac{(x - \mu_2)^2}{2\sigma^2}\right] \leq \pi_1 \exp\left[-\frac{(x - \mu_1)^2}{2\sigma^2}\right]$$

$$\Leftrightarrow \frac{\pi_2}{\pi_1} \leq \exp\left[-\frac{(x - \mu_1)^2}{2\sigma^2} + \frac{(x - \mu_2)^2}{2\sigma^2}\right] \Leftrightarrow \log\left(\frac{\pi_2}{\pi_1}\right) \leq \frac{(x - \mu_2)^2 - (x - \mu_1)^2}{2\sigma^2}$$

$$\Leftrightarrow 2\sigma^2 \log\left(\frac{\pi_2}{\pi_1}\right) \leq 2(\mu_1 - \mu_2)x + \mu_2^2 - \mu_1^2$$

と書き直せる。この不等式を満たす x の領域が $\mathcal{R}_1$ である。

ベイズ判別法

例えば、 $\pi_1 = \pi_2 = 1/2$ の時、この式は

$$\mu_1^2 - \mu_2^2 \leq 2(\mu_1 - \mu_2)x \Leftrightarrow (\mu_1 - \mu_2)(\mu_1 + \mu_2) \leq 2(\mu_1 - \mu_2)x$$

となる。よって $\mu_1 - \mu_2 > 0$ の場合は $\mathcal{R}_1$ は

$$\frac{\mu_1 + \mu_2}{2} \leq x$$

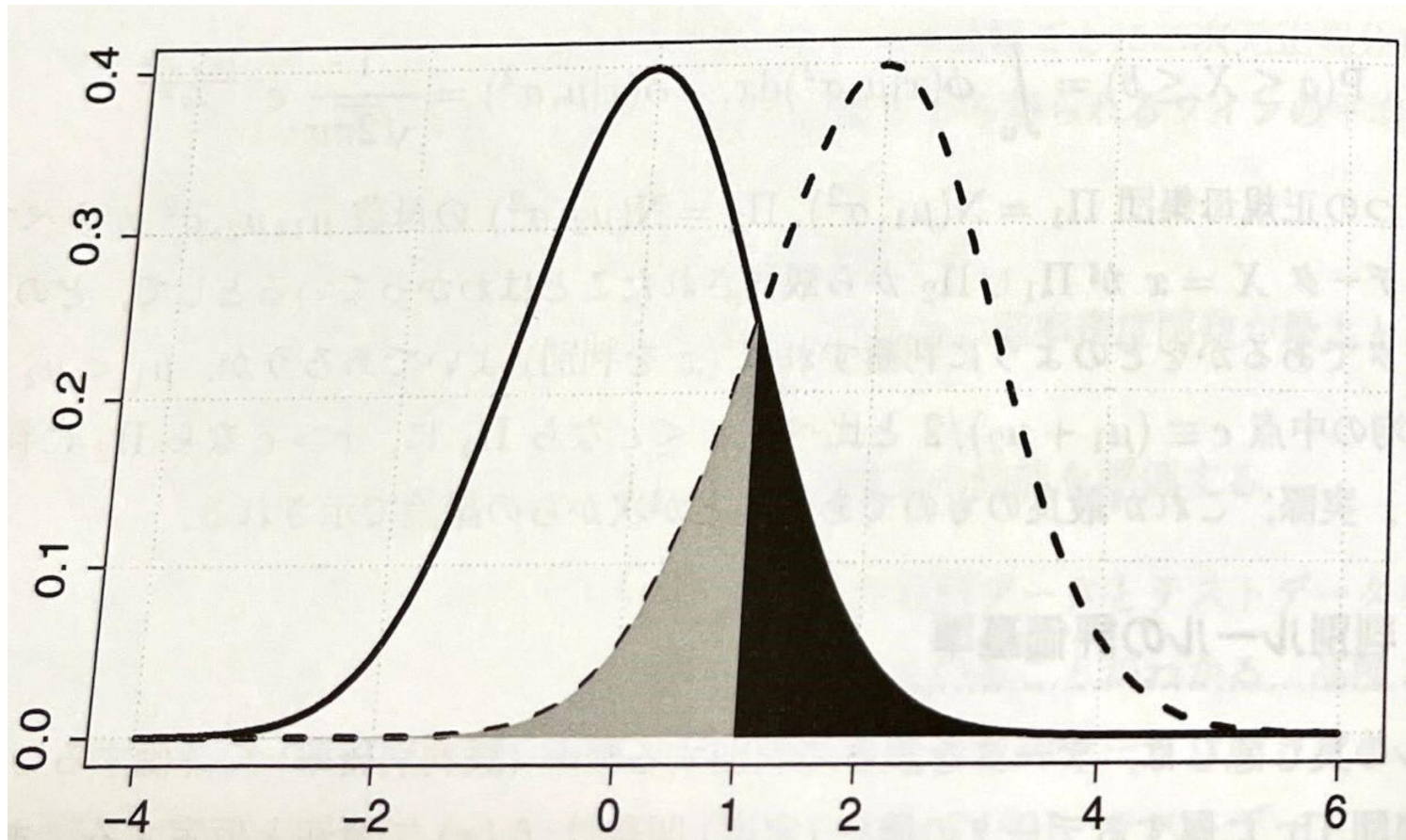
を満たす領域となり、 $\mu_1 - \mu_2 < 0$ の場合は $\mathcal{R}_1$ は

$$\frac{\mu_1 + \mu_2}{2} \geq x$$

を満たす領域となる。 $\mu_1 - \mu_2$ の符号によって $\mathcal{R}_1$ の領域が異なることに注意。また σ^2 に依存しなくなることに注意。

ベイズ判別法

$\Pi_1 = N(\mu_1, 1)$, $\Pi_2 = N(\mu_2, 1)$ である正規母集団に対して、事前確率が共に $\pi_1 = \pi_2 = 0.5$ であり、 $\mu_1 = 0$ および $\mu_2 = 2$ の場合を図示したものである。



ベイズ判別法

■ データが d 次元の場合

判別のために与えられるデータ $\mathbf{x}$ が d 次元ベクトル $\mathbf{x} = [x_1, x_2, \dots, x_d]^T$ の場合を考える。この場合、ベイズルールは多変量分布の密度関数を用いて以下のようになる。

$\mathbf{x}$ は密度関数 $f_1(\mathbf{x})$ を持つ分布 Π_1 , または密度関数 $f_2(\mathbf{x})$ を持つ分布 Π_2 からの実現値とする。 $\mathbf{x}$ が Π_k ($k=1,2$) からの実現値である事前確率を π_k とする。この時、ベイズルール $\{\mathcal{R}_1^*, \mathcal{R}_2^*\}$ は、先ほど同様、

$$\mathcal{R}_1^* = \{ \mathbf{x} \in \mathbb{R} \mid \pi_2 f_2(\mathbf{x}) \leq \pi_1 f_1(\mathbf{x}) \},$$

$$\mathcal{R}_2^* = \{ \mathbf{x} \in \mathbb{R} \mid \pi_2 f_2(\mathbf{x}) > \pi_1 f_1(\mathbf{x}) \},$$

となることを示すことができる。密度関数が1変量から多変量に変わったただけであとは同じである。

ベイズ判別法

■ 事後確率

Y を個体が Π_k ($k=1, 2$) に属するときは $Y = k$ となる変数とし、**ラベル**とよぶことにする。

データとして $\mathbf{X} = \mathbf{x}$ が観測された時に $Y = k$ である確率はベイズの定理より

$$P(Y = k \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})}, k = 1, 2$$

となる(とあるが、ここでは Y は離散変数で $\mathbf{x}$ が連続変数なので、実際にはやや込みいった議論が必要だが、ここでは適当にごまかしている)。
この確率を($Y = k$ となる)**事後確率**とよぶ。

ベイズ判別法

後に詳しく見るようにベイズルールは事後確率を最大化している。よってベイズルールは**事後確率最大化ルール**とも言える。

事後確率の分子と分母を $\pi_k f_k(\mathbf{x})$ で割ると

$$P(Y = k | \mathbf{X} = \mathbf{x}) = \frac{1}{1 + \lambda_j g_j(\mathbf{x})}, \quad g_j(\mathbf{x}) = \frac{f_j(\mathbf{x})}{f_k(\mathbf{x})}, \quad \lambda_j = \frac{\pi_j}{\pi_k}, \quad j, k = 1, 2, j \neq k,$$

となる。 $g_j(\mathbf{x})$ と λ_j はそれぞれ密度関数と事前確率の比なので、これはつまり、事後確率は事前確率の比と密度関数の比が分かれば計算でき、**事前確率や密度関数の値自体には依存していない**ことを意味している。

ベイズ判別法

■ 誤判別確率を最小化するというルール of 注意点

ベイズ判別法では誤判別確率を最小化するという基準であるが、2群の場合、判別間違いの重要度が異なる場合がある。

例えば、ある患者を

群1: 糖尿病 か 群2: 糖尿病でない

と判別する問題を考える。この場合、間違って群1とするよりも間違って群2とする間違いの方が重大な間違いとなる。

このように間違いの重要度が異なる時、損失を加重平均するという方法が考えられる。これは事前確率 π_k が加重されることになる。

ベイズ判別法

■ 線形判別 - ベイズ判別法の特例ケース

ベイズ判別において多変量分布を多変量正規分布、密度関数が

$$\phi(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right\}, k = 1, 2,$$

で与えられる分布(期待値ベクトル $\boldsymbol{\mu}_k = [\mu_{k,1}, \dots, \mu_{k,d}]^T$, 分散共分散行列 $\boldsymbol{\Sigma}$)、を用いた場合(**$\boldsymbol{\Sigma}$ が共通であることに注意**)、**線形判別**と呼ばれる以下の判別基準が得られる。

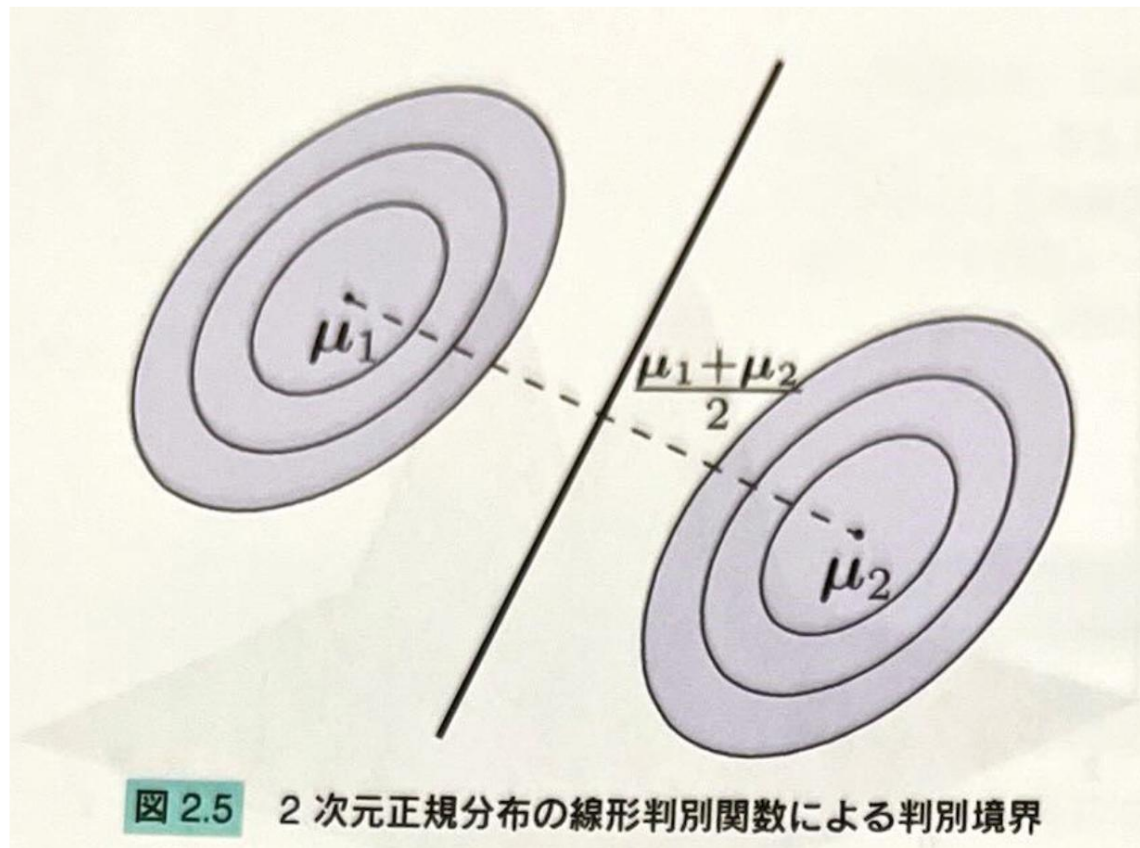
$$\begin{aligned} \mathcal{R}_1^* &= \{ \mathbf{x} \in \mathbb{R}^d \mid L(\mathbf{x}) + c \geq 0 \}, \quad L(\mathbf{x}) = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ \mathcal{R}_2^* &= \{ \mathbf{x} \in \mathbb{R}^d \mid L(\mathbf{x}) + c < 0 \}, \end{aligned}$$

ここで $c = \log(\pi_1 / \pi_2)$ は事前確率 π_k に依存して決まる定数である。

これは例えば、 $\pi_1 = \pi_2 = 1/2$ の時、 $c = \log[(1/2) / (1/2)] = 0$ となる。

ベイズ判別法

線形判別の判別境界は以下ようになる。



ベイズ判別法

■ マハラノビス距離と誤判別確率

このベイズ判別法（線形判別）における誤差判別確率を求めてみよう。
まず**マハラノビス距離**と呼ばれる距離を定義する。

2つの正規母集団 $\Pi_k = N_d(\mu_k, \Sigma)$, $k=1, 2$ の**マハラノビス距離**は

$$\Delta(\mu_1, \mu_2; \Sigma) = \sqrt{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)}$$

と定義される。

この時、線形判別の誤判別確率は以下で与えられる。

ベイズ判別法

■ 定理2.3

事前確率 π_k の正規母集団 $\Pi_k = N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k=1,2$ について、線形判別の誤判別確率は

$$e(\Delta; \pi_1) = \pi_1 \Phi\left(-\frac{\Delta}{2} - \frac{c}{\Delta}\right) + \pi_2 \Phi\left(-\frac{\Delta}{2} + \frac{c}{\Delta}\right)$$

ただし、 $\Delta = \Delta(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{\Sigma})$, $c = \log(\pi_1/\pi_2)$ および $\Phi(x)$ は標準正規分布の分布関数。特に $\pi_1 = 0.5$ の時は $e(\Delta; 1/2) = \Phi(-\Delta/2)$ に等しい。

ベイズ判別法

■ 2 次判別関数

先ほどは共通の分散共分散行列 Σ を持つ2つ正規母集団の判別について考えた。ここでは必ずしも共通の Σ を持たない場合のベイズ判別ルールを考える。以下のようになる。

$$\mathcal{R}_1^* = \{ \mathbf{x} \in \mathbb{R}^d \mid \Delta^2(\mathbf{x}, \boldsymbol{\mu}_2; \boldsymbol{\Sigma}_2) - \Delta^2(\mathbf{x}, \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1) + c^* \geq 0 \},$$

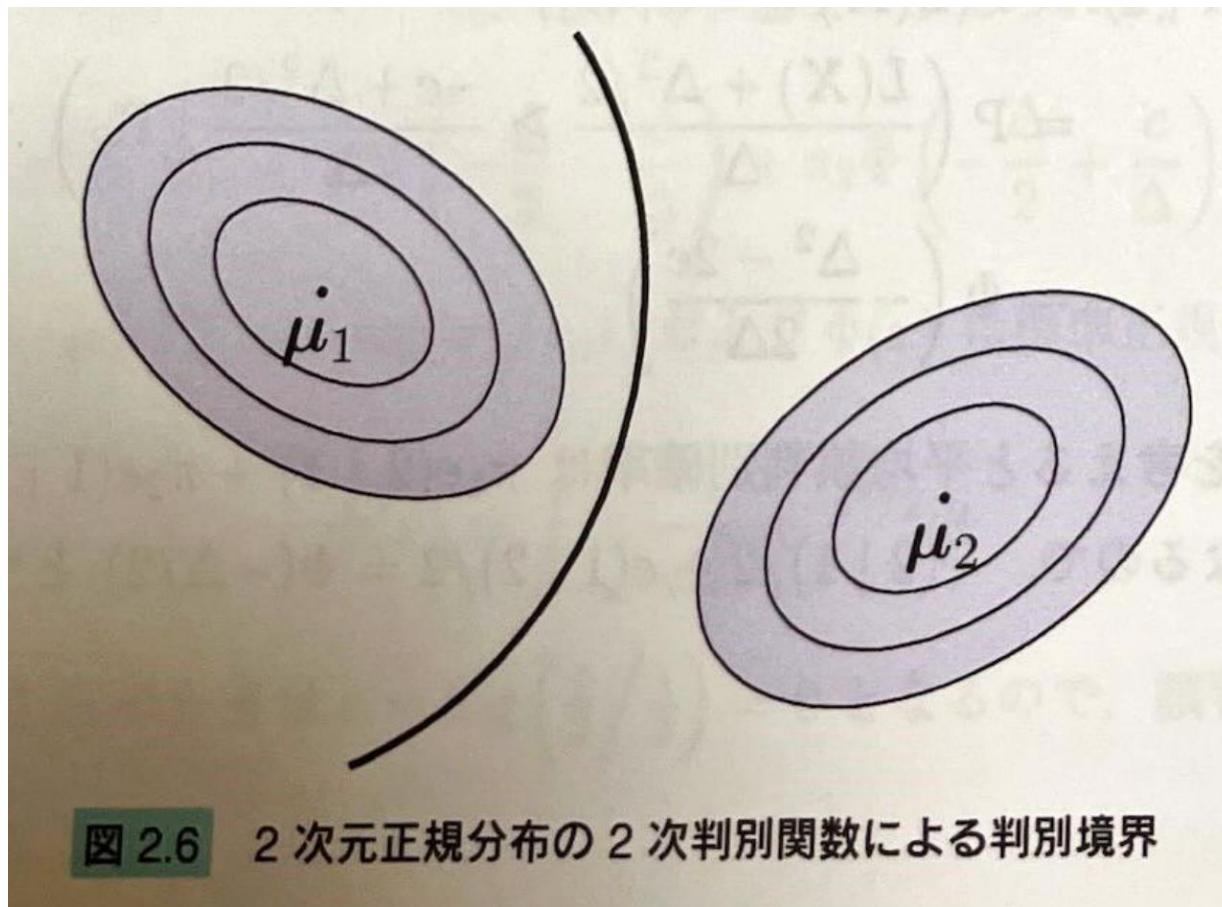
$$\mathcal{R}_2^* = \{ \mathbf{x} \in \mathbb{R}^d \mid \Delta^2(\mathbf{x}, \boldsymbol{\mu}_2; \boldsymbol{\Sigma}_2) - \Delta^2(\mathbf{x}, \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1) + c^* < 0 \},$$

ここで、 $c^* = \log(|\boldsymbol{\Sigma}_2| / |\boldsymbol{\Sigma}_1|) + 2\log(\pi_1 / \pi_2)$ である。

$Q(\mathbf{x}) = \Delta^2(\mathbf{x}, \boldsymbol{\mu}_2; \boldsymbol{\Sigma}_2) - \Delta^2(\mathbf{x}, \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1)$ とすると、上記のルールは $Q(\mathbf{x}) + c^*$ の符号によって判別する。 $Q(\mathbf{x})$ は2次判別関数と呼ばれ、この判別手法を **2次判別** という。

ベイズ判別法

2次判別の判別境界は以下のようなになる。



ベイズ判別法

■ 2 次判別関数の誤判別確率の上限

2次判別分析の誤判別確率は明示的には表現できないが、その上限は以下のチャーノフ上限によって与えられる。

$$\begin{aligned} \text{誤判別確率} \leq & \sqrt{\pi_1 \pi_2} \left[\frac{|\boldsymbol{\Sigma}_1| |\boldsymbol{\Sigma}_2|}{|(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) / 2|} \right]^{1/4} \\ & \times \exp \left[-\frac{1}{8} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \right] \end{aligned}$$

ベイズ判別法

■ 多群の場合のベイズ判別法

ここまでは2群に判別する問題を考えてきたが、群の数が一般に $g (\geq 3)$ の場合はどのようにしたらよいだろうか？

以下のような状況を考える(2群の場合の自然な拡張)。

- (1) 母集団 Π_k , $k=1, \dots, g$ に属する d 次元データ $\mathbf{x} = [x_1, x_2, \dots, x_d]^T$ の (d 変量)確率密度関数は $f_k(\mathbf{x})$ で既知であるとする。
- (2) データ $\mathbf{x}$ が Π_k から得られる事前確率は $\pi_k > 0$, $k = 1, 2, \dots, g$ であるとする。

ベイズ判別法

この時 $\mathbf{x}$ がとりうる領域 $\mathbb{R}^d$ を

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \dots \cup \mathcal{R}_g = \mathbb{R}^d, \quad \mathcal{R}_i \cap \mathcal{R}_j = \Phi, \quad i \neq j,$$

のような、互いに交わらない g 個の領域 $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_g$ に分割し、 $\mathbf{x} \in \mathcal{R}_k$ であれば、 $\mathbf{x}$ は Π_k に属すると判別する。

これを判別ルール $\{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_g\}$ とする。

実際には、 $g - 1$ 個の領域を決めれば残り1つは自動的に決まることに注意。
この時、以下の定理が得られる。

ベイズ判別法

定理2.5: 誤判別確率を最小とする判別ルール $\{\mathcal{R}_1^*, \mathcal{R}_2^*, \dots, \mathcal{R}_g^*\}$ (ベイズルール)は

$$\mathcal{R}_k^* = \left\{ \mathbf{x} \in \mathbb{R}^d \mid k = \operatorname{argmax}_{\ell=1,2,\dots,g} \{ \pi_\ell f_\ell(\mathbf{x}) \} \right\}, k = 1, 2, \dots, g$$

で与えられる。

この判別ルールでは、 $\mathbf{x}$ が群 k と判別される $\mathbf{x}$ の領域 $\mathcal{R}_k^*$ は $\pi_k f_k(\mathbf{x})$ が最も大きくなる、すなわち、すべての $j \neq k$ に対して、 $\pi_k f_k(\mathbf{x}) \geq \pi_j f_j(\mathbf{x})$ となる $\mathbf{x}$ の領域となっている。

実際に判別する際には、与えられた $\mathbf{x}$ に対して $\pi_\ell f_\ell(\mathbf{x})$, $\ell = 1, \dots, g$ を計算し、一番大きくなる k に属すると判断する。

ベイズ判別法

この領域 $\mathcal{R}_\ell^*$, $\ell=1,\dots,g$ が判別ルールを満足するか考えてみる。

$\pi_\ell f_\ell(\mathbf{x})$, $\ell=1,\dots,g$ に対して、所与の(どのような) $\mathbf{x}$ の値に対しても、必ずどれかの k において、 $\pi_k f_k(\mathbf{x})$ が他のすべての $k \neq j$ に対して $\pi_k f_k(\mathbf{x}) \geq \pi_j f_j(\mathbf{x})$ となる。よって最初の条件(領域の和集合が $\mathbb{R}^d$)は満たされる。

次に等号が成立するような $\mathbf{x}$ の値に対しては、どの k に属するかあらかじめ決めておくとなれば、2番目の条件(領域は互いに排反)も満たされる(ただし、その場合は決め方によって条件を満足するルールが複数存在する)。

よって、この判別ルールは2つの条件を満足している。

ベイズ判別法

■ 定理2.5 の証明

判別ルール $\{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_g\}$ による誤判別確率は

$$\begin{aligned}\sum_{k=1}^g \sum_{\ell \neq k} \pi_k P(X \in R_\ell | \Pi_k) &= \sum_{k=1}^g \pi_k \sum_{\ell \neq k} P(X \in R_\ell | \Pi_k) \\ &= \sum_{k=1}^g \pi_k [1 - P(X \in R_k | \Pi_k)] \\ &= \sum_{k=1}^g \pi_k \left[1 - \int_{\mathcal{R}_k} f_k(\mathbf{x}) d\mathbf{x} \right] \\ &= 1 - \sum_{k=1}^g \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x}\end{aligned}$$

で与えられる。

ベイズ判別法

■ 定理2.5 の証明

これを $\{\mathcal{R}_1, \dots, \mathcal{R}_g\}$ について最小化するには $\sum_{k=1}^g \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x}$ を最大にする $\{\mathcal{R}_1, \dots, \mathcal{R}_g\}$ を選べばよいが、先ほどのベイズルールのもとでは、

$$\sum_{k=1}^g \int_{\mathcal{R}_k^*} \pi_k f_k(\mathbf{x}) d\mathbf{x} \geq \sum_{k=1}^g \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x}$$

が成り立つので、ベイズルールの下でこの部分が最大化されていることがわかる。よって定理2.5が成り立つ。

ベイズ判別法

$\sum_{k=1}^g \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x}$ は任意の j に対して

$$\begin{aligned} \sum_{k=1}^g \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x} &= \sum_{k \neq j} \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x} + \int_{\mathbb{R}^d - \bigcup_{k \neq j} \mathcal{R}_k} \pi_j f_j(\mathbf{x}) d\mathbf{x} \\ &= \sum_{k \neq j} \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x} + \left[\int_{\mathbb{R}^d} \pi_j f_j(\mathbf{x}) d\mathbf{x} - \int_{\bigcup_{k \neq j} \mathcal{R}_k} \pi_j f_j(\mathbf{x}) d\mathbf{x} \right] \\ &= \sum_{k \neq j} \int_{\mathcal{R}_k} \pi_k f_k(\mathbf{x}) d\mathbf{x} + \left[\pi_j - \sum_{k \neq j} \int_{\mathcal{R}_k} \pi_j f_j(\mathbf{x}) d\mathbf{x} \right] \\ &= \pi_j + \sum_{k \neq j} \int_{\mathcal{R}_k} [\pi_k f_k(\mathbf{x}) - \pi_j f_j(\mathbf{x})] d\mathbf{x} \end{aligned}$$

と書き直せる。これは $g=2$ の時に 2群の場合に使用した表現になる。

ベイズ判別法

■ 実際のデータを使った正規母集団の判別法

実際の場面で、母集団分布が正規分布に従うと仮定できても、その平均や分散が既知であることはありえない。実際にはそれらの未知パラメーターをデータから推定する必要がある。

教師データから未知パラメーターを推定し、試験データによってそのパフォーマンスを評価することが通常である。

以下にワインのデータに対する適用例を紹介する。

ベイズ判別法

■ ワインデータへの適用例

ワインの種類を判別する問題を考える。

データとして 178本のワイン(品種1: 59本、品種2: 71本、品種3: 48本)について

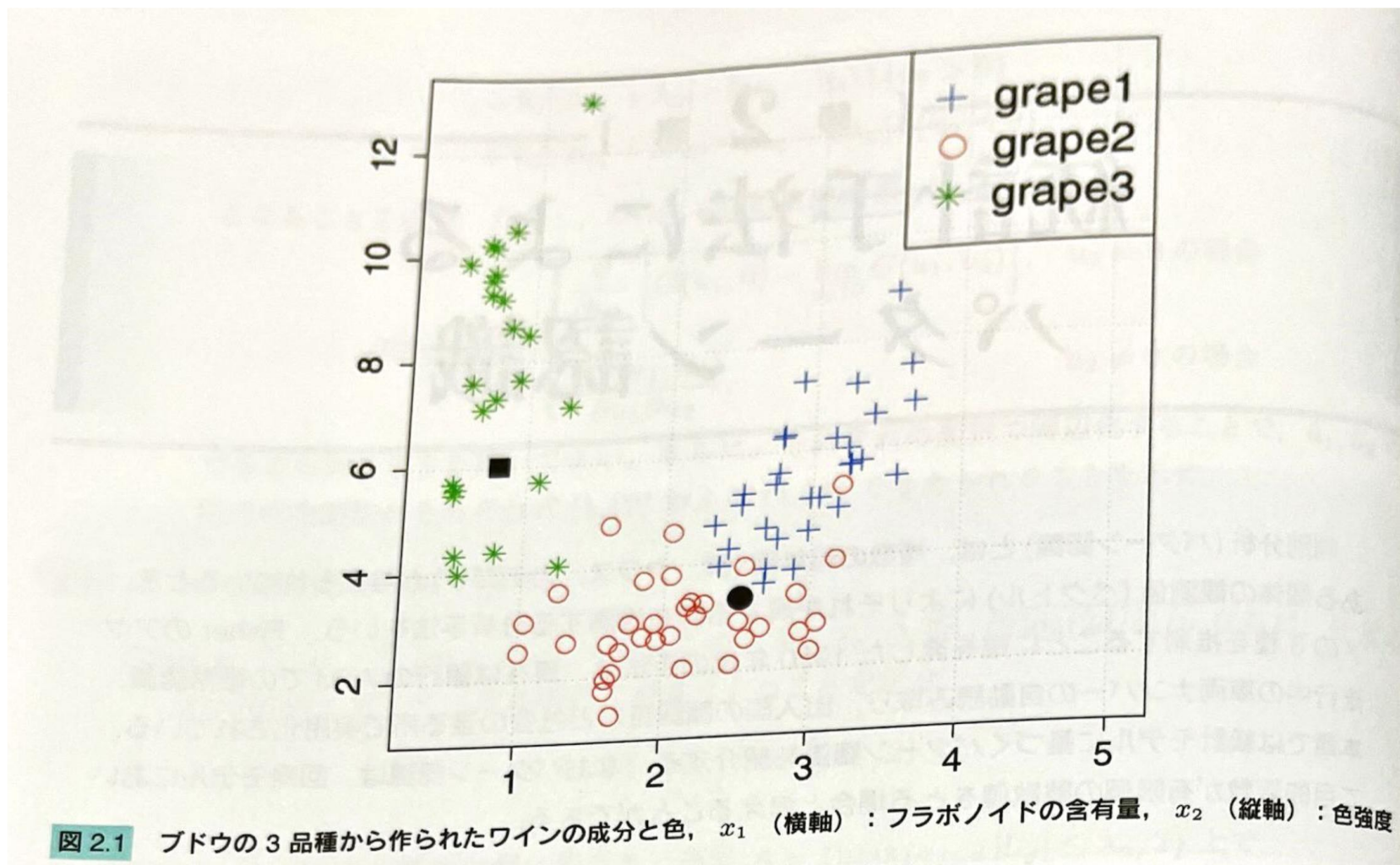
x_1 : フラボノイドの含有量、および x_2 : 色強度
のデータがある。

データは2次元 ($d = 2$) である。
またワインの種類が3種類であるため多群判別($g = 3$)になる。

この2次元データをプロットすると、

ベイズ判別法

■ ワインデータへの適用例



のようになる。

ベイズ判別法

■ ワインデータへの適用例

品種ごとに無作為に2分割し、教師サンプルとテストサンプルを以下のようにする

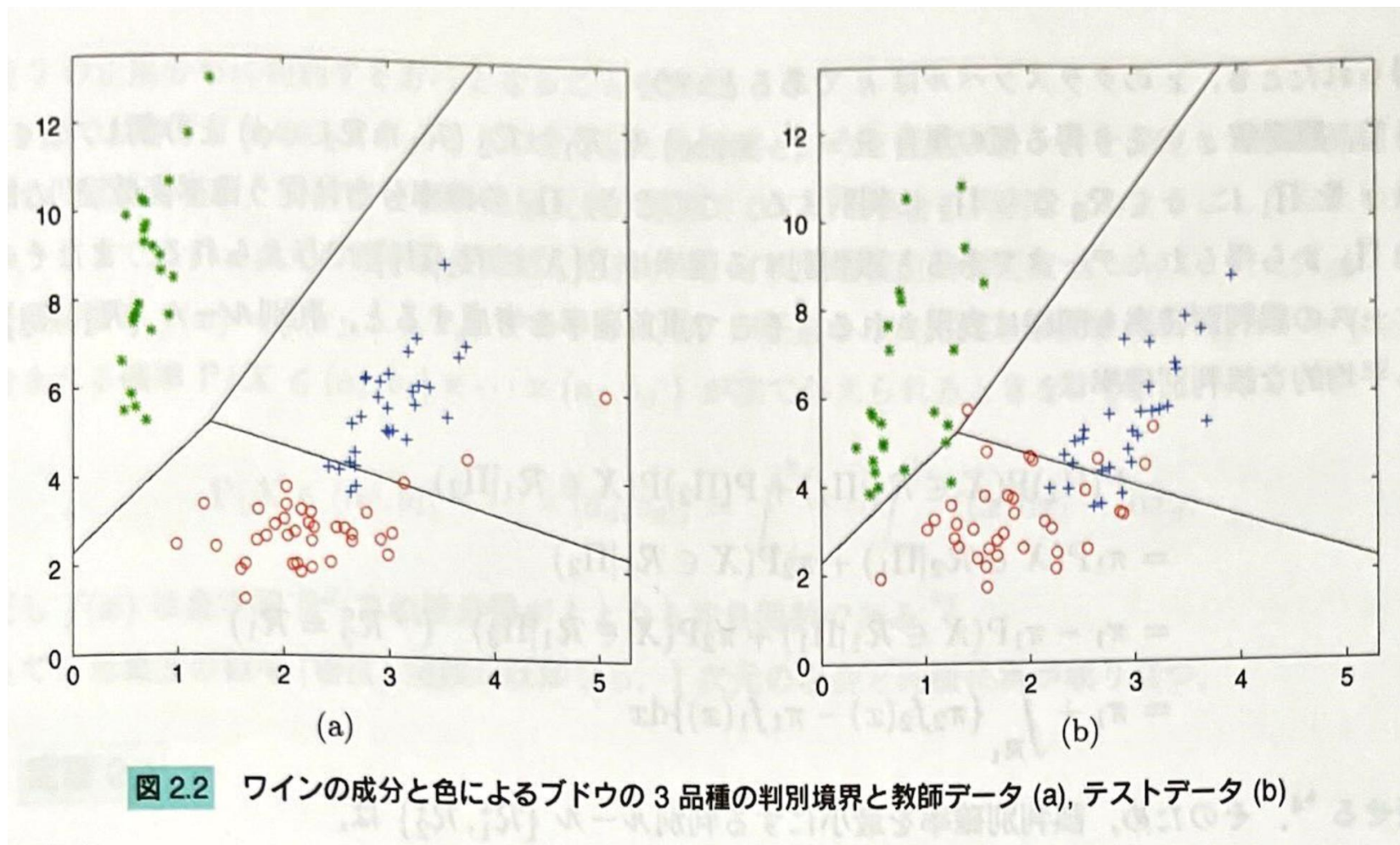
教師サンプル 品種1: 29 本、 品種2: 36本、 品種3: 24本

テストサンプル 品種1: 30本、 品種2: 35本、 品種3: 24 本

教師サンプルより μ と Σ を推定する。これらの推定値を用いて線形判別を適用し、教師データの判別結果とテストデータの判別結果は以下の図のようになる

ベイズ判別法

■ 線形判別の結果



ベイズ判別法

■ 線形判別の結果

判別境界は3本の直線で与えられている。

品種3の教師データは完全に正しく判別されているが、品種1 と品種2の教師データはそれぞれ 4 本と3本が誤って判別されている。よって教師エラー率は $7/89$ となる。

テストデータについて合計10本が誤って判別されている。テストエラー率は $10/89$ となっている。

ベイズ判別法

■ 2次判別の結果
2次判別の結果は

のようになる。

