

判別分析

1. 一変量の場合

第1群のデータ数 n_1 , 平均 μ_1 , 不偏分散 σ_1^2

第2群のデータ数 n_2 , 平均 μ_2 , 不偏分散 σ_2^2

とする。ここでは、 $\mu_1 < \mu_2$ とする。

2群間の平均 $\mu = \frac{\mu_1 + \mu_2}{2}$, 分散 $\sigma^2 = \frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1 + n_2 - 2}$ として、

(1) $\sigma_1^2 = \sigma_2^2$ とみなせる場合

入力値を x として、

$x \leq \mu$ なら第1群とみなす。そうでない場合は第2群とみなす。

第1群の平均値からの距離（マハラノビス距離）は

$$\frac{(x - \mu_1)^2}{\sigma^2}$$

第2群の平均値からの距離（マハラノビス距離）は

$$\frac{(x - \mu_2)^2}{\sigma^2}$$

で計算される。

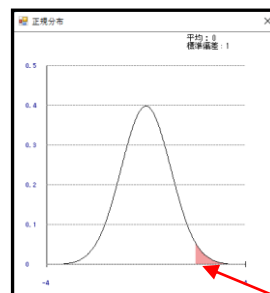
この場合の 誤判別の確率

P_1 : 第1群のデータなのに、第2群と誤判別される確率は

標準正規分布において、

$$\frac{x - \mu_1}{\sigma} > \frac{\mu - \mu_1}{\sigma}$$

となる 確率である。



$$\frac{\mu - \mu_1}{\sigma}$$

P_2 : 第2群のデータなのに、第1群と誤判別される確率は

は P_1 に等しい。

(2) $\sigma_1^2 \neq \sigma_2^2$ の場合

第1群の平均値からの距離（マハラノビス距離） d_1 は

$$d_1 = \frac{(x - \mu_1)^2}{\sigma_1^2}$$

第2群の平均値からの距離（マハラノビス距離） d_2 は

$$d_2 = \frac{(x - \mu_2)^2}{\sigma_2^2}$$

で計算される。

$d_1 < d_2$ ならば 第1群に、
そうでなければ、 第2群 とみなす。

この場合の 誤判別の確率 は 以下のように計算する。

P_1 : 第1群のデータなのに、第2群と誤判別される確率 は

第1群のデータ $\{x_i^{(1)}\}$ の中で、上記の判定手順で 第2群に属する
と誤判定される割合。

P_2 : 第2群のデータなのに、第1群と誤判別される確率 は

第2群のデータ $\{x_i^{(2)}\}$ の中で、上記の判定手順で 第1群に属する
と誤判定される割合。

2. 多変量の場合

変量数を P とする。

第1群のデータ数 n_1 として、 データを、

$$\begin{bmatrix} x_{1,1}^{(1)} & \cdots & x_{1,P}^{(1)} \\ \vdots & \ddots & \vdots \\ x_{n_1,1}^{(1)} & \cdots & x_{n_1,P}^{(1)} \end{bmatrix}$$

第2群のデータ数 n_2 として、 データを、

$$\begin{bmatrix} x_{1,1}^{(2)} & \cdots & x_{1,P}^{(2)} \\ \vdots & \ddots & \vdots \\ x_{n_2,1}^{(2)} & \cdots & x_{n_2,P}^{(2)} \end{bmatrix}$$

とする。

$$\{\mu\}^{(1)} = \begin{pmatrix} \mu_1^{(1)} \\ \vdots \\ \mu_P^{(1)} \end{pmatrix} \quad \mu_j^{(1)} = \frac{1}{n_1} \sum_{k=1}^{n_1} x_{k,j}^{(1)}, \quad j = 1 \sim P$$

$$\{\mu\}^{(2)} = \begin{pmatrix} \mu_1^{(2)} \\ \vdots \\ \mu_P^{(2)} \end{pmatrix} \quad \mu_j^{(2)} = \frac{1}{n_2} \sum_{k=1}^{n_2} x_{k,j}^{(2)}, \quad j = 1 \sim P$$

$$\Sigma_{(1)} = \begin{bmatrix} \sigma_{1,1}^{(1)} & \cdots & \sigma_{1,P}^{(1)} \\ \vdots & \ddots & \vdots \\ \sigma_{P,1}^{(1)} & \cdots & \sigma_{P,P}^{(1)} \end{bmatrix} \quad \sigma_{i,j}^{(1)} = \frac{1}{n_1-1} \sum_{k=1}^{n_1} (x_{k,i}^{(1)} - \mu_i^{(1)})(x_{k,j}^{(1)} - \mu_j^{(1)}) \quad i, j = 1 \sim P$$

$$\Sigma_{(2)} = \begin{bmatrix} \sigma_{1,1}^{(2)} & \cdots & \sigma_{1,P}^{(2)} \\ \vdots & \ddots & \vdots \\ \sigma_{P,1}^{(2)} & \cdots & \sigma_{P,P}^{(2)} \end{bmatrix} \quad \sigma_{i,j}^{(2)} = \frac{1}{n_2-1} \sum_{k=1}^{n_2} (x_{k,i}^{(2)} - \mu_i^{(2)})(x_{k,j}^{(2)} - \mu_j^{(2)}) \quad i, j = 1 \sim P$$

$$\{\mu\} = \frac{1}{2} \{ \{\mu\}^{(1)} + \{\mu\}^{(2)} \}$$

$$\{d\} = \{ \{\mu\}^{(1)} - \{\mu\}^{(2)} \}$$

1つの判別対象を $\{X\} = \begin{Bmatrix} x_1 \\ \vdots \\ x_P \end{Bmatrix} X$ として、2つのマハラノビス距離

$$D_1^2 = \{X - \mu^{(1)}\}^t \{\Sigma_{(1)}\}^{-1} \{X - \mu^{(1)}\}$$

$$D_2^2 = \{X - \mu^{(2)}\}^t \{\Sigma_{(2)}\}^{-1} \{X - \mu^{(2)}\}$$

を計算する。

(1) $\Sigma_{(1)} = \Sigma_{(2)}$ とみなせる場合

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \cdots & \sigma_{1,P} \\ \vdots & \ddots & \vdots \\ \sigma_{P,1} & \cdots & \sigma_{P,P} \end{bmatrix}, \quad \sigma_{ij} = \frac{(n_1 - 1) \sigma_{ij}^{(1)} + (n_2 - 1) \sigma_{ij}^{(2)}}{n_1 + n_2 - 2} \quad i, j = 1 \sim P$$

を分散の不偏推定量とみなす。

$$\begin{aligned} D_2^2 - D_1^2 &= \{X - \mu^{(2)}\}^t \{\Sigma_{(2)}\}^{-1} \{X - \mu^{(2)}\} - \{X - \mu^{(1)}\}^t \{\Sigma_{(1)}\}^{-1} \{X - \mu^{(1)}\} \\ &= \{X - \mu^{(2)}\}^t \{\Sigma\}^{-1} \{X - \mu^{(2)}\} - \{X - \mu^{(1)}\}^t \{\Sigma\}^{-1} \{X - \mu^{(1)}\} \\ &= 2\{X - \mu\}^t \{a\} \quad \text{但し} \quad \{a\} = \{\Sigma\}^{-1} \{d\} \end{aligned}$$

として、 $z = \frac{1}{2}(D_2^2 - D_1^2) = \{X - \mu\}^t \{a\}$ を計算し、

$z > 0$ ならば 第1群に判別

$z \leq 0$ ならば 第2群に判別

この場合の 誤判別の確率 は 以下のように計算する。

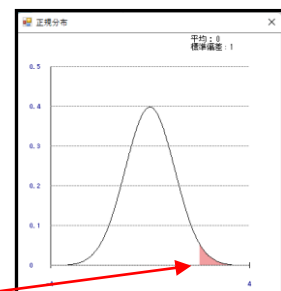
$$\Delta^2 = \{\mu^{(1)} - \mu^{(2)}\}^t \{\Sigma\}^{-1} \{\mu^{(1)} - \mu^{(2)}\} \quad \text{として、}$$

P_1 : 第1群のデータなのに、第2群と誤判別

される確率は標準正規分布において、右図の赤い部分、つまり $u = \Delta/2$ より右側の赤い部分の面積。

P_2 : 第2群のデータなのに、第1群と誤判別

される確率は P_1 に等しい。



$\frac{\Delta}{2}$

(2) $\Sigma_{(1)} \neq \Sigma_{(2)}$ の場合

$D_2^2 \geq D_1^2$ ならば 第1群に判別、 $D_2^2 < D_1^2$ ならば 第2群に判別。

この場合の 誤判別の確率 は 以下のように計算する。

P 1 : 第1群のデータ $\{x_i^{(1)}\}$ の内で、上記の判定手順で 第2群に属すると誤判定される割合。

P 2 : 第2群のデータ $\{x_i^{(2)}\}$ の内で、上記の判定手順で 第1群に属すると誤判定される割合。

次に、 $\Sigma_{(1)} = \Sigma_{(2)}$ であるかどうかの判断について

$$\Sigma_{(1)} = \begin{bmatrix} \sigma_{1,1}^{(1)} & \cdots & \sigma_{1,P}^{(1)} \\ \vdots & \ddots & \vdots \\ \sigma_{P,1}^{(1)} & \cdots & \sigma_{P,P}^{(1)} \end{bmatrix} \quad \sigma_{ij}^{(1)} = \frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (x_{k,i}^{(1)} - \mu_i^{(1)})(x_{k,j}^{(1)} - \mu_j^{(1)}) \quad i, j = 1 \sim P$$

$$\Sigma_{(2)} = \begin{bmatrix} \sigma_{1,1}^{(2)} & \cdots & \sigma_{1,P}^{(2)} \\ \vdots & \ddots & \vdots \\ \sigma_{P,1}^{(2)} & \cdots & \sigma_{P,P}^{(2)} \end{bmatrix} \quad \sigma_{ij}^{(2)} = \frac{1}{n_2 - 1} \sum_{k=1}^{n_2} (x_{k,i}^{(2)} - \mu_i^{(2)})(x_{k,j}^{(2)} - \mu_j^{(2)}) \quad i, j = 1 \sim P$$

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \cdots & \sigma_{1,P} \\ \vdots & \ddots & \vdots \\ \sigma_{P,1} & \cdots & \sigma_{P,P} \end{bmatrix}, \quad \sigma_{ij} = \frac{(n_1 - 1) \sigma_{ij}^{(1)} + (n_2 - 1) \sigma_{ij}^{(2)}}{n_1 + n_2 - 2} \quad i, j = 1 \sim P$$

として、

$$W = \left| \Sigma_{(1)} \right|^{\frac{n_1}{2}} * \left| \Sigma_{(2)} \right|^{\frac{n_2}{2}} \bigg/ \left| \Sigma \right|^{\frac{n_1 + n_2}{2}}$$

$$\chi_{W^2} = -2 * \log_e(W)$$

は 自由度 $\frac{P(P+1)}{2}$ の χ^2 分布に従う。

従って、上記 χ_{W^2} が 帰無仮説 ($\Sigma_{(1)} = \Sigma_{(2)}$) の棄却域にあれば、帰無仮説を棄却し、そうでなければ 帰無仮説を 棄却できない。