多変量解析　～判別分析　～

発表日　5月12日

発表者　　田島　勇樹

判別分析とは？

- 2つの母集団を設定して、あるサンプルがどちらの母集団に属するかを推測する方法
- 母集団への所属がわかっているサンプルとその変数の値に基づいて判別方式を構成する。
- 判別方式を用いて所属不明のサンプルがどちらの母集団に所属しているかを判別。

判別分析の具体例

右の表は、健常者と患者に対する
検査値ｘ１（量的変数）、ｘ２（量的変数）
のデータ

このデータから、

・患者かどうかを検査値1と検査値2より
判別できるか

・どちらの変数のほうが判別能力があるか

・判別できるとすればその精度は
どのくらいか

・検査値ｘ１、ｘ２に数値を入れたときどのように判別されるか

などを検討する。

サンプル No.	健常者・患者	検査値１ x1	検査値２ x2
1	健常者	50	15.5
2	健常者	69	18.4
3	健常者	93	26.4
4	健常者	76	22.9
5	健常者	88	18.6
6	患者	43	16.9
7	患者	56	21.6
8	患者	38	12.2
9	患者	21	16
10	患者	25	10.5

解析の流れ

1. 健常者を母集団[1]、患者を母集団[2]とする。
   母集団[1]と母集団[2]における変数の確率分布を母平均（ベクトル）が異なる正規分布と想定

   変数の値からそれぞれの母集団への距離としてマハラノビスの距離の2乗を求める。
   マハラノビスの距離の2乗値の小さい母集団へ判別するという判別方式を定める。

2. 誤判別の確率を求め、得られた判別方式の精度を評価。

3. 変数選択を行い、有用な変数を選択。

4. 得られた判別方式を利用して、どちらの母集団に属するのか不明なサンプルの判別を行う。

マハラノビスの距離と判別方式(1)

- ここでは量的変数が1つだけの場合を説明する。
- 「検査値1　χ1」を使用して説明する。

母集団1におけるx_1の確率分布→ $N(\mu_1^{[1]}, \sigma^2)$
母集団2におけるx_2の確率分布→ $N(\mu_1^{[2]}, \sigma^2)$
母分散は同じである。

サンプルの測定値　からそれぞれの母集団への距離
→マハラノビスの距離の2乗

マハラノビスの距離と判別方式(2)

- マハラノビスの距離の2乗の定義

$$D^{[1]2} = \frac{\left(x_1 - \mu_1^{[1]}\right)^2}{\sigma^2} \qquad D^{[2]2} = \frac{\left(x_1 - \mu_1^{[2]}\right)^2}{\sigma^2}$$

- 1次元正規分布 $N(\mu,\sigma^2)$の確率密度関数とマハラノビスの距離の2乗の対応関係

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{D^2}{2}\right\}$$

- マハラノビスの距離を用いた判別方式

$D^{[1]2} \le D^{[2]2}$ ⇔ 母集団[1]に属する

$D^{[1]2} > D^{[2]2}$ ⇔ 母集団[2]に属する

マハラノビスの距離と判別方式(3)

マハラノビスの距離の2乗の差

$$D^{[2]2} - D^{[1]2} = \frac{x_1^2 - 2\mu_1^{[2]}x_1 + \mu_1^{[2]2} - x_1^2 + 2\mu_1^{[1]}x_1 - \mu_1^{[1]2}}{\sigma^2}$$

$$= \frac{2(\mu_1^{[1]} - \mu_1^{[2]})}{\sigma^2}\left(x_1 - \frac{\mu_1^{[1]} + \mu_1^{[2]}}{2}\right)$$

これを2で割ったものを線形判別関数という。

$$z = \frac{D^{[2]2} - D^{[1]2}}{2} = \frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)}{\sigma^2}\left(x_1 - \overline{\mu}\right)$$

ただし、

$$\overline{\mu} = \frac{\mu_1^{[1]} + \mu_1^{[2]}}{2}$$

マハラノビスの距離と判別方式(4)

- 線形判別方式を書き換えたもの

$z \geq 0 <=> \quad D^{[1]2} \leq D^{[2]2} <=>$ 母集団[1]に属する

$z < 0 <=> \quad D^{[1]2} > D^{[2]2} <=>$ 母集団[2]に属する

サンプルの値を線形判別関数の推定式$\hat{z}$に代入して

得られた値　→スコア (判別得点)

マハラノビスの距離と判別方式(5)

- 誤判別の確率[1]

$\delta = \mu_1^{[1]} - \mu_1^{[2]} > 0$ と仮定する（逆向きの場合でも同様の結果を得る）

母集団[1]のサンプル→母集団[2]のサンプルと誤判別する確率

$x_1 \sim N(\mu_1^{[1]}, \sigma^2)$ →z の表現 $z \sim N(E(z), V(z))$ にする

$$E(z) = \frac{\mu_1^{[1]} - \mu_1^{[2]}}{\sigma^2}\left(\mu_1^{[1]} - \frac{\mu_1^{[1]} + \mu_1^{[2]}}{2}\right)$$

$$= \frac{\mu_1^{[1]} - \mu_1^{[2]}}{\sigma^2} \times \frac{\mu_1^{[1]} - \mu_1^{[2]}}{2}$$

$$= \frac{\delta^2}{2\sigma^2}$$

$$V(z) = V\left(\frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)}{\sigma^2}\left(x_1 - \overline{\mu}\right)\right)$$

$$= V\left(\frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)}{\sigma^2} x_1\right)$$

$$= \frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)}{\sigma^2} V(x_1) \frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)}{\sigma^2}$$

$$= \frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)^2}{\sigma^2} = \frac{\delta^2}{\sigma^2}$$

よって $z \sim N\left(\frac{\delta^2}{2\sigma^2}, \frac{\delta^2}{\sigma^2}\right)$ となる

マハラノビスの距離と判別方式(6)

- 誤判別の確率[2]

$z \sim N(\frac{\delta^2}{2\sigma^2}, \frac{\delta^2}{\sigma^2})$ のとき

$$\Pr(z < 0) = \Pr\left(\frac{z - \frac{\delta^2}{2\sigma^2}}{\frac{\delta}{\sigma}} < \frac{-\frac{\delta^2}{2\sigma^2}}{\frac{\delta}{\sigma}} \right) = \Pr(u < -\frac{\delta}{2\sigma}) = \Pr(u > \frac{\delta}{2\sigma})$$

ただし $u \sim N(0,1^2)$ である

母集団[2]のサンプル→母集団[1]のサンプルと誤判別する確率
同様に計算すると

$$x_1 \sim N(\mu_1^{[1]}, \sigma^2) \rightarrow z \sim N(-\frac{\delta^2}{2\sigma^2}, \frac{\delta^2}{\sigma^2})$$ となる

よって $\Pr(u = \frac{\delta}{2\sigma}) = 0$ だから $\Pr(z < 0) = \Pr(z \geq 0)$

マハラノビスの距離と判別方式(7)

- 判別表

右図より

健常者を患者と誤判別する割合：1/5＝0．2

患者を健常者と誤判別する割合：1/5＝0．2

データ結果	判別結果		計
	健常者	患者	
健常者	4	1	5
患者	1	4	5
計	5	5	１０

$\mu_1^{[1]}$と$\mu_1^{[2]}$の差が大きい→誤判別の確率は小さくなる

→判別が正確になる

2つの母平均のマハラノビスの距離の2乗

$$D_{x_1}^2([1],[2]) = \frac{\left(\mu_1^{[1]} - \mu_1^{[2]}\right)^2}{\sigma^2}$$

これを判別効率という

マハラノビスの距離と判別方式(8)

- 変数選択

→変数が判別に寄与しているかどうか検討

$$F_0 = \frac{\left(n^{[1]} + n^{[2]} - p - r - 1\right) n^{[1]} n^{[2]} \left\{ \hat{D}_{x(p+r)}([1],[2]) - \hat{D}_{x(p)}([1],[2]) \right\}}{r\left\{ \left(n^{[1]} + n^{[2]} - 2\right)\left(n^{[1]} + n^{[2]}\right) + n^{[1]} n^{[2]} \hat{D}^2_{x(p)}([1],[2]) \right\}}$$

ただし

$$\hat{D}^2_{x(p+r)}([1],[2]) = \hat{D}^2_{x_1 x_2 \cdots x_p x_{p+1} \cdots x_r}([1],[2])$$

$$\hat{D}^2_{x(p)}([1],[2]) = \hat{D}^2_{x_1 x_2 \cdots x_p}([1],[2])$$

p:変数を追加する前の変数の個数

r:追加する変数の個数

F_0はr個の追加する変数が判別に寄与しないというもとで $F(r, n^{[1]} + n^{[2]} - p - r - 1)$ に従う

ただし「F_0値が2以上であるならその変数が判別に寄与する」というのが目安

マハラノビスの距離と判別方式(9)

マハラノビスの距離の2乗について

$$D^2_{x_1 x_2 \cdots x_p}([1],[2]) = \delta \hat{\Sigma}^{-1} \delta$$

ただし

$$\delta = \mu^{[1]} - \mu^{[2]} = \begin{bmatrix} \mu_1^{[1]} - \mu_1^{[2]} \\ \mu_2^{[1]} - \mu_2^{[2]} \\ \bullet \\ \bullet \\ \mu_p^{[1]} - \mu_p^{[2]} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1p} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{p1} & \sigma^{p2} & \cdots & \sigma^{pp} \end{bmatrix}$$

分子にある2つの判別効率の差が大きい→変数が多く取り入れられている。

マハラノビスの距離と判別方式(１０)

･変数が2個以上の場合の解析方法
　以下では変数が2個の場合を説明する。

データおよびスコアと判別結果

サンプルNo	健常者･患者	検査値1x1	検査値2x2	スコア	判別結果
1	健常者	50	15.5	-0.516	患者
2	健常者	69	18.4	2.809	健常者
3	健常者	93	26.4	5.561	健常者
4	健常者	76	22.9	2.888	健常者
5	健常者	88	18.6	7.037	健常者
6	患者	43	16.9	-2.566	患者
7	患者	56	21.6	-1.197	患者
8	患者	38	12.2	-2.126	患者
9	患者	21	16.0	-7.237	患者
10	患者	25	10.5	-4.496	患者

変数は「検査値1 x1」、「検査値2 x2」の2つである。

マハラノビスの距離と判別方式(１１)

母集団[1]におけるx=$[x_1, x_2]$の確率分布として

$$\mu^{[1]} = \begin{bmatrix} \mu_1^{[1]} \\ \mu_1^{[2]} \end{bmatrix} \quad 、 \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

の正規分布 $N(\mu^{[1]}, \Sigma)$ を仮定する。

母集団[2]におけるx=$[x_1, x_2]$の確率分布として

$$\mu^{[1]} = \begin{bmatrix} \mu_1^{[1]} \\ \mu_1^{[2]} \end{bmatrix} \quad 、 \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

の正規分布 $N(\mu^{[2]}, \Sigma)$ を仮定する。Σ は同じであるとする。

ここで

$$\Sigma^{-1} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{bmatrix}$$

マハラノビスの距離と判別方式(１２)

- マハラノビスの距離の２乗

$$
\begin{aligned}
D^{[k]2} &= (x-\mu^{[k]})'\Sigma^{-1}(x-\mu^{[k]}) \\
&= \left[x_1-\mu_1^{[k]}, x_2-\mu_2^{[k]}\right]\begin{bmatrix}\sigma^{11} & \sigma^{12}\\ \sigma^{21} & \sigma^{22}\end{bmatrix}\begin{bmatrix}x_1-\mu_1^{[k]}\\ x_2-\mu_2^{[k]}\end{bmatrix} \\
&= (x_1-\mu_1^{[k]})^2\sigma^{11} + \left(x_2-\mu_2^{[k]}\right)^2\sigma^{22} + 2\left(x_1-\mu_1^{[k]}\right)\left(x_2-\mu_2^{[k]}\right)\sigma^{12} \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2}(x_i-\mu_i^{[k]})(x_j-\mu_j^{[k]})\sigma^{ij} \qquad (k=1,2)
\end{aligned}
$$

$D^{[k]2}$はxから母集団[k]の母平均ベクトル$\mu^{[k]}$までの距離を測る量

Σが単位行列ならユークリッド距離になる
通常はΣは単位行列でないのでマハラノビスの汎距離という

マハラノビスの距離と判別方式(１３)

- 2次元正規分布 $N(\mu,\Sigma)$ の確率密度関数とマハラノビスの距離の2乗の対応関係

$$f(x)=\frac{1}{2\pi\sqrt{|\Sigma|}}\exp\left\{-\frac{D^2}{2}\right\}$$

マハラノビスの距離を用いての判別方式

$D^{[1]2} \le D^{[2]2}$　　＜＝＞母集団[1]に属する

$D^{[1]2} > D^{[2]2}$　　＜＝＞母集団[2]に属する

マハラノビスの距離の差

$$D^{[2]2}-D^{[1]2}=2\left[\mu_1^{[1]}-\mu_1^{[2]},\mu_2^{[1]}-\mu_2^{[2]}\right]\begin{bmatrix}\sigma^{11} & \sigma^{12}\\ \sigma^{21} & \sigma^{22}\end{bmatrix}\begin{bmatrix}x_1-\overline{\mu}_1\\ x_2-\overline{\mu}_2\end{bmatrix}$$

ただし、

$$\overline{\mu}_1=\frac{\mu_1^{[1]}+\mu_1^{[2]}}{2},\overline{\mu}_2=\frac{\mu_2^{[1]}+\mu_2^{[2]}}{2}$$

マハラノビスの距離と判別方式(１４)

マハラノビスの距離の差を２で割った

$$z = \left[\mu_1^{[1]} - \mu_1^{[2]}, \mu_2^{[1]} - \mu_2^{[2]}\right] \begin{bmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{bmatrix} \begin{bmatrix} x_1 - \overline{\mu}_1 \\ x_2 - \overline{\mu}_2 \end{bmatrix}$$

線形判別関数という。

$z \geq 0$ <=> $D^{[1]2} \leq D^{[2]2}$ <=>母集団[1]に属する

$z < 0$ <=> $D^{[1]2} > D^{[2]2}$ <=>母集団[2]に属する

・誤判別の確率

$\delta = \mu_1^{[1]} - \mu_1^{[2]} > 0$ と定義する。

母集団[1]のサンプル→母集団[2]のサンプルと誤判別する確率

前回と同様に計算して $x \sim N(\mu^{[1]}, \Sigma)$のもとにして $z \sim N(\frac{\delta' \Sigma^{-1} \delta}{2}, \delta' \Sigma^{-1} \delta)$ とする

マハラノビスの距離と判別方式(１５)

母集団[2]に属すると誤判別する確率

$$\Pr(z<0)=\Pr\left(\frac{z-\delta'\Sigma^{-1}\delta/2}{\sqrt{\delta'\Sigma^{-1}\delta}}<\frac{-\delta'\Sigma^{-1}\delta/2}{\sqrt{\delta'\Sigma^{-1}\delta}}\right)$$
$$=\Pr\left(u<-\sqrt{\delta'\Sigma^{-1}\delta}/2\right)$$
$$=\Pr\left(u>\sqrt{\delta'\Sigma^{-1}\delta}/2\right)$$

となる。同様に母集団[2]に属するのに母集団[1]に属すると誤判別する確率

$x\sim N(\mu^{[2]},\Sigma)$ をもとにして $z\sim N(-\frac{\delta'\Sigma^{-1}\delta}{2},\delta'\Sigma^{-1}\delta)$ とする

$$\Pr(z\geq 0)=\Pr\left(\frac{z+\delta'\Sigma^{-1}\delta/2}{\sqrt{\delta'\Sigma^{-1}\delta}}\geq\frac{\delta'\Sigma^{-1}\delta/2}{\sqrt{\delta'\Sigma^{-1}\delta}}\right)$$
$$=\Pr\left(u\geq\sqrt{\delta'\Sigma^{-1}\delta}/2\right)$$

つまり

$$\Pr(z<0)=\Pr(z\geq 0)$$

マハラノビスの距離と判別方式(１６)

- 判別表を右に記す。

健常者を患者と誤判別する割合：1/5＝0．2

患者を健常者と誤判別する割合：0/5＝0

データ結果	判別結果		計
	健常者	患者	
健常者	4	1	5
患者	0	5	5
計	4	6	１０

2変数のとき、2つの母平均ベクトル間のマハラノビスの距離の2乗の定義→判別効率

$$D^2_{x_1x_2}([1],[2])(\mu^{[1]}-\mu^{[2]})'\Sigma^{-1}(\mu^{[1]}-\mu^{[2]})=\delta\Sigma^{-1}\delta$$

マハラノビスの距離と判別方式(１７)

- 多変数の変数選択

・判別効率（誤判別しにくいか）をそれぞれ調べる。

・それぞれの変数だけで F_0 値を求めて判別に寄与しているかを確かめる。

・上の２つの条件でもっとも効率の高い変数を判別方式に取り入れ、そのあとに残った変数の中で高い順番に変数を追加していく。

例　２変数の場合

「検査値１ｘ１」と「検査値２ｘ２」の場合で考えると・・・

$$x_1 : \hat{D}^2_{x_1}([1],[2]) = \frac{(\hat{\mu}_1^{[1]} - \mu_1^{[2]})^2}{\hat{\sigma}^2_{11}} = 6.106 \qquad x_2 : \hat{D}^2_{x_2}([1],[2]) = \frac{(\hat{\mu}_2^{[1]} - \mu_2^{[2]})^2}{\hat{\sigma}^2_{22}} = 1.302$$

「検査値１」のほうが「検査値２」よりも判別効率が大きい。

次に F_0 値を求める。

それぞれ　x_1**は** $F_0 = 15.27, x_2$**は** $F_0 = 3.255$**となる。**

マハラノビスの距離と判別方式(１８)

以上よりまず、「検査値1」を判別方式に取り入れる。
次に「検査値2」を判別方式に追加する価値があるか調べる。

検査値1があるためp=1となる。また、

$\hat{D}^2_{x(p)}([1],[2]) = \hat{D}^2_{x_1}([1],[2]) = 6.106$　$\hat{D}^2_{x(p+r)}([1],[2]) = \hat{D}^2_{x_1x_2}([1],[2]) = 7.068$

であるため、 x_2**の**F_0**値は**

$$F_0 = \frac{(5+5-1-1-1)5\times 5\{7.068-6.106\}}{1\{(5+5-2)(5+5)+5\times 5\times 6.106\}} = 0.724$$

よってを x_1**を取り込んだ後では、** x_2**は健常者と患者の判別 に寄与しない**
ということになる。