データ解析演習問題

2006.11.13

提出期限: 2006 年 11 月 16 日 (木) 12:00
提出場所: システム棟 5F レポート提出 BOX

学籍番号:

氏名:

A. 判別分析のまとめ

A.1.

判別分析では, 表 A.1 のようなデータを扱う. n 個の個体は, グループ P とグループ Q の 2 つのグループに分割されており, 個体番号 1 から m はグループ P に, 個体番号 $m+1$ から n まではグループ Q に属している.

表 A.1: 判別分析のデータ

No.	x	y	P/Q
1	x_1	y_1	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_m	y_m	P
$m+1$	x_{m+1}	y_{m+1}	Q
$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	x_j	y_j	Q
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	Q

判別分析の目的は, これらの 2 つのグループがなるべく遠ざかって見えるように, 合成変数

$$z = ax + by + c \tag{A.1}$$

を求めることである. z を x と y の関数として見るときは, これを線形判別関数と呼ぶ.

$n_P = m, n_Q = n - m$ とする. すなわち, n_P と n_Q は, それぞれ, グループ P とグループ Q に属する個体の数である. $\bar{z}_P$ と $\bar{z}_Q$ を

$$\bar{z}_P = \frac{1}{n_P}\sum_{i=1}^{m} z_i, \quad \bar{z}_Q = \frac{1}{n_Q}\sum_{j=m+1}^{n} z_j, \tag{A.2}$$

で定義する. すなわち, $\bar{z}_P$ はグループ P における z の平均であり, $\bar{z}_Q$ はグループ Q における z の平均である.

z の変動 S_T

$$S_T = \sum_{k=1}^{n} (z_k - \bar{z})^2 \tag{A.3}$$

は,

$$S_T = S_B + SW \tag{A.4}$$

と分解される. ここで, S_W と S_B は, それぞれ,

$$S_W = \sum_{i=1}^{m} (z_i - \bar{z}_P)^2 + \sum_{j=m+1}^{n} (z_j - \bar{z}_Q)^2 \tag{A.5}$$

$$S_B = n_P(\bar{z}_P - \bar{z})^2 + n_Q(\bar{z}_Q - \bar{z})^2 \tag{A.6}$$

と定義される. (なぜ, 式 (A.4) が成り立つのか復習しておきなさい.) S_B は 2 つのグループがどれだけ離れているかを表していると解釈できる. したがって, S_B は群間変動と呼ばれる. また, S_W は, グループ P 内での z の変動とグループ Q 内での z の変動の和である. したがって, S_W は群内変動と呼ばれる.

これらの観察から, 「2 つのグループがなるべく遠ざかって見える」ように, 合成変数 $z = ax + by + c$ を求めることは, 相関比

$$\eta^2 = \frac{S_B}{S_T} \tag{A.7}$$

を最大にするような a, b, c を求めることであることが分かる.

A.2.

S_T と S_B は, それぞれ, a, b を用いて,

$$S_T = n(a^2 s_x^2 + 2ab s_{xy} + s_y^2 b^2) \tag{A.8}$$

$$S_B = n_P\{a(\bar{x}_P - \bar{x}) + b(\bar{y}_P - \bar{y})\}^2 + n_Q\{a(\bar{x}_Q - \bar{x}) + b(\bar{y}_Q - \bar{y})\}^2 \tag{A.9}$$

というように書ける. (なぜそうなるのか復習しておきなさい.) ここで,

$$\bar{x}_P = \frac{1}{n_P}\sum_{i=1}^{m} x_i, \quad \bar{x}_Q = \frac{1}{n_Q}\sum_{j=m+1}^{n} x_j, \tag{A.10}$$

$$\bar{y}_P = \frac{1}{n_P}\sum_{i=1}^{m} y_i, \quad \bar{y}_Q = \frac{1}{n_Q}\sum_{j=m+1}^{n} y_j, \tag{A.11}$$

である.

したがって, 相関比 η^2 も

$$\eta^2 = \frac{n_P\{a(\bar{x}_P - \bar{x}) + b(\bar{y}_P - \bar{y})\}^2 + n_Q\{a(\bar{x}_Q - \bar{x}) + b(\bar{y}_Q - \bar{y})\}^2}{n(a^2 s_x^2 + 2ab s_{xy} + s_y^2 b^2)} \tag{A.12}$$

と書き直せる. さらに, $t = \frac{a}{b}$ と置けば,

$$\eta^2 = \frac{n_P\{t(\bar{x}_P - \bar{x}) + (\bar{y}_P - \bar{y})\}^2 + n_Q\{t(\bar{x}_Q - \bar{x}) + (\bar{y}_Q - \bar{y})\}^2}{n(t^2 s_x^2 + 2ts_{xy} + s_y^2)} \tag{A.13}$$

と書けるので, 方程式 $\frac{d\eta^2}{dt} = 0$ の 2 つの解のうちの一つが求める $t = \frac{a}{b}$ となる. (詳細はスライドを見よ).

z の分散 s_z^2 は 1 に等しいという仮定を置くと, a と b が決定される.

さらに, 直線 $0 = ax + by + c$ が $(\bar{x}_P, \bar{y}_P)$ と $(\bar{x}_Q, \bar{y}_Q)$ の中点を通るように c を決めれば, c も決定される.

A.3.

こうして求められた, 線形判別関数 $z = ax + by + c$ を用いて, 各個体が P と Q のいずれのグループに分類されるかを判定することができる. すなわち, もし

$$z_k = ax_k + by_k + c > 0 \tag{A.14}$$

ならば第 k 番目の個体は, P に属すると判定して,

$$z_k = ax_k + by_k + c < 0 \tag{A.15}$$

ならば第 k 番目の個体は, Q に属すると判定する. ($z_k = 0$ のときは, どちらとも言えない.)

B.

表 B.1 に示すような 2 変数 x, y に関する判別分析のデータが得られているとする. このデータに対して, 線形判別関数

$$z = ax + by + c \tag{B.1}$$

を考える. 以下の設問に答えよ.

B.1.

a, b, c を求めて, その結果を小数点以下第 3 位まで記入せよ.

$$a = -0.053, \quad b = 0.487 \quad c = 0.165 \tag{B.2}$$

B.2.

設問 1 で求められた a, b, c に対して, 変動 S_T, 群間変動 S_B, 群内変動 S_W, 相関比 η^2 を求め, その結果を小数点以下第 3 位まで記入せよ.

$$S_T = 10.000, \quad S_B = 5.527 \quad S_W = 4.473, \quad \eta^2 = 0.553 \tag{B.3}$$

表 B.1: x, y に関するデータ

No.	x	y	P/Q	z
1	28.0	5.3	P	1.26
2	64.0	7.9	P	0.62
3	35.0	5.5	P	0.99
4	42.0	7.0	P	1.35
5	72.0	6.5	P	-0.49
6	52.0	5.8	Q	0.23
7	72.0	7.1	Q	-0.20
8	64.0	3.3	Q	-1.63
9	90.0	7.1	Q	-1.16
10	80.0	6.4	Q	-0.97

B.3.

No. 1〜10 の各個体に対する判別得点 z_k を求めて, 表 B.1 の z の列に小数点以下第 2 位まで記入せよ. (あるいは, Excel の表を印刷したものを, 添付せよ.)

B.4.

判別的中率はいくつか? この結果はどのように評価できるか?
0.8. やや良い.

C.

本講義「データ解析」についての感想, 要望, 質問等があれば記せ.