Multivariate Analysis - 2nd Test

4 Feb 2021 - 1st Serie 2020/2021 -

Group I

1. $\underset{\sim}{x} \in \mathbb{R}^p$, $\underset{\sim}{x}_k = \frac{1}{n_k} \sum_{\underset{\sim}{x}_h \in C_k} x_h$

(a) $$\frac{1}{n_k} \sum_{\underset{\sim}{x}_h, \underset{\sim}{x}_{h'} \in C_k} \sum_{j=1}^{p} (x_{hj} - x_{h'j})^2 =$$

$$= \frac{1}{n_k} \sum_{\underset{\sim}{x}_h, \underset{\sim}{x}_{h'}} \sum_{j=1} (x_{hj} - \bar{x}_j + \bar{x}_j - x_{h'j})^2$$

$$= \frac{1}{n_k} \sum_{\underset{\sim}{x}_h, \underset{\sim}{x}_{h'}} \sum_{j=1} \left[(x_{hj} - \bar{x}_j)^2 + (x_{h'j} - \bar{x}_j)^2 + 2 (x_{hj} - \bar{x}_j)(x_{h'j} - \bar{x}_j) \right]$$

$$= \frac{1}{n_k} \sum_{\underset{\sim}{x}_h \in C_k} \sum_{j=1}^{p} n_k (x_{hj} - \bar{x}_j)^2 + \frac{1}{n_k} \sum_{\underset{\sim}{x}_{h'} \in C_k} \sum_{j=1}^{p} (x_{h'j} - \bar{x}_j)^2 n_k$$

$$+ 2 \sum_{j=1}^{p} \sum_{\underset{\sim}{x}_h \in C_k} (x_h - \bar{x}_j) \underbrace{\sum_{\underset{\sim}{x}_{h'} \in C_k} (x_{h'j} - \bar{x}_j)}_{=0} =$$

$$= 2 \sum_{\underset{\sim}{x}_h \in C_k} \sum_{j=1}^{p} (x_{hj} - \bar{x}_j)^2$$

QED

(b) Version A

2 (a) $C_1 = \{x_1, x_3, x_5\}$ and $C_2 = \{x_2, x_4, x_6\}$

$\bar{x}_{1.c_1} = \frac{-2-2+1}{3} = -1$ $\qquad$ $\bar{x}_{2.c_1} = \frac{-1-2+2}{3} = 1$

$\bar{x}_{1.c_2} = -2.5$ $\qquad$ $\bar{x}_{2.c_2} = 2.0$

2(b) 1st iteration:

$d^2(x_i, \bar{x}_{c_j})$

	$\bar{x}_j$	x_1	x_2	x_3	x_4	x_5
C1	(-1,1)	5	5	2	10	5
C2	(-2.5, 2)	9.25	4.25	0.25	4.25	12.25

$C_1 = \{x_1, x_5\}$. $\qquad$ $C_2 = \{x_2, x_3, x_4\}$

2nd iteration:

$d^2(x_i, \bar{x}_{c_j})$

	$\bar{x}_{c_j}$	x_1	x_2	x_3	x_4	x_5
C1	(-0.5, 0.5)	4.5	6.5	4.5	14.5	4.5
C2	$(-\frac{7}{3}, 2)$	9.11	4.44	0.11	4.11	11.11

$C_1 = \{x_1, x_5\}$. $\qquad$ $C_2 = \{x_2, x_3, x_4\}$

Version B

2 (a) $C_1 = \{x_1, x_3, x_5\}$ and $C_2 = \{x_2, x_4, x_6\}$

$\bar{x}_{1.C_1} = \frac{5-2+1}{3} = \frac{4}{3}$ $\bar{x}_{2.C_1} = \frac{-3+2+2}{3} = \frac{1}{3}$

$\bar{x}_{1.C_2} = \frac{-2-2}{2} = -2$ $\bar{x}_{2.C_2} = \frac{-4+4}{2} = 0$

2(b) 1st iteration:

$d^2(x_i, \bar{x}_{c_j})$

$\bar{x}_j$	x_1	x_2	x_3	x_4	x_5
C1 $(\frac{4}{3}, \frac{1}{3})$	24.56	29.89	13.89	24.56	2.89
C2 $(-2, 0)$	58	16	4	16	13

$C_1 = \{x_1, x_5\}$. $C_2 = \{x_2, x_3, x_4\}$

2nd iteration:

$d^2(x_i, \bar{x}_{c_j})$

$\bar{x}_{c_j}$	x_1	x_2	x_3	x_4	x_5
C1 $(3, -1/2)$	10.25	37.25	31.25	45.25	10.25
C2 $(-\frac{7}{3}, 2)$	78.78	36.11	0.11	4.11	11.11

$C_1 = \{x_1, x_5\}$. $C_2 = \{x_2, x_3, x_4\}$

2(c) $C_1 = \{\underline{x}_1, \underline{x}_3, \underline{x}_5\}$ $C_2 = \{\underline{x}_2, \underline{x}_4\}$

Version A:

$$d(\underline{x}_1, \underline{x}_3) = \sqrt{(-2+2)^2 + (-1-2)^2} = \sqrt{9} = 3$$

$$d(\underline{x}_1, \underline{x}_5) = \sqrt{(-2-1)^2 + (-1-2)^2} = \sqrt{18} = 3\sqrt{2}$$

$$d(\underline{x}_1, \underline{x}_2) = \sqrt{(-2+3)^2 + (-1-0)^2} = \sqrt{2}$$

$$d(x_1, x_4) = \sqrt{(-2+2)^2 + (-1-4)^2} = \sqrt{25} = 5$$

$$b(1) = \min_{i \neq 1} \frac{1}{n_j} \sum_j d(\underline{x}_1, \underline{x}_j) = (5 + \sqrt{2})/2 \simeq 3.207107$$

$$a(1) = \frac{1}{n_1 - 1} \sum_i d(\underline{x}_1, \underline{x}_j) = \frac{3 + 3\sqrt{2}}{2} \simeq 3.62132$$

$$s(1) = \frac{b(1) - a(1)}{\max(b(1), a(1))} = \frac{3.207107 - 3.62132}{3.62132}$$

$$= -0.1144 < 0.25$$

thus, object x_1 is badly classified in this cluster. Note that this is expected since we are dealing with an initial partition and the final one is quite different.

Version B:

$\underline{x}_1 = (5, -3)$

$d(\underline{x}_1, \underline{x}_3) = \sqrt{74}$

$d(\underline{x}_1, \underline{x}_5) = \sqrt{41}$

$d(\underline{x}_1, \underline{x}_2) = \sqrt{50}$

$d(x_1, x_4) = \sqrt{98}$

$$b(1) = \min_{i \neq 1} \frac{1}{n_j} \sum_j d(\underline{x}_1, \underline{x}_j) = \frac{1}{2}\left(\sqrt{50} + \sqrt{98}\right) = 8.4853$$

$$a(1) = \frac{1}{n_1 - 1} \sum_{j \neq 1} d(\underline{x}_1, x_j) = \frac{1}{2}\left(\sqrt{74} + \sqrt{41}\right) = 7.5027$$

$$s(1) = \frac{b(1) - a(1)}{\max(b(1), a(1))} = \frac{8.4853 - 7.5027}{8.4853}$$

$$= 0.1158 < 0.25$$

thus, object x_1 is badly classified in this cluster. Note that this is expected since we are dealing with an initial partition and the final one is quite different.

Group II

1. Deduce (. . .)

$$\frac{f_{X|Y=1}(x)}{f_{X|Y=0}(x)} \geq \frac{1-p}{p}$$ then assign x to $\{Y=1\}$

$$\Leftrightarrow \quad \frac{d_1 \gamma_1^{d_1} x^{d_1-1} e^{-(\gamma_1 x)^{d_1}}}{d_0 \gamma_0^{d_0} x^{d_0-1} e^{-(d_0 x)^{d_1}}} \geq \frac{1-p}{p}$$

$$\Leftrightarrow \quad d_1 = d_0 = d$$

$$\frac{\cancel{d} \gamma_1^{d} x^{d-1} e^{(-\gamma_1 x)^d}}{\cancel{d} \gamma_0^{d} x^{d-1} e^{(-\gamma_0 x)^d}} \geq \frac{1-p}{p}$$

$$\Leftrightarrow \quad \left(\frac{\gamma_1}{\gamma_0}\right)^d \exp\left\{ -(\gamma_1 x)^d + (\gamma_0 x)^d \right\} \geq \frac{1-p}{p}$$

$$\Leftrightarrow \quad \exp\left\{ -(\gamma_1 x)^d + (\gamma_0 x)^d \right\} \geq \frac{1-p}{p} \left(\frac{\gamma_0}{\gamma_1}\right)^d$$

$$\Leftrightarrow \quad -(\gamma_1 x)^d + (\gamma_0 x)^d \geq \log\left(\frac{1-p}{p}\right)$$

$$\Leftrightarrow \quad x^d \underbrace{(\gamma_0^d - \gamma_1^d)}_{<0} \geq \log\left(\frac{1-p}{p}\right)$$

$$\Leftrightarrow \quad x^d \leq \frac{\log(\text{—})}{\gamma_0^d - \gamma_1^d}$$

(E) $$x \leq \left[\frac{\log\left(\frac{1-p}{p}\right) + d \log\left(\frac{\sigma_0}{\sigma_1}\right)}{\sigma_0^{d} - \sigma_1^{d}} \right]^{1/d} = \xi$$

Classification Rule:

If $$x \leq \left[\frac{\log\left(\frac{1-p}{p}\right) + d \log\left(\frac{\sigma_0}{\sigma_1}\right)}{\sigma_0^{d} - \sigma_1^{d}} \right]^{1/d} \Rightarrow \text{assign } x \text{ to } \{Y=1\}$$

otherwise assign x to $\{Y=0\}$.

(b.A) $d=1$, $\sigma_0=1$, $\sigma_1=10$, $p=0.4$

$$\xi = \frac{\log\left(\frac{0.6}{0.4}\right) + \log\left(\frac{1}{10}\right)}{1-10} = \frac{\log(1.5) - \log(10)}{-9} = \frac{\log(10) - \log(1.5)}{9}$$

$$= 0.2108$$

(i) $$Se = P(\text{classify } \{Y=1\} \mid Y=1) = P(X \leq \xi \mid Y=1)$$

$$= F_{X|Y=1}(\xi) = 1 - \exp\{-(10\xi)\} = 0.8785$$

(iii) $$PPV = P(Y=1 \mid \text{classify } x \; Y=1) = \frac{P(\text{classif } Y=1 \mid Y=1)\, P(Y=1)}{P(\text{classif } Y=1 \mid Y=1) P(Y=1) + P(\text{classif } Y=1 \mid Y=0)(1-p)}$$

$$= \frac{Se\, p}{Se\, p + (1-Sp)(1-p)} \simeq 0.7550$$

$$1 - Sp = P(X \leq \xi \mid Y=0) = 1 - \exp(-\xi) = 0.1901$$

(b.B)

(i) $$Sp = P(X > \xi \mid Y=0) = 1 - P(X \leq \xi \mid Y=0) = 1 - [1 - \exp(-\xi)]$$

$$= \exp(-\xi) = 0.8100$$

(ii) $NPV = P(Y=0 \mid \text{classif } Y=0) = \dfrac{P(\text{classif } Y=0 \mid Y=0)\, P(Y=0)}{P(\text{classif } Y=0 \mid Y=1)\, P(Y=1) + P(\text{classif } Y=0 \mid Y=0)(1-p)}$

$$= \frac{Sp(1-p)}{(1-Se)p + Sp(1-p)} = 0.9091$$

2. (A) Not necessarly.

(B) Wrong.