

SAI SỐ BAYES VÀ KHOẢNG CÁCH GIỮA HAI HÀM MẬT ĐỘ XÁC SUẤT TRONG PHÂN LOẠI HAI TỔNG THỂ

Võ Văn Tài⁽¹⁾, Phạm Gia Thụ⁽²⁾, Tô Anh Dũng⁽³⁾

(1) Trường Đại học Cần Thơ

(2) Trường Đại học Moncton, Canada

(3) Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

1. GIỚI THIỆU

Trong thực tế có nhiều vấn đề đòi hỏi chúng ta phải giải quyết bài toán phân loại hai tổng thể H_1 và H_2 . Có nhiều cách khác nhau để giải quyết bài toán phân loại này như kiểu phân loại dựa vào khoảng cách Metric đã được đề cập bởi Forgy (1965), Mac Queen (1967), E.Dilay (1972). Đó cũng là phân tích phân biệt của R.A. Fisher (1936), P.C. Mahalanobis (1936) (xem [5]). Các phương pháp này có nhược điểm là không xác định được xác suất của sai lầm trong phân loại.

Một phương pháp phân loại khác dựa trên hàm mật độ xác suất của hai tổng thể, đó là phương pháp Bayes. Phương pháp này có thể tính được xác suất sai lầm tối thiểu trong phân loại. Giả sử trên hai tổng thể ta quan sát biến ngẫu nhiên X , gọi $f_1(x)$, $f_2(x)$ là hàm mật độ xác suất của hai tổng thể. Nếu ta không quan tâm đến xác suất tiên nghiệm v của H_1 thì sai số Bayes được xác định

$$P_e = \int_R \min\{f_1(x), (1 - f_2(x))\} dx, \quad \text{và nếu quan tâm đến } v \text{ thì}$$

$$P_e = \int_R \min\{v \cdot f_1(x), (1 - v) f_2(x)\} dx. P_e \text{ đã được chứng minh là xác suất sai lầm nhỏ nhất trong}$$

phân loại. Như vậy phương pháp Bayes đã giải quyết được vấn đề quan trọng trong lý thuyết phân loại, đó là việc tính sai số trong phân loại. Tuy nhiên, trong thực tế việc tính kết quả cụ thể gặp nhiều khó khăn, bởi việc xác định hàm mật độ xác suất, việc giải phương trình và việc tính các tích phân. Trong bài viết này chúng tôi quan tâm đến việc xác định sai số Bayes, tìm hàm mật

độ xác suất cho tổng của hai loại sai lầm trên khoảng $(0, \frac{1}{4})$, từ đó xác định khoảng cách L^1 của

hai hàm mật độ theo Lissack và Fu (1976). Các vấn đề được xem xét chi tiết cho phân phối chuẩn, phân phối mũ và phân phối Beta.

2. SAI SỐ BAYES TRONG PHÂN LOẠI HAI TỔNG THỂ

2.1. Hai tổng thể với hàm mật độ xác suất $f_1(x)$ và $f_2(x)$ có một đỉnh

2.1.1. Khi không quan tâm đến xác suất tiên nghiệm v của H_1

Phương trình $f_1(x) - f_2(x) = 0$ có thể có một nghiệm hoặc nhiều nghiệm. Giả sử $f_1(x)$ và $f_2(x)$ là hàm số chỉ có một đỉnh thì phương trình trên nếu có nghiệm chỉ có thể có một nghiệm hoặc hai nghiệm.

Nếu phương trình trên có một nghiệm x_0 thì ta có phân tích nhận dạng như sau: một phần tử với quan sát y được xếp vào H_1 nếu $y \leq x_0$ và xếp vào H_2 nếu $y > x_0$.

Đặt $h(x) = \min\{f_1(x), f_2(x)\}$, khi đó:

$t = P(H_2|H_1) = \int_{x > x_0} h(x) dx$: xác suất phân loại một phần tử vào H_2 khi thật sự nó thuộc H_1 .

$d = P(H_1|H_2) = \int_{x \leq x_0} h(x) dx$: xác suất phân loại một phần tử vào H_1 khi thật sự nó thuộc H_2 .

Nếu phương trình có hai nghiệm x_1 và x_2 (giả sử $x_1 < x_2$) thì một phần tử với quan sát y sẽ được xếp vào H_1 nếu $x_1 \leq y \leq x_2$ và xếp vào H_2 nếu $y \notin [x_1, x_2]$. Khi đó:

$$t = \int_{\{x > x_2\} \cup \{x < x_1\}} h(x) dx \quad d = \int_{x_1 \leq x \leq x_2} h(x) dx$$

Trong cả hai trường hợp ta có xác suất của phân loại sai lầm là $e = P_e = t + d$. Chúng ta chứng minh được bất kỳ sự chọn lựa nào khác x_0 hoặc x_1 và x_2 trong phân tích nhận dạng đều dẫn đến một xác suất sai lầm lớn hơn P_e , nghĩa là phân loại Bayes có xác suất sai lầm tối thiểu.

2.1.2. Khi quan tâm đến xác suất tiên nghiệm v (hằng số) của H_1

Đặt $k_1(x) = vf_1(x)$, $k_2 = (1-v)f_2(x)$, khi đó phương trình

$$k_1(x) = k_2(x) \text{ hay } \ln \frac{f_1(x)}{f_2(x)} = \frac{1-v}{v}$$

có thể có một nghiệm x'_0 hoặc hai nghiệm x'_1 và x'_2 . Phân tích nhận dạng được xác định như trường hợp a). Khi đó xác suất sai lầm trong phân loại trở thành t_1 và d_1 với

$$t_1 = \int_{R_1} k_1(x) dx \quad d_1 = \int_{R_2} k_2(x) dx$$

trong đó $R_1 = \{x \mid k_1(x) \geq k_2(x)\}$ và $R_2 = \{x \mid k_1(x) < k_2(x)\}$. Miền R_1 và R_2 được xác định từ x'_0 hoặc x'_1 và x'_2 ở trên. Khi đó xác suất sai lầm trong phân loại $P_e = t_1 + d_1$ cũng là nhỏ nhất.

Khi xác suất tiên nghiệm trong phân loại hai tổng thể là như nhau $v = \frac{1}{2}$ thì

$$t_1 = \frac{1}{2} \int_{\{x > x_2\} \cup \{x < x_1\}} h(x) dx \quad d_1 = \frac{1}{2} \int_{x_1 \leq x \leq x_2} h(x) dx$$

P_e xác định ở trên cũng là xác suất sai lầm tối thiểu.

2.1.3. Khi v là biến ngẫu nhiên với hàm mật độ xác suất tiên nghiệm biết trước

Phân tích nhận dạng và sai số Bayes trong trường hợp này được xác định như trường hợp b) bằng việc thay v bởi kỳ vọng của phân phối tiên nghiệm của v .

2.1.4. Trường hợp không có sai lầm ($e = t = d = 0$)

Trường hợp này xảy ra khi $f_1(x)$ và $f_2(x)$ không cắt nhau, khi đó ta có thể ước lượng tỷ lệ của H_1 trong tổng thể $H_1 \cup H_2$ bằng cách giả sử tỷ lệ này ban đầu có phân phối tiên nghiệm Beta và lấy một mẫu từ tổng thể chung qua định lý dưới đây.

Định lý 1: Lấy n phần tử quan sát từ tổng thể trộn $H_1 \cup H_2$. Gọi X_i là đại lượng ngẫu nhiên ứng với quan sát thứ i mà $X_i = 1$ nếu phần tử quan sát thuộc H_1 và $X_i = 0$ nếu phần tử quan sát không thuộc H_1 . Giả sử $P(X_i = 1) = h$ và h có phân phối tiên nghiệm Beta(a, b), khi đó ta có các kết quả sau:

1) Hàm mật độ xác suất hậu nghiệm của h là

$$j^{(n)}(h) = \text{Beta}(h, a + y, b + n - y) \text{ trong đó } y = \sum_{i=1}^n X_i \quad (1)$$

2) Kỳ vọng hậu nghiệm của h là

$$m^{(n)}(h) = \frac{a+y}{a+b+n} \quad (2)$$

3) Phương sai hậu nghiệm của h là

$$Var^{(n)}(h) = \frac{(a+y)(b+n-y)}{(a+b+n)^2(a+b+n+1)} \quad (3)$$

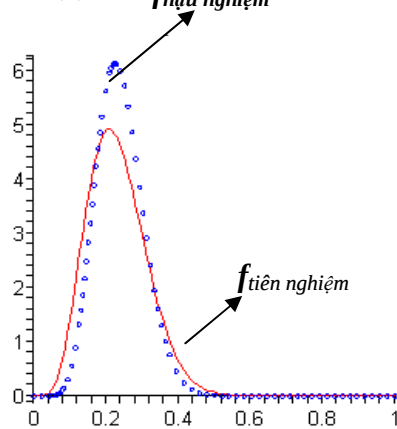
Đây là kết quả vận dụng định lý 1 ([2], trang 321) trong phân loại.

Ví dụ 1: Giả sử h không tính được chính xác, nhưng nó có phân phối tiên nghiệm Beta(6,20). Thực hiện một mẫu gồm 16 quan sát từ hai tổng thể H1 và H2 ta thấy có 4 phần tử thuộc H1 và 12 phần tử thuộc H2, khi đó:

Hàm mật độ xác suất của h theo (1) là Beta(10,32).

Kỳ vọng hậu nghiệm của h theo (2) là 0.238.

Phương sai hậu nghiệm của h theo (3) là $C_{\text{hậu nghiệm}}$



Hình 1. Đồ thị hàm mật độ xác suất tiên nghiệm và hậu nghiệm của h (Beta(6,20))

2.2. Hai tổng thể có phân phối chuẩn và Beta

Trong phần này ta không quan tâm đến v hoặc giả sử $v = \frac{1}{2}$. Xem xét xác suất sai lầm trong phân loại hai tổng thể cho hai trường hợp: Hai tổng thể có biến ngẫu nhiên phân phối chuẩn và phân phối Beta.

2.2.1. Hai tổng thể có phân phối chuẩn

Giả sử $X_1 \sim N(m_1, s_1^2)$, $X_2 \sim N(m_2, s_2^2)$, ta có hai trường hợp:

Trường hợp 1: Hai trung bình khác nhau $m_1 < m_2$.

Nếu $s_1 = s_2 = s$ thì phương trình $f_1(x) - f_2(x) = 0$ có một nghiệm $x_0 = \frac{m_1 + m_2}{2}$

Khi đó ta có $t = d = 1 - \Phi(x)$ với $x = \frac{m_2 - m_1}{2s}$ và $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$

Nếu $s_1 \neq s_2$ thì phương trình $f_1(x) - f_2(x) = 0$ có hai nghiệm sau:

$$x_i = \frac{(m_1 s_2^2 - m_2 s_1^2) \pm s_1 s_2 \sqrt{(m_1 - m_2)^2 + K}}{s_2^2 - s_1^2}, \quad i=1, 2$$

trong đó, $K = \frac{2(s_2^2 - s_1^2) \ln \frac{s_2}{s_1}}{s_1}$, và nếu $x_1 \leq x_2$ thì

$$t = 1 - \Phi\left(\frac{x_2 - m_1}{s_1}\right) + \Phi\left(\frac{x_1 - m_1}{s_1}\right); \quad d = \Phi\left(\frac{x_2 - m_2}{s_2}\right) - \Phi\left(\frac{x_1 - m_2}{s_2}\right)$$

Trường hợp 2: $m_1 = m_2$.

Nếu $s_1 \neq s_2$. Trường hợp này phương trình $f_1(x) - f_2(x) = 0$ có hai nghiệm

$$x_i = m \pm s_1 s_2 \sqrt{E} \quad \text{với} \quad E = \frac{2 \ln \frac{s_1}{s_2}}{s_2^2 - s_1^2} \geq 0, \quad i=1, 2$$

$$\text{Khi đó} \quad t = 1 - \Phi(s_2 \sqrt{E}) + \Phi(-s_2 \sqrt{E}), d = \Phi(s_1 \sqrt{E}) - \Phi(-s_1 \sqrt{E}).$$

Nếu $s_1 = s_2$. Trong trường hợp này ta có $e = t = d = 1$.

Ví dụ 2: Trên hai tổng thể H_1 và H_2 ta quan sát biến ngẫu nhiên X_1 và X_2 lần lượt có phân phối chuẩn $X_1 \sim N(5, 92)$, $X_2 \sim N(18, 62)$. Nếu ta không quan tâm đến xác suất tiên nghiệm thì phương trình $f_1(x) = f_2(x)$ có hai nghiệm $x_1 = 11.198$, $x_2 = 45.602$. Vì vậy trong phân tích nhận dạng Bayes nếu kết quả quan sát là $11.198 \leq x \leq 45.602$ thì quan sát đó được xếp vào H_1 , ngược lại ta sẽ xếp nó vào H_2 . Trong phân tích nhận dạng này

$$t = P(H_2 | H_1) = \int_{11.198}^{45.602} f_1(x) dx = 0.2455,$$

$$d = P(H_1 | H_2) = \int_{-\infty}^{11.198} f_2(x) dx + \int_{45.602}^{+\infty} f_2(x) dx = 0.1285$$

và xác suất sai lầm trong phân loại là $e = t + d = 0.3739$.

Nếu $v = \frac{1}{2}$ thì xác suất sai lầm $e = \frac{0.3739}{2} = 0.18695$.

2.2.2. Hai tổng thể có phân phối Beta

Giả sử $X_1 \sim \text{Beta}(a_1, b_1)$; $X_2 \sim \text{Beta}(a_2, b_2)$

$$\text{Xét phương trình } f_1(x) = f_2(x) \Leftrightarrow \frac{x^{a_1-1}(1-x)^{b_1-1}}{B(a_1, b_1)} = \frac{x^{a_2-1}(1-x)^{b_2-1}}{B(a_2, b_2)}$$

$$\Leftrightarrow x^{a_1-a_2}(1-x)^{b_1-b_2} = \frac{B(a_1, b_1)}{B(a_2, b_2)}$$

$$\Leftrightarrow x^a(1-x)^b = A$$

$$a = a_1 - a_2; b = b_1 - b_2; A = \frac{B(a_1, b_1)}{B(a_2, b_2)}.$$

Trong đó,

$$\text{Đặt } k = \frac{a}{b}; B = \sqrt[b]{A} > 0 \text{ khi đó phương trình trên trở thành}$$

$$xk - xk+1 = B \quad (4)$$

Phương trình (4) có thể giải được trên máy tính, ta tìm được hoành độ giao điểm của hai hàm mật độ $f_1(x)$ và $f_2(x)$ và từ đó ta có thể tính được $t = P(H_2 | H_1)$ và $d = P(H_1 | H_2)$. Việc tính t và d dẫn đến việc tính tích phân của hàm Beta

$$F(x) = \frac{1}{B(a, b)} \int_0^x x^{a-1}(1-x)^{b-1} dx \quad (5)$$

Tích phân (5) theo Robert J. Boik (1988) tính được

$$F(x) = K_{x,a,b} {}_2F_1\left(\begin{matrix} 1-b, 1; a+1; \\ 1-x \end{matrix}\right) \quad (6)$$

trong đó, $K_{x,a,b} = \frac{x^a(1-x)^{b-1}}{aB(a, b)}$, ${}_2F_1(a, b; c; x) = \sum_{n=0}^{\infty} \frac{(a, n)(b, n)}{(c, n)} \frac{x^n}{n!}$, với (a, n) là hệ số Pochhammer (xem [3]).

Và theo Tretter và Walster (xem [4]), dùng tính toán gần đúng cấp n nhận được

$$1 + \frac{a_1}{b_1 + \frac{a_2}{b_2 + \frac{a_3}{b_3 + \frac{a_4}{\dots + \frac{a_n}{b_n}}}}}$$

${}_2F_1 \approx$

trong đó,

$$a_1 = \frac{af(b-1)}{b(a+1)} \quad a_n = \frac{a^2 f^2(n-1)(a+b+n-2)(a+n-1)(b-n)}{b^2(a+2n-3)(a+2n-2)^2(a+2n-1)}; \quad n \geq 2$$

$$b_n = \frac{2(a.f + 2b)n^2 + 2(af + 2b)(a-1)n + ab(a-2-af)}{b(a+2n-2)(a+2n)}; \quad n \geq 1$$

$$f = \frac{bx}{a(1-x)}$$

Nhận xét. Trong trường hợp đặc biệt $a_1 = b_1 = p$ và $a_2 = b_2 = q$, hai đồ thị của các hàm số đều đối xứng với nhau qua đường $x = \frac{1}{2}$.

Nếu $p = q$ thì (4) sẽ có vô số nghiệm.

Nếu $p \neq q$ thì (4) trở thành $x^2 - x + B = 0$ và hai đồ thị của các hàm số sẽ cắt nhau tại hai

điểm đối xứng qua $x = \frac{1}{2}$: $x_1 = \frac{1 - \sqrt{1 - 4B}}{2}$; $x_2 = \frac{1 + \sqrt{1 - 4B}}{2}$.

3. KHOẢNG CÁCH L^1 GIỮA $vf_1(x)$ VÀ $(1-v)f_2(x)$

Trong phần này ta coi v là biến ngẫu nhiên và như vậy t và d cùng với Pe cũng là biến ngẫu nhiên. Theo Lissack và Fu thì $2Pe = 1 - J_1(H_1, H_2|v)$ với $Z = J_1(H_1, H_2|v)$ là khoảng cách L^1 giữa $vf_1(x)$ và $(1-v)f_2(x)$. Từ mối quan hệ này, khi không biết về $f_1(x)$ và $f_2(x)$ cũng như v nhưng chúng ta có thông tin về hai xác suất sai lầm t và d là hai biến ngẫu nhiên độc lập, chúng ta có thể tìm được hàm mật độ xác suất của Z .

3.1 Hàm tổng của hai biến ngẫu nhiên độc lập trên $(0, \frac{1}{4})$

Định lý 2: Giả sử X_1 và X_2 là hai biến ngẫu nhiên độc lập trên $(0, \frac{1}{4})$ có hàm mật độ xác suất lần lượt là $f_1(x)$, $f_2(x)$. Xét $Y = X_1 + X_2$, khi đó hàm mật độ xác suất của Y có dạng:

$$g(y) = \begin{cases} \int_0^y f_1(t) f_2(y-t) dt & \text{khi } 0 < y \leq \frac{1}{4} \\ \int_{y-\frac{1}{4}}^{\frac{1}{4}} f_1(t) f_2(y-t) dt & \text{khi } \frac{1}{4} < y \leq \frac{1}{2} \\ 0 & \text{khi } y \notin (0, \frac{1}{2}) \end{cases}$$

Chúng minh

$$g(y) = \int_{-\infty}^{+\infty} f_1(y-x) f_2(x) dx$$

Ta có

Vì X_2 là biến ngẫu nhiên trên $(0, \frac{1}{4})$ nghĩa là $f_2(x) = 0 \quad \forall x \notin (0, \frac{1}{4})$, nên

$$g(y) = \int_0^{\frac{1}{4}} f_1(y-x) f_2(x) dx$$

Đặt $t = y - x$, $dt = -dx$; khi $x = 0$, $t = y$; khi $x = \frac{1}{4}$, $t = y - \frac{1}{4}$. Từ đó,

$$g(y) = \int_y^{\frac{1}{4}} f_1(t) f_2(y-t) (-dt) = \int_{y-\frac{1}{4}}^y f_1(t) f_2(y-t) dt$$

Vì X_1 và $X_2 \in (0, \frac{1}{4})$ nên $y \in (0, \frac{1}{2})$.

$$\text{Nếu } 0 < y \leq \frac{1}{4} \text{ thì } y - \frac{1}{4} \leq 0 \text{ nên } g(y) = \int_{y-\frac{1}{4}}^y f_1(t)f_2(y-t)dt = \int_0^y f_1(t)f_2(y-t)dt \quad (7)$$

$$\text{Nếu } \frac{1}{4} < y < \frac{1}{2} \text{ thì } y - \frac{1}{4} > 0 \text{ nên } g(y) = \int_{y-\frac{1}{4}}^y f_1(t)f_2(y-t)dt = \int_{y-\frac{1}{4}}^{\frac{1}{4}} f_1(t)f_2(y-t)dt \quad (8)$$

3.2 Một số trường hợp cụ thể của $Y = X_1 + X_2$

3.2.1. X_1, X_2 là hai biến ngẫu nhiên độc lập có phân phối Beta trên $(0, \frac{1}{4})$

Giả sử $X_1 \sim \text{Beta}(a_1, b_1; 0, \frac{1}{4})$; $X_2 \sim \text{Beta}(a_2, b_2; 0, \frac{1}{4})$ với $a_1, b_1, a_2, b_2 > 0$.

Theo Pham-Gia và Turkkan (xem [3]) ta có kết quả:

$$\text{Nếu } 0 < y \leq \frac{1}{4} \text{ thì } g(y) = H_1 4^{a_1+a_2} y^{a_1+a_2-1} (1-4y)^{b_1-1} \cdot F_D^{(2)}(a_2, 1-b_1, 1-b_2; a_1+a_2; \frac{4y}{4y-1}, 4y) \quad (9)$$

với $H_1 = \frac{\Gamma(a_1+b_1)\Gamma(a_2+b_2)}{\Gamma(a_1+a_2)\Gamma(b_1)\Gamma(b_2)}$; $F_D^{(2)}$ là hàm siêu bội với hai biến số.

$$\begin{aligned} \text{Nếu } \frac{1}{4} < y < \frac{1}{2} \text{ thì} \\ g(y) &= H_2 2^{b_1+b_2+1} (1-2y)^{b_1+b_2-1} (4y-1)^{a_2-1} \cdot F_D^{(2)}(b_2, 1-a_1, 1-a_2; b_1+b_2; 2-4y, \frac{4y-2}{4y-1}) \\ (10) \quad & \text{với} \\ H_2 &= \frac{\Gamma(a_1+b_1)\Gamma(a_2+b_2)}{\Gamma(b_1+b_2)\Gamma(a_1)\Gamma(a_2)}. \end{aligned}$$

3.2.2. X_1, X_2 là hai biến ngẫu nhiên độc lập có phân phối mũ cắt trên $(0, \frac{1}{4})$

Giả sử $X_1 \sim \text{Exp}(b_1; 0, \frac{1}{4})$, $X_2 \sim \text{Exp}(b_2; 0, \frac{1}{4})$ với $b_1, b_2 \in \mathbb{R}^+$.

Trong phần này ta có thể đổi vai trò của X_1 và X_2 cho nhau để luôn giả sử $b_1 \geq b_2$.

Khi $b_1 > b_2$,

$$\text{Nếu } 0 < y \leq \frac{1}{4} \text{ thì } g(y) = \frac{b_1 b_2}{ab(b_1 - b_2)} [e^{-b_2 y} - e^{-b_1 y}]. \quad (11)$$

$$\text{Nếu } \frac{1}{4} < y < \frac{1}{2} \text{ thì } g(y) = \frac{b_1 b_2}{ab(b_1 - b_2)} \left[e^{-\frac{b_2 - b_1 + 4b_1 y}{4}} - e^{-\frac{b_1 - b_2 + 4b_2 y}{4}} \right]. \quad (12)$$

$$\text{với } b_1, b_2 > 0 \text{ và } a = \int_0^{\frac{1}{4}} f_1(x)dx = 1 - e^{-\frac{b_1}{4}}; b = \int_0^{\frac{1}{4}} f_2(x)dx = 1 - e^{-\frac{b_2}{4}}.$$

Khi $b_1 = b_2 = c$,

$$\text{Nếu } 0 < y \leq \frac{1}{4} \text{ thì } g(y) = \left(\frac{c}{d}\right)^2 y \cdot e^{-cy}. \quad (13)$$

$$\text{Nếu } \frac{1}{4} < y \leq \frac{1}{2} \text{ thì } g(y) = \left(\frac{c}{d}\right)^2 \left(\frac{1}{2} - y\right) e^{-cy}, \quad (14)$$

$$\text{với } d = \int_0^{\frac{1}{4}} f_1(x)dx = \int_0^{\frac{1}{4}} ce^{-cx}dx = 1 - e^{-\frac{c}{4}}.$$

Chứng minh.

Khi $b_1 > b_2$ vì $X_1 \sim \text{Exp}(b_1; 0, \frac{1}{4})$; $X_2 \sim \text{Exp}(b_2; 0, \frac{1}{4})$ nên trên $(0, \frac{1}{4})$

$$f_1(x) = \frac{b_1}{a} e^{-b_1 x}; f_2(x) = \frac{b_2}{b} e^{-b_2 x}$$

$$\text{Nếu } 0 < y \leq \frac{1}{4}, \text{ theo (7) thì } g(y) = \int_0^y f_1(t)f_2(y-t)dt = \frac{b_1 b_2}{ab} e^{-b_2 y} \int_0^y e^{-(b_1 - b_2)t} dt$$

$$\text{Vì } b_1 > b_2 \text{ nên } g(y) = \frac{b_1 b_2}{ab(b_1 - b_2)} e^{-b_2 y} [1 - e^{-(b_1 - b_2)y}] = \frac{b_1 b_2}{ab(b_1 - b_2)} [e^{-b_2 y} - e^{-b_1 y}]$$

Nếu $\frac{1}{4} < y < \frac{1}{2}$, tương tự như trên ta có:

$$\begin{aligned} g(y) &= \frac{b_1 b_2}{ab} e^{-b_2 y} \int_{y-\frac{1}{4}}^{\frac{1}{4}} e^{-(b_1 - b_2)t} dt = \frac{b_1 b_2}{ab(b_1 - b_2)} e^{-b_2 y} \left[e^{-(b_1 - b_2)(y-\frac{1}{4})} - e^{-(b_1 - b_2)/4} \right] \\ &= \frac{b_1 b_2}{ab(b_1 - b_2)} \left[e^{-\frac{b_2 - b_1 + 4b_1 y}{4}} - e^{-\frac{b_1 - b_2 + 4b_2 y}{4}} \right] \end{aligned}$$

Khi $b_1 = b_2 = c$, ta có $a = b = d$, vì vậy:

$$\text{Nếu } 0 < y \leq \frac{1}{4} \text{ thì } g(y) = \frac{b_1 b_2}{ab} e^{-b_2 y} \int_0^y 1 dt = \left(\frac{c}{d}\right)^2 y \cdot e^{-cy}.$$

$$\text{Nếu } \frac{1}{4} < y < \frac{1}{2} \text{ thì } g(y) = \frac{b_1 b_2}{ab} e^{-b_2 y} \int_{y-\frac{1}{4}}^{\frac{1}{4}} 1 dt = \left(\frac{c}{d}\right)^2 \left(\frac{1}{2} - y\right) e^{-cy}.$$

2.2.3. Nếu X_1, X_2 là hai biến ngẫu nhiên độc lập có phân phối chuẩn cắt trên $(0, \frac{1}{4})$

Giả sử $X_1 \sim N(m_1, s_1^2; 0, \frac{1}{4})$, $X_2 \sim N(m_2, s_2^2; 0, \frac{1}{4})$ với $m_1, m_2 \in R$ và $s_1, s_2 \in R^+$.

Nếu $0 < y \leq \frac{1}{4}$ thì

$$g(y) = K_1 e^{-By^2 + Cy} \left[\Phi\left(\frac{s_2}{s_1 \sqrt{s_1^2 + s_2^2}} y + K_2\right) - \Phi\left(-\frac{s_1}{s_2 \sqrt{s_1^2 + s_2^2}} y + K_2\right) \right] \quad (15)$$

Trong đó a, b, A, B, C, K1, K2 là các hằng số có dạng

$$A = \frac{1}{2s_1^2} + \frac{1}{2s_2^2}; \quad B = \frac{1}{2(s_1^2 + s_2^2)}; \quad C = \frac{m_1 + m_2}{s_1^2 + s_2^2}$$

$$K_1 = \frac{1}{ab\sqrt{2p(s_1^2 + s_2^2)}} e^{-\frac{(m_1+m_2)^2}{2(s_1^2+s_2^2)}}; \quad K_2 = \frac{m_2 s_1^2 - m_1 s_2^2}{s_1 s_2 \sqrt{s_1^2 + s_2^2}}$$

$$a = \int_0^{1/4} \frac{1}{s_1 \sqrt{2p}} e^{-\frac{(x-m_1)^2}{2s_1^2}} dx = \Phi\left(\frac{1-4m_1}{4s_1}\right) - \Phi\left(-\frac{m_1}{s_1}\right)$$

$$b = \int_0^{1/4} \frac{1}{s_2 \sqrt{2p}} e^{-\frac{(x-m_2)^2}{2s_2^2}} dx = \Phi\left(\frac{1-4m_2}{4s_2}\right) - \Phi\left(-\frac{m_2}{s_2}\right)$$

Nếu $\frac{1}{4} < y < \frac{1}{2}$ thì

$$g(y) = K_1 e^{-By^2 + Cy} \left[\Phi\left(-\frac{s_1}{s_2 \sqrt{s_1^2 + s_2^2}} y + K_2 + \frac{\sqrt{s_1^2 + s_2^2}}{4s_1 s_2}\right) - \Phi\left(\frac{s_2}{s_1 \sqrt{s_1^2 + s_2^2}} y + K_2 - \frac{\sqrt{s_1^2 + s_2^2}}{4s_1 s_2}\right) \right] \quad (16)$$

Chứng minh

Theo (7) nếu $0 < y \leq \frac{1}{4}$ thì

$$g(y) = \int_0^y f_1(t) f_2(y-t) dt$$

$$= \frac{1}{2pabs_1 s_2} \int_0^y e^{-\frac{(t-m_1)^2}{2s_1^2}} \cdot e^{-\frac{(y-t-m_2)^2}{2s_2^2}} dt = \frac{1}{2pabs_1 s_2} e^Q \int_0^y e^{-(At^2 + Pt)} dt$$

Trong đó a, b, A được xác định như trên, và

$$P = \frac{m_2}{s_2^2} - \frac{m_1}{s_1^2} - \frac{y}{s_2^2}; \quad Q = \frac{m_2 y}{s_2^2} - \frac{m_1^2}{2s_1^2} - \frac{m_2^2}{2s_2^2} - \frac{y^2}{2s_2^2}$$

$$V\ddot{t} - (At^2 + Pt) = \frac{-\left(\sqrt{2At} + \frac{P}{\sqrt{2A}}\right)^2}{2} + \frac{P^2}{4A} \quad \text{nên} \quad \int_0^y e^{-(At^2+Pt)} dt = e^{\frac{P^2}{4A}} \cdot \int_0^y e^{-\frac{(\sqrt{2At} + \frac{P}{\sqrt{2A}})^2}{2}} dt$$

Đặt $u = \sqrt{2A}t + \frac{P}{\sqrt{2A}}, du = \sqrt{2A} dt$; khi $t = 0, u = \frac{P}{\sqrt{2A}}$;
 $t = y, u = \sqrt{2A}y + \frac{P}{\sqrt{2A}}$.

Từ đó,

$$\begin{aligned} \int_0^y e^{-(At^2+Pt)} dt &= \frac{e^{\frac{P^2}{4A}}}{\sqrt{2A}} \int_{\frac{P}{\sqrt{2A}}}^{\sqrt{2A}y + \frac{P}{\sqrt{2A}}} e^{-\frac{u^2}{2}} du \\ &= \frac{\sqrt{2p} e^{\frac{P^2}{4A}}}{\sqrt{2A}} \frac{1}{\sqrt{2p}} \int_{\frac{P}{\sqrt{2A}}}^{\sqrt{2A}y + \frac{P}{\sqrt{2A}}} e^{-\frac{u^2}{2}} du \\ &= \frac{\sqrt{p} e^{\frac{P^2}{4A}}}{\sqrt{A}} \left[\Phi\left(\sqrt{2A}y + \frac{P}{\sqrt{2A}}\right) - \Phi\left(\frac{P}{\sqrt{2A}}\right) \right] \end{aligned}$$

Thế tích phân này vào g(y) ta có

$$\begin{aligned} g(y) &= \frac{1}{2abs_1s_2\sqrt{pA}} e^{\frac{P^2}{4A}+Q} \left[\Phi\left(\sqrt{2A}y + \frac{P}{\sqrt{2A}}\right) - \Phi\left(\frac{P}{\sqrt{2A}}\right) \right] \\ V\ddot{t} \quad \frac{P^2}{4A} + Q &= -\frac{y^2}{2(s_1^2 + s_2^2)} + \frac{(m_1 + m_2)y}{s_1^2 + s_2^2} - \frac{(m_1 + m_2)^2}{2(s_1^2 + s_2^2)} \quad \text{nên} \\ \frac{1}{2abs_1s_2\sqrt{pA}} e^{\frac{P^2}{4A}+Q} &= K_1 e^{-By^2+Cy} \end{aligned}$$

$$UP = \sqrt{2A}y + \frac{P}{\sqrt{2A}} = \frac{y}{s_1^2 \sqrt{\frac{s_1^2 + s_2^2}{s_1^2 s_2^2}}} + \frac{m_2 s_1^2 - m_1 s_2^2}{s_1^2 s_2^2 \sqrt{\frac{s_1^2 + s_2^2}{s_1^2 s_2^2}}} = \left[\frac{s_2}{s_1 \sqrt{s_1^2 + s_2^2}} \right] y + K_2$$

$$LP = \frac{P}{\sqrt{2A}} = -\frac{y}{s_2^2 \sqrt{\frac{s_1^2 + s_2^2}{s_1^2 s_2^2}}} + \frac{m_2 s_1^2 - m_1 s_2^2}{s_1^2 s_2^2 \sqrt{\frac{s_1^2 + s_2^2}{s_1^2 s_2^2}}} = -\left[\frac{s_1}{s_2 \sqrt{s_1^2 + s_2^2}} \right] y + K_2$$

Thay các kết quả trên vào g(y) ta có (15).

Tương tự, nếu $\frac{1}{4} < y < \frac{1}{2}$ thì

$$\begin{aligned} \int_{y-\frac{1}{4}}^{\frac{1}{4}} e^{-(At^2+Pt)} dt &= \frac{e^{\frac{P^2}{4A}}}{\sqrt{2A}} \int_{\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}}^{\frac{A+2P}{2\sqrt{2A}}} e^{-\frac{u^2}{2}} du = \frac{\sqrt{2p} e^{\frac{P^2}{4A}}}{\sqrt{2A}} \frac{1}{\sqrt{2p}} \int_{\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}}^{\frac{A+2P}{2\sqrt{2A}}} e^{-\frac{u^2}{2}} du \\ &= \frac{\sqrt{p} e^{\frac{P^2}{4A}}}{\sqrt{A}} \left[\Phi\left(\frac{A+2P}{2\sqrt{2A}}\right) - \Phi\left(\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}\right) \right] \\ \text{Khi đó } g(y) &= K_1 e^{-By^2+Cy} \cdot \left[\Phi\left(\frac{A+2P}{2\sqrt{2A}}\right) - \Phi\left(\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}\right) \right] \\ \text{UP} &= \frac{A+2P}{2\sqrt{2A}} \\ &= -\frac{y}{s_2^2 \sqrt{\frac{s_1^2+s_2^2}{s_1^2 s_2^2}}} + \frac{s_1^2+s_2^2+4m_2 s_1^2-4m_1 s_2^2}{4s_1^2 s_2^2 \sqrt{\frac{s_1^2+s_2^2}{s_1^2 s_2^2}}} = -\frac{s_1}{s_2 \sqrt{s_1^2+s_2^2}} y + K_2 + \frac{\sqrt{s_1^2+s_2^2}}{4s_1 s_2} \\ \text{LP} &= \frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}} \\ &= \frac{y}{s_1^2 \sqrt{\frac{s_1^2+s_2^2}{s_1^2 s_2^2}}} + \frac{-s_1^2-s_2^2+4m_2 s_1^2-4m_1 s_2^2}{4s_1^2 s_2^2 \sqrt{\frac{s_1^2+s_2^2}{s_1^2 s_2^2}}} = \frac{s_2}{s_1 \sqrt{s_1^2+s_2^2}} y + K_2 - \frac{\sqrt{s_1^2+s_2^2}}{4s_1 s_2} \end{aligned}$$

Thế các kết quả trên vào $g(y)$ ta có (16).

Nhận xét. Tùy theo giá trị của m và s hàm mật độ xác suất của luật chuẩn cắt trên khoảng $(0, \frac{1}{4})$ có thể có rất nhiều hình dạng khác nhau nên (15) và (16) có thể cho những hình dạng rất phong phú của hàm mật độ xác suất của tổng hai biến ngẫu nhiên.

4. MỘT SỐ TRƯỜNG HỢP CỤ THỂ VỀ HÀM MẬT ĐỘ CỦA Z

Ta có $Z = 1 - 2P_e = 1 - 2y$ ($y = t + d$, có hàm mật độ xác suất $g(y)$). Vì hàm ngược của Z là $y = \frac{1-Z}{2}$ và $y'_Z = -\frac{1}{2}$ nên hàm mật độ xác suất của Z là $h(z) = |y'_Z| g\left(\frac{1-z}{2}\right)$.

Thế hàm mật độ xác suất $g(y)$ lần lượt vào các kết quả trên về tổng của hai hàm mật độ xác suất trên $(0, \frac{1}{4})$ ta có các kết quả sau:

4.1 t và d là hai biến ngẫu nhiên độc lập có phân phối Beta trên $(0, \frac{1}{4})$

Giả sử $t \sim \text{Beta}(a_1, b_1; 0, \frac{1}{4}); d \sim \text{Beta}(a_2, b_2; 0, \frac{1}{4})$

Nếu $0 < z < \frac{1}{2}$ thì

$$h(z) = H_2 2^{b_1+b_2} z^{b_1+b_2-1} (1-2z)^{a_2-1} \cdot F_D^{(2)}(b_2, 1-a_1, 1-a_2; b_1+b_2; 2z, \frac{2z}{2z-1}) \quad (17)$$

Nếu $\frac{1}{2} \leq z < 1$ thì

$$h(z) = H_1 2^{a_1+a_2} (1-z)^{a_1+a_2-1} (2z-1)^{b_1-1} \cdot F_D^{(2)}(a_2, 1-b_1, 1-b_2; a_1+a_2; \frac{2-2z}{1-2z}, 2-2z) \quad (18)$$

Đây là kết quả đã được tác giả T. Pham-Gia trình bày trong [3].

4.2. t và d là hai biến ngẫu nhiên độc lập có phân phối mũ cắt trên $(0, \frac{1}{4})$

Giả sử $t \sim \text{Exp}(b_1; 0, \frac{1}{4}); d \sim \text{Exp}(b_2; 0, \frac{1}{4})$.

Khi $b_1 > b_2$:

$$\text{Nếu } 0 < z < \frac{1}{2} \text{ thì } h(z) = \frac{b_1 b_2}{2ab(b_1 - b_2)} e^{-\frac{b_2+b_2}{2}} \left[e^{\frac{b_1 z}{2}} - e^{\frac{b_2 z}{2}} \right] \quad (19)$$

$$\text{Nếu } \frac{1}{2} \leq z < 1 \text{ thì } h(z) = \frac{b_1 b_2}{2ab(b_1 - b_2)} \left[e^{-\frac{b_2(1-z)}{2}} - e^{-\frac{b_1(1-z)}{2}} \right] \quad (20)$$

Khi $b_1 = b_2$:

$$\text{Nếu } 0 < z < \frac{1}{2} \text{ thì } h(z) = \left(\frac{c}{2d} \right)^2 z e^{-\frac{c(1-z)}{2}} \quad (21)$$

$$\text{Nếu } \frac{1}{2} \leq z < 1 \text{ thì } h(z) = \left(\frac{c}{2d} \right)^2 (1-z) e^{-\frac{c(1-z)}{2}} \quad (22)$$

4.3. t và d là hai biến ngẫu nhiên độc lập có phân phối chuẩn cắt trên $(0, \frac{1}{4})$

Giả sử $t \sim N(m_1, s_1^2; 0, \frac{1}{4})$ và $d \sim N(m_2, s_2^2; 0, \frac{1}{4})$.

Nếu $0 < z < \frac{1}{2}$ thì

$$h(z) = \frac{1}{2} K_1 e^{\frac{2C-B}{4}} e^{-\frac{B}{4}z^2 + \frac{B-C}{2}z} \left[\Phi \left(-\frac{s_1(1-z)}{2s_2\sqrt{s_1^2 + s_2^2}} + K_2 + \frac{\sqrt{s_1^2 + s_2^2}}{4s_1s_2} \right) - \Phi \left(\frac{s_2(1-z)}{2s_1\sqrt{s_1^2 + s_2^2}} + K_2 - \frac{\sqrt{s_1^2 + s_2^2}}{4s_1s_2} \right) \right] \quad (23)$$

Nếu $\frac{1}{2} \leq z < 1$ thì

$$h(z) = \frac{1}{2} K_1 e^{\frac{2C-B}{4}} e^{-\frac{B}{4}z^2 + \frac{B-C}{2}z} \left[\Phi \left(\frac{s_2(1-z)}{2s_1\sqrt{s_1^2 + s_2^2}} + K_2 \right) - \Phi \left(-\frac{s_1(1-z)}{2s_2\sqrt{s_1^2 + s_2^2}} + K_2 \right) \right] \quad (24)$$

5. KẾT LUẬN

- Về mặt lý thuyết có thể xác định được xác suất sai lầm tối thiểu trong phân loại hai tổng thể, nhưng việc tìm biểu thức giải tích cụ thể cho sai lầm này không phải là vấn đề đơn giản. Bài viết đã xem xét biểu thức giải tích cụ thể cho sai lầm này của hai phân phối chuẩn và beta.

- Xác định được biểu thức cụ thể cho hàm mật độ xác suất của tổng hai loại sai lầm phân loại khi giả sử mỗi sai lầm có phân phối chuẩn, mũ, beta trên $(0, 1/4)$, từ đó xác định khoảng cách của hai hàm mật độ xác suất.

- Vấn đề của bài viết có thể mở rộng cho nhiều tổng thể và có thể dùng một phần mềm toán nào đó để hỗ trợ.

TÀI LIỆU THAM KHẢO

- [1]. Andrew R. Webb,, *Statistical Pattern Recognition* John Wiley, London, (1999).
- [2]. Morris H.Degroot, *Probability and Statistics*, Addison-Wesley, United State, (1986).
- [3]. Pham-Gia T., Turkkan, N.and Bekker, A., *Bayesian Analysis in the L^1 – Norm of the Mixing*, Proportion using Discriminant Analysis, Metrika, (2005).
- [4]. Robert J.Boik and James F. Robison-Cox, *Derivatives of the Incomplete Beta Function*, Montana State University –Bozema, Montana, (1988).
- [5]. Tô Cẩm Tú, *Phân tích số liệu nhiều chiều*, NXB Khoa học và Kỹ thuật, Hà Nội, (2003).