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

(Bài nhận ngày 11 tháng 06 năm 2007, hoàn chỉnh sửa chữa ngày 18 tháng 09 năm 2007)

TÓM TẮT: *Bài báo quan tâm đến sai số trong phân loại hai tổng thể H_1 và H_2 bằng phương pháp Bayes. Thiết lập hàm mật độ xác suất cho tổng của hai loại sai lầm trong phân loại khi giả sử mỗi sai lầm có hàm mật độ xác suất trên (0,1/4), từ đó xác định khoảng cách L^1 giữa hai hàm mật độ xác suất theo Lissack và Fu. Các kết quả được xem xét cụ thể cho các phân phối chuẩn, mũ và beta.*

Từ khóa*: Sai số Bayes, khoảng cách L^1, phân phối chuẩn, mũ, beta.*

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 tiền nghiệm v của H_1 thì sai số Bayes được xác định $P_e = \int_R \min\{f_1(x),(1-f_2(x)\}dx$, và nếu quan tâm đến v thì $P_e = \int_R \min\{v.f_1(x),(1-v)f_2(x)\}dx$. P_e 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 H1

Phương trình f1(x) - f2(x) = 0 có thể có một nghiệm hoặc nhiều nghiệm. Giả sử f1(x) và f2(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 x0 thì ta có phân tích nhận dạng như sau: một phần tử với quan sát y được xếp vào H1 nếu $y \leq x0$ và xếp vào H2 nếu y > x0.

Đặt h(x) = min{f1(x), f2(x)}, khi đó:

τ = P(H2|H1) = $\int_{x>x_0} h(x)dx$: xác suất phân loại một phần tử vào H2 khi thật sự nó thuộc H1.

δ = P(H1|H2) = $\int_{x\leq x_0} h(x)dx$: xác suất phân loại một phần tử vào H1 khi thật sự nó thuộc H2.

Nếu phương trình có hai nghiệm x1 và x2 (giả sử x1 < x2) thì một phần tử với quan sát y sẽ được xếp vào H1 nếu $x_1 \leq y \leq x_2$ và xếp vào H2 nếu $y \notin [x_1, x_2]$. Khi đó:

$$\tau = \int_{\{x>x_2\}\cup\{x<x_1\}} h(x)dx \quad \text{và} \quad \delta = \int_{x_1\leq x\leq x_2} h(x)dx$$

Trong cả hai trường hợp ta có xác suất của phân loại sai lầm là ε = Pe = $\tau + \delta$. Chúng ta chứng minh được bất kỳ sự chọn lựa nào khác x0 hoặc x1 và x2 trong phân tích nhận dạng đều dẫn đến một xác suất sai lầm lớn hơn Pe, 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 H1

Đặt k1(x) = vf1(x), k2 = (1-v)f2(x), khi đó phương trình

k1(x) = k2(x) 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 τ_1 và δ_1 với

$$\tau_1 = \int_{R_1} k_1(x)dx \quad \text{và} \quad \delta_1 = \int_{R_2} k_2(x)dx$$

trong đó R1 = {x | k1(x)$\geq$k2(x) }và R2 = {x | k1(x)< k2(x) }. Miền R1 và R2 được xác định từ x'0 hoặc x'1 và x'2 ở trên. Khi đó xác suất sai lầm trong phân loại Pe = $\tau_1 + \delta_1$ cũng là nhỏ nhất.

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

$$\tau_1 = \frac{1}{2}\int_{\{x>x_2\}\cup\{x<x_1\}} h(x)dx \quad \text{và} \quad \delta_1 = \frac{1}{2}\int_{x_1\leq x\leq x_2} h(x)dx$$

Pe 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 ($\varepsilon = \tau = \delta = 0$)

Trường hợp này xảy ra khi f1(x) và f2(x) không cắt nhau, khi đó ta có thể ước lượng tỷ lệ của H1 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 Xi là đại lượng ngẫu nhiên ứng với quan sát thứ i mà Xi = 1 nếu phần tử quan sát thuộc H1 và Xi = 0 nếu phần tử quan sát không thuộc H1. Giả sử P (Xi = 1) = η và η có phân phối tiền nghiệm Beta(α, β), khi đó ta có các kết quả sau:

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

$$\varphi^{(n)}(\eta) = Beta(\eta, \alpha + y, \beta + n - y) \text{ trong đó } y = \sum_{i=1}^{n} X_i \quad (1)$$

2) Kỳ vọng hậu nghiệm của η là

$$\mu^{(n)}(\eta) = \frac{\alpha + y}{\alpha + \beta + n} \quad (2)$$

3) Phương sai hậu nghiệm của η là

$$Var^{(n)}(\eta) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \quad (3)$$

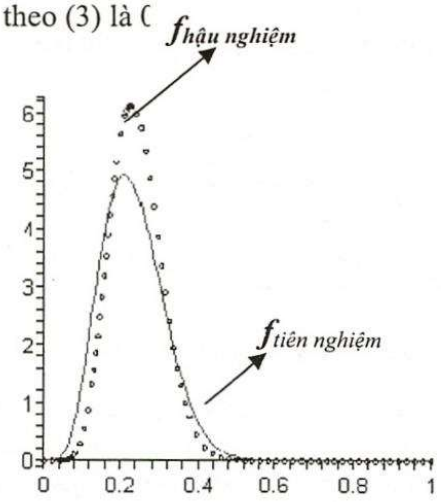
Đây là kết quả vận dụng định lý 1 ([2], trang 321) trong phân loại.

Ví dụ 1: Giả sử η 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 η theo (1) là Beta(10,32).

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

Phương sai hậu nghiệm của η theo (3) là (



Hình 1.Đồ thị hàm mật độ xác suất tiên nghiệm và hậu nghiệm của η (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 $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$, ta có hai trường hợp:

Trường hợp 1: Hai trung bình khác nhau $\mu_1 < \mu_2$.

Nếu $\sigma_1 = \sigma_2 = \sigma$ thì phương trình f1(x) - f2(x) = 0 có một nghiệm $x_0 = \frac{\mu_1 + \mu_2}{2}$

Khi đó ta có $\tau = \delta = 1 - \Phi(\xi)$ với $\xi = \frac{\mu_2 - \mu_1}{2\sigma}$ và $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$

Nếu $\sigma_1 \neq \sigma_2$ thì phương trình f1(x) - f2(x) = 0 có hai nghiệm sau:

$$x_i = \frac{(\mu_1\sigma_2^2 - \mu_2\sigma_1^2) \pm \sigma_1\sigma_2\sqrt{(\mu_1 - \mu_2)^2 + K}}{\sigma_2^2 - \sigma_1^2}, \quad i = 1, 2$$

trong đó, $K = 2(\sigma_2^2 - \sigma_1^2)\ln\frac{\sigma_2}{\sigma_1} \geq 0$, và nếu $x_1 \leq x_2$ thì

$$\tau = 1 - \Phi\left(\frac{x_2 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{x_1 - \mu_1}{\sigma_1}\right); \quad \delta = \Phi\left(\frac{x_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{x_1 - \mu_2}{\sigma_2}\right)$$

Trường hợp 2: $\mu_1 = \mu_2$.

Nếu $\sigma_1 \neq \sigma_2$. Trường hợp này phương trình f1(x) - f2(x) = 0 có hai nghiệm

$$x_i = \mu \pm \sigma_1\sigma_2\sqrt{E} \text{ với } E = \frac{2\ln\frac{\sigma_1}{\sigma_2}}{\sigma_2^2 - \sigma_1^2} \geq 0, \quad i = 1,2$$

Khi đó $\tau = 1 - \Phi(\sigma_2\sqrt{E}) + \Phi(-\sigma_2\sqrt{E}), \delta = \Phi(\sigma_1\sqrt{E}) - \Phi(-\sigma_1\sqrt{E})$.

Nếu $\sigma_1 = \sigma_2$. Trong trường hợp này ta có $\varepsilon = \tau = \delta = 1$.

Ví dụ 2: Trên hai tổng thể H1 và H2 ta quan sát biến ngẫu nhiên X1 và X2 lần lượt có phân phối chuẩn X1 ~ N(5, 92), X2 ~ N (18, 62). Nếu ta không quan tâm đến xác suất tiên nghiệm thì phương trình f1(x) = f2(x) có hai nghiệm x1 = 11.198, x2 = 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 H1, ngược lại ta sẽ xếp nó vào H2. Trong phân tích nhận dạng này

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

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

và xác suất sai lầm trong phân loại là $\varepsilon = \tau + \delta = 0.3739$.

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

2.2.2.Hai tổng thể có phân phối Beta

Giả sử $X_1 \sim$ Beta(α_1, β_1); X2 $\sim$ $Beta(\alpha_2, \beta_2)$

Xét phương trình f1(x) = f2(x) $\Leftrightarrow \frac{x^{\alpha_1 - 1}(1-x)^{\beta_1 - 1}}{B(\alpha_1, \beta_1)} = \frac{x^{\alpha_2 - 1}(1-x)^{\beta_2 - 1}}{B(\alpha_2, \beta_2)}$

$$\Leftrightarrow x^{\alpha_1 - \alpha_2}(1-x)^{\beta_1 - \beta_2} = \frac{B(\alpha_1, \beta_1)}{B(\alpha_2, \beta_2)}$$

$$\Leftrightarrow x^{\alpha}(1-x)^{\beta} = A$$

Trong đó, $\alpha = \alpha_1 - \alpha_2; \beta = \beta_1 - \beta_2; A = \frac{B(\alpha_1, \beta_1)}{B(\alpha_2, \beta_2)}$.

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

xk - xk+1 = B (4)

Phương trình (4) có thể giải được trên máy tính, ta tìm được hoành độ giao điểm của hai hàm mật độ f1(x) và f2(x) và từ đó ta có thể tính được $\tau = P(H_2 \mid H_1)$ và $\delta = P(H_1 \mid H_2)$. Việc tính τ và δ dẫn đến việc tính tích phân của hàm Beta

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

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

F(x) = $K_{x,\alpha,\beta}$ 2F1$(1-\beta, 1; \alpha + 1; \frac{-x}{1-x})$ (6)

trong đó, $K_{x,\alpha,\beta} = \frac{x^{\alpha}(1-x)^{\beta - 1}}{\alpha B(\alpha, \beta)}$, 2F1 (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

$$2F1 \approx \cfrac{1+\cfrac{a_1}{b_1+\cfrac{a_2}{b_2+\cfrac{a_3}{b_3+\cfrac{a_4}{\cdots\cdots}}}}}{+\cfrac{a_n}{b_n}}$$

trong đó,

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

$$b_n = \frac{2(\alpha.f+2\beta)n^2 + 2(\alpha f+2\beta)(\alpha-1)n + \alpha\beta(\alpha-2-\alpha f)}{\beta(\alpha+2n-2)(\alpha+2n)}; \quad n \geq 1$$

$$f = \frac{\beta x}{\alpha(1-x)}$$

Nhận xét. Trong trường hợp đặc biệt $\alpha_1 = \beta_1 = p$ và $\alpha_2 = \beta_2 = q$, 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 x2 - 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 τ và δ cùng với Pe cũng là biến ngẫu nhiên. Theo Lissack và Fu thì 2Pe =1 - J1(H1,H2|v) với Z = J1(H1,H2|v) là khoảng cách L1 giữa vf1(x) và (1-v)f2(x). Từ mối quan hệ này, khi không biết về f1(x) và f2(x) cũng như v nhưng chúng ta có thông tin về hai xác suất sai lầm τ và δ 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 X1 và X2 là hai biến ngẫu nhiên độc lập trên $(0, \frac{1}{4})$ có hàm mật độ xác suất lần lượt là f1(x), f2(x). Xét Y = X1 + X2 , khi đó hàm mật độ xác suất của Y có dạng:

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

Chứng minh

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

Vì X2 là biến ngẫu nhiên trên $(0,\frac{1}{4})$ nghĩa là f2 (x) = 0 $\forall x\notin(0,\frac{1}{4})$, 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,

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

Vì X1 và X2 $\in (0;\frac{1}{4})$ nên y $\in (0;\frac{1}{2})$.

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

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

3.2 Một số trường hợp cụ thể của Y = X1 + X2

3.2.1.X1, X2 là hai biến ngẫu nhiên độc lập có phân phối Beta trên $(0,\frac{1}{4})$

Giả sử $X_1 \sim Beta(\alpha_1, \beta_1; 0, \frac{1}{4}); X_2 \sim Beta(\alpha_2, \beta_2; 0, \frac{1}{4})$ với $\alpha_1, \beta_1, \alpha_2, \beta_2 > 0$.

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

Nếu $0 < y \le \frac{1}{4}$ thì

$$g(y) = H_1 4^{\alpha_1+\alpha_2} y^{\alpha_1+\alpha_2-1}(1-4y)^{\beta_1-1}.F_D^{(2)}(\alpha_2, 1-\beta_1, 1-\beta_2; \alpha_1+\alpha_2; \frac{4y}{4y-1}, 4y) \quad (9)$$

với $H_1 = \frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_1+\alpha_2)\Gamma(\beta_1)\Gamma(\beta_2)}$; $F_D^{(2)}$ là hàm siêu bội với hai biến số.

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

g(y)

$$= H_2 2^{\beta_1+\beta_2+1}(1-2y)^{\beta_1+\beta_2-1}(4y-1)^{\alpha_2-1}.F_D^{(2)}(\beta_2, 1-\alpha_1, 1-\alpha_2; \beta_1+\beta_2; 2-4y, \frac{4y-2}{4y-1}) \quad (10)$$

với $H_2 = \frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_2+\beta_2)}{\Gamma(\beta_1+\beta_2)\Gamma(\alpha_1)\Gamma(\alpha_2)}$.

3.2.2.X1, X2 là hai biến ngẫu nhiên độc lập có phân phối mũ cắt trên $(0, \frac{1}{4})$

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

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

Khi $b_1 > b_2$,

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

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

với $b_1, b_2 > 0$ 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 $b_1 = b_2 = c$,

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

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

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

$$f_1(x) = \frac{b_1}{a}e^{-b_1x}; f_2(x) = \frac{b_2}{b}e^{-b_2x}$$

Nếu $0 < y \le \frac{1}{4}$, theo (7) thì $g(y) = \int_0^y f_1(t)f_2(y-t)dt = \frac{b_1b_2}{ab}e^{-b_2y}\int_0^y e^{-(b_1-b_2)t}dt$

Vì $b_1 > b_2$ nên $g(y) = \frac{b_1b_2}{ab(b_1-b_2)}e^{-b_2y}\left[1 - e^{-(b_1-b_2)y}\right] = \frac{b_1b_2}{ab(b_1-b_2)}\left[e^{-b_2y} - e^{-b_1y}\right]$

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

$$g(y) = \frac{b_1b_2}{ab}e^{-b_2y}\int_{y-\frac{1}{4}}^{\frac{1}{4}} e^{-(b_1-b_2)t}dt = \frac{b_1b_2}{ab(b_1-b_2)}e^{-b_2y}\left[e^{-(b_1-b_2)(y-\frac{1}{4})} - e^{-(b_1-b_2)/4}\right]$$

$$= \frac{b_1b_2}{ab(b_1-b_2)}\left[e^{-\frac{b_2-b_1+4b_1y}{4}} - e^{-\frac{b_1-b_2+4b_2y}{4}}\right]$$

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

Nếu $0 < y \le \frac{1}{4}$ thì $g(y) = \frac{b_1b_2}{ab}e^{-b_2y}\int_0^y 1dt = \left(\frac{c}{d}\right)^2 y.e^{-cy}$.

Nếu $\frac{1}{4} < y < \frac{1}{2}$ thì $g(y) = \frac{b_1b_2}{ab}e^{-b_2y}\int_{y-\frac{1}{4}}^{\frac{1}{4}} 1dt = \left(\frac{c}{d}\right)^2 (\frac{1}{2} - y)e^{-cy}$.

2.2.3.Nếu X1, X2 là hai biến ngẫu nhiên độc lập có phân phối chuẩn cắt trên $(0, \frac{1}{4})$

Giả sử $X1 \sim N(\mu_1, \sigma_1^2; 0, \frac{1}{4})$, $X2 \sim N(\mu_2, \sigma_2^2; 0, \frac{1}{4})$ với $\mu_1, \mu_2 \in R$ và $\sigma_1, \sigma_2 \in R^+$.

Nếu $0 < y \le \frac{1}{4}$ thì

$$g(y) = K_1e^{-By^2+Cy}.\left[\Phi\left(\frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}}y + K_2\right) - \Phi\left(-\frac{\sigma_1}{\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}y + K_2\right)\right] \quad (15)$$

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

$$A = \frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2};\ B = \frac{1}{2(\sigma_1^2+\sigma_2^2)};\ C = \frac{\mu_1+\mu_2}{\sigma_1^2+\sigma_2^2}$$

$$K1 = \frac{1}{ab\sqrt{2\pi(\sigma_1^2+\sigma_2^2}}e^{-\frac{(\mu_1+\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}};\ K_2 = \frac{\mu_2\sigma_1^2-\mu_1\sigma_2^2}{\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}$$

$$a = \int_0^{1/4}\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}dx = \Phi\left(\frac{1-4\mu_1}{4\sigma_1}\right)-\Phi\left(\frac{-\mu_1}{\sigma_1}\right)$$

$$b = \int_0^{1/4}\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}dx = \Phi\left(\frac{1-4\mu_2}{4\sigma_2}\right)-\Phi\left(\frac{-\mu_2}{\sigma_2}\right)$$

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

$$g(y) = K_1 e^{-By^2+Cy}\cdot\left[\Phi\left(-\frac{\sigma_1}{\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}y + K_2 + \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{4\sigma_1\sigma_2}\right) - \Phi\left(\frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}}y + K_2 - \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{4\sigma_1\sigma_2}\right)\right] \quad (16)$$

Chứng minh

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

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

$$= \frac{1}{2\pi ab\sigma_1\sigma_2}\int_0^y e^{-\frac{(t-\mu_1)^2}{2\sigma_1^2}}.e^{-\frac{(y-t-\mu_2)^2}{2\sigma_2^2}}dt \qquad = \frac{1}{2\pi ab\sigma_1\sigma_2}e^{Q}\int_0^y e^{-(At^2+Pt)}.dt$$

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

$$P = \frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2} - \frac{y}{\sigma_2^2};\ Q = \frac{\mu_2 y}{\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{2\sigma_2^2} - \frac{y^2}{2\sigma_2^2}$$

Vì $-(At2 + Pt) = \frac{-\left(\sqrt{2A}t + \frac{P}{\sqrt{2A}}\right)^2}{2} + \frac{P^2}{4A}$ nên $\int_0^y e^{-(At^2+Pt)}dt = e^{\frac{P^2}{4A}}.\int_0^y e^{-\frac{-(\sqrt{2A}t+\frac{P}{\sqrt{2A}})^2}{2}}dt$.

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

$t = y$, $u = \sqrt{2A}\,y + \dfrac{P}{\sqrt{2A}}$.

Từ đó,

$$\int_0^y e^{-(At^2+Pt)}\,dt = \frac{e^{\frac{P^2}{4A}}}{\sqrt{2A}} \int_{\frac{P}{\sqrt{2A}}}^{\sqrt{2A}y+\frac{P}{\sqrt{2A}}} e^{-\frac{u^2}{2}}\,du$$

$$= \frac{\sqrt{2\pi}e^{\frac{P^2}{4A}}}{\sqrt{2A}} \frac{1}{\sqrt{2\pi}} \int_{\frac{P}{\sqrt{2A}}}^{\sqrt{2A}y+\frac{P}{\sqrt{2A}}} e^{-\frac{u^2}{2}}\,du$$

$$= \frac{\sqrt{\pi}e^{\frac{P^2}{4A}}}{\sqrt{A}} \left[\Phi\left(\sqrt{2A}y + \frac{P}{\sqrt{2A}} \right) - \Phi\left(\frac{P}{\sqrt{2A}} \right) \right]$$

Thế tích phân này vào g(y) ta có

$$g(y) = \frac{1}{2ab\sigma_1\sigma_2\sqrt{\pi A}} e^{\frac{P^2}{4A}+Q} \left[\Phi\left(\sqrt{2A}y + \frac{P}{\sqrt{2A}} \right) - \Phi\left(\frac{P}{\sqrt{2A}} \right) \right]$$

Vì $\dfrac{P^2}{4A} + Q = -\dfrac{y^2}{2(\sigma_1^2+\sigma_2^2)} + \dfrac{(\mu_1+\mu_2)y}{\sigma_1^2+\sigma_2^2} - \dfrac{(\mu_1+\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}$ nên

$$\frac{1}{2ab\sigma_1\sigma_2\sqrt{\pi A}} e^{\frac{P^2}{4A}+Q} = K_1 e^{-By^2+Cy}$$

$$\text{UP} = \sqrt{2A}y + \frac{P}{\sqrt{2A}} = \frac{y}{\sigma_1^2\sqrt{\dfrac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} + \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2\sigma_2^2\sqrt{\dfrac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} = \left[\frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}} \right] y + K_2$$

$$\text{LP} = \frac{P}{\sqrt{2A}} = -\frac{y}{\sigma_2^2\sqrt{\dfrac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} + \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2\sigma_2^2\sqrt{\dfrac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} = -\left[\frac{\sigma_1}{\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}} \right] y + K_2$$

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

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

$$\int_{y-\frac{1}{4}}^{\frac{1}{4}} e^{-(At^2+Pt)}dt = \frac{e^{\frac{P^2}{4A}}}{\sqrt{2A}} \int_{\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}}^{\frac{A+2P}{2\sqrt{2A}}} e^{-\frac{u^2}{2}}du = \frac{\sqrt{2\pi}e^{\frac{P^2}{4A}}}{\sqrt{2A}}\frac{1}{\sqrt{2\pi}} \int_{\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}}^{\frac{A+2P}{2\sqrt{2A}}} e^{-\frac{u^2}{2}}du$$

$$= \frac{\sqrt{\pi}e^{\frac{P^2}{4A}}}{\sqrt{A}}\left[\Phi\left(\frac{A+2P}{2\sqrt{2A}}\right) - \Phi\left(\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}\right)\right]$$

Khi đó $g(y) = K_1 e^{-By^2+Cy}.\left[\Phi\left(\frac{A+2P}{2\sqrt{2}A}\right) - \Phi\left(\frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}\right)\right]$.

UP $= \frac{A+2P}{2\sqrt{2A}}$

=

$$-\frac{y}{\sigma_2^2\sqrt{\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} + \frac{\sigma_1^2+\sigma_2^2+4\mu_2\sigma_1^2-4\mu_1\sigma_2^2}{4\sigma_1^2\sigma_2^2\sqrt{\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} = -\frac{\sigma_1}{\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}y + K_2 + \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{4\sigma_1\sigma_2}$$

LP $= \frac{2A(y-\frac{1}{4})+P}{\sqrt{2A}}$.

=

$$\frac{y}{\sigma_1^2\sqrt{\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} + \frac{-\sigma_1^2-\sigma_2^2+4\mu_2\sigma_1^2-4\mu_1\sigma_2^2}{4\sigma_1^2\sigma_2^2\sqrt{\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}} = \frac{\sigma_2}{\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}}y + K_2 - \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{4\sigma_1\sigma_2}$$

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

Nhận xét. Tùy theo giá trị của μ và σ 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 VỀ HÀM MẬT ĐỘ CỦA *Z*

Ta có $Z = 1 - 2P_e = 1 - 2y$ ($y = \tau + \delta$, 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 τ và δ là hai biến ngẫu nhiên độc lập có phân phối Beta trên $(0,\frac{1}{4})$

Giả sử $\tau \sim Beta(\alpha_1,\beta_1;0,\frac{1}{4}); \delta \sim Beta(\alpha_2,\beta_2;0,\frac{1}{4})$

Nếu $0 < z < \frac{1}{2}$ thì

$$h(z) = H_2 2^{\beta_1+\beta_2} z^{\beta_1+\beta_2-1}(1-2z)^{\alpha_2-1}.F_D^{(2)}(\beta_2,1-\alpha_1,1-\alpha_2;\beta_1+\beta_2;2z,\frac{2z}{2z-1}) \quad (17)$$

Nếu $\frac{1}{2} \le z < 1$ thì

$$h(z) = H_1 2^{\alpha_1+\alpha_2}(1-z)^{\alpha_1+\alpha_2-1}(2z-1)^{\beta_1-1}.F_D^{(2)}(\alpha_2,1-\beta_1,1-\beta_2;\alpha_1+\alpha_2;\frac{2-2z}{1-2z},2-2z) \quad (18)$$

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

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

Giả sử $\tau \sim$ Exp(b1; 0, $\frac{1}{4}$) ; $\delta \sim$ Exp(b2; 0, $\frac{1}{4}$) .

Khi b1 > b2 :

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

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

Khi b1 = b2 :

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

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

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

Giả sử $\tau \sim N(\mu_1,\sigma_1^2;0,\frac{1}{4})$ và $\delta \sim N(\mu_2,\sigma_2^2;0,\frac{1}{4})$.

Nếu $0 < z < \frac{1}{2}$ thì

$$h(z) = \frac{1}{2}K_1 e^{\frac{2C-B}{4}} e^{-\frac{B}{4}z^2+\frac{B-C}{2}z} \cdot \left[\Phi\left(-\frac{\sigma_1(1-z)}{2\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}} + K_2 + \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{4\sigma_1\sigma_2} \right) - \right.$$

$$\left. - \Phi\left(\frac{\sigma_2(1-z)}{2\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}} + K_2 - \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{4\sigma_1\sigma_2} \right) \right] \quad (23)$$

Nếu $\frac{1}{2} \le z < 1$ thì

$$h(z) = \frac{1}{2}K_1 e^{\frac{2C-B}{4}} e^{-\frac{B}{4}z^2+\frac{B-C}{2}z} \cdot \left[\Phi\left(\frac{\sigma_2(1-z)}{2\sigma_1\sqrt{\sigma_1^2+\sigma_2^2}} + K_2 \right) - \Phi\left(-\frac{\sigma_1(1-z)}{2\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}} + K_2 \right) \right] \quad (24)$$

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.

BAYES ERROR AND DISTANCE BETWEEN TWO PROBABILITY DISTRIBUTION FUNCTIONS IN CLASSIFICATION OF TWO POPULATIONS

Vo Van Tai[(1)], Pham Gia Thu[(2)], To Anh Dung[(3)]
(1) Can tho University
(2) Moncton University, Canada
(3) University of Natural Sciences, VNU-HCM

ABSTRACT: *The article looks at the error in classification of two populations H_1 and H_2 by Bayesian method. Establishing probability distribution function for sum of two misclassification probabilities in classification when supposing they have a distribution in (0,1/4), from there to consider L^1-distance between two probability distribution functions by Liassak and Fu. This problem is considered by objectifying for normal, exponential and beta distributions.*

Keywords: *Bayes error, L^1 - distance, normal distribution, exponential, beta.*

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).