ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

\- - - - - - - - - - - - - - - - - - -

TRẦN ANH TUẤN

THỐNG KÊ BAYES NHIỀU CHIỀU VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC

HÀ NỘI - 2015

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

- - - - - - - - - - - - - - - - - - -

TRẦN ANH TUẤN

THỐNG KÊ BAYES NHIỀU CHIỀU VÀ ỨNG DỤNG

Chuyên ngành: Lí thuyết xác suất và thống kê Toán

Mã số: 60 46 01 06

LUẬN VĂN THẠC SĨ KHOA HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC
GS.TSKH. ĐẶNG HÙNG THẮNG

HÀ NỘI - 2015

Mục lục

Lời cảm ơn

Trước khi trình bày nội dung chính của luận văn, Tác giả xin bày tỏ lòng biết ơn sâu sắc tới GS. TSKH. Đặng Hùng Thắng người Thầy đáng kính đã luôn tận tình chỉ bảo giúp đỡ tác giả trong suốt thời gian qua.

Mặc dù có nhiều cố gắng, song trong quá trình thực hiện luận văn Tác giả không tránh khỏi những thiếu sót. Vì vậy, Tác giả rất mong nhận được ý kiến đóng góp của Thầy Cô và bạn bè đồng nghiệp, để luận văn được hoàn thiện hơn.

Tác giả xin chân thành cảm ơn!

Hà Nội, ngày 10 tháng 10 năm 2015.

Học viên

Trần Anh Tuấn

Lời nói đầu

Hiện tại thống kê có hai trường phái: Thống kê tần suất và thống kê Bayes. Thống kê tần suất đã ra đời trước, là phương pháp phổ biến hiện nay. Nó dựa trên những kết quả quan sát mẫu của hiện tại mà không cần để ý đến những thông tin, dữ liệu đã biết trước. Thống kê Bayes dựa trên những thông tin dữ liệu đã biết trước về vấn đã quan sát để suy luận cho những thống kê hiện tại.

Suy luận Bayes được sử dụng rất rộng rãi trong tất cả các ngành nghề như y học, kinh tế, tin học,... Đặc biệt trong xác suất và thống kê hiện nay nó đóng vai trò cũng hết sức quan trọng. Hiện tại chúng ta tìm được một số biểu thức giải tích hậu nghiệm cụ thể khi giả sử tiên nghiệm là các hàm mật độ xác suất thông dụng như Beta, mũ, chuẩn,... Trong thống kê sử dụng định lí Bayes cho ước lượng và kiểm định tham số thống kê, cũng như các bài toán phân loại ngày nay trở nên phổ biến.

Trong đề tài luận văn này, tác giả trình bày một số kiến thức cơ bản về thống kê Bayes nhiều chiều và mô hình hồi quy Bayes đồng thời đưa ra một số ứng dụng cơ bản của hồi quy Bayes.

Luận văn của tác giả được chia làm 3 chương.

Chương 1. *Các phân phối xác suất nhiều chiều quan trọng.*

Chương 2. *Mở đầu về thống kê Bayes nhiều chiều.*

Chương 3. *Hồi quy Bayes và áp dụng.*

Chương 1

Các phân phối xác suất nhiều chiều quan trọng

1.1 Phân phối nhiều chiều

1.1.1 Phân phối chuẩn nhiều chiều

Một biến ngẫu nhiên tuân theo quy luật p-biến ngẫu nhiên phân phối chuẩn nhiều chiều với vectơ kì vọng μ

và ma trận hiệp phương sai Σ được kí hiệu là

$$x|\mu, \Sigma \sim N(\mu, \Sigma), \tag{1.1}$$

ở đây tham số (μ, Σ) được cho bởi

$$p(x|\mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \tag{1.2}$$

1.1.2 Phân phối Student nhiều chiều t

Một biến ngẫu nhiên tuân theo t-phân phối Student nhiều chiều được kí hiệu là

$$t|\nu, t_0, \Sigma, \phi^2 \sim t(\nu, t_0, \Sigma, \phi^2), \tag{1.3}$$

ở đây tham số $(\nu, t_0, \Sigma, \phi^2)$ được cho bởi

$$p(t|\nu, t_0, \Sigma, \phi^2) = \frac{k_t (\phi^2)^{-\frac{\nu}{2}} |\Sigma|^{-\frac{1}{2}}}{\left[\phi^2 + \frac{1}{\nu}(t - t_0)'\Sigma^{-1}(t - t_0))\right]^{\frac{\nu+p}{2}}}, \tag{1.4}$$

ở đây

$$k_t = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu\pi)^{\frac{p}{2}}\Gamma\left(\frac{\nu}{2}\right)}, \tag{1.5}$$

1.2 Phân phối của ma trận ngẫu nhiên

1.2.1 Phân phối chuẩn ma trận

Một ma trận ngẫu nhiên có phân phối chuẩn ma trận $n \times p$ được kí hiệu

$$X|M, \Sigma, \Phi \sim N(M, \Phi \otimes \Sigma) \tag{1.6}$$

ở đây (M, Σ, Φ) là các tham số của phân phối trên với

$$X \in \mathbb{R}^{x \times p},\ M \in \mathbb{R}^{x \times p},\ \Sigma, \Phi > 0. \tag{1.7}$$

các ma trận Σ và Φ thường được gọi là ma trận hiệp phương sai trong và giữa.

1.2.2 Phân phối Wishart

Một $p \times p$ ma trận đối xứng G tuân theo phân phối Wishart được kí hiệu

$$G|\Upsilon, p, \nu_0 \sim W(\Upsilon, p, \nu_0) \tag{1.8}$$

ở đây tham số (Υ, p, ν_0) được cho bởi

$$p(G|\Upsilon, p, \nu_0) = k_W |\Upsilon|^{-\frac{\nu_0}{2}} |G|^{\frac{\nu_0 - p - 1}{2}} e^{-\frac{1}{2} tr \Upsilon^{-1} G}, \tag{1.9}$$

ở đây

$$k_W^{-1} = 2^{\frac{\nu_0 p}{2}} \pi^{\frac{p(p-1)}{2}} \prod_{j=1}^{p} \Gamma\left(\frac{\nu_0 + 1 - j}{2}\right) \tag{1.10}$$

1.2.3 Phân phối Wishart nghịch đảo

Một $p \times p$ ma trận ngẫu nhiên Σ tuân theo phân phối Wishart nghịch đảo được kí hiệu

$$\Sigma|Q, p, \nu \sim IW(Q, p, \nu) \tag{1.11}$$

ở đây tham số $(\Sigma|Q, p, \nu)$ được cho bởi

$$p(\Sigma|\nu, Q) = k_{IW} |Q|^{\frac{\nu-p-1}{2}} |\Sigma|^{-\frac{\nu}{2}} e^{-\frac{1}{2} tr \Sigma^{-1} Q}, \tag{1.12}$$

ở đây

$$k_{IW}^{-1} = 2^{\frac{(\nu-p-1)p}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{\nu - p - j}{2}\right) \tag{1.13}$$

1.2.4 Phân phối ma trận T

Một biến ngẫu nhiên T tuân theo T-phân phối ma trận Student được kí hiệu là

$$T|\nu, T_0, \Sigma, \Phi \sim T(\nu, T_0, \Sigma, \Phi), \tag{1.14}$$

ở đây tham số (ν, T_0, Σ, Φ) được cho

$$p(T|\nu, T_0, \Sigma, \Phi) = k_T \frac{|\Phi|^{\frac{\nu}{2}}|\Sigma|^{-\frac{n}{2}}}{\left|\Phi + \frac{1}{\nu}(T - T_0)\Sigma^{-1}(T - T_0)'\right|^{\frac{\nu+p}{2}}}, \tag{1.15}$$

ở đây

$$k_T = \frac{\Pi_{j=1}^{n}\Gamma\left(\frac{\nu+p+1-j}{2}\right)}{(\nu\pi)^{\frac{np}{2}}\Pi_{j=1}^{n}\Gamma\left(\frac{\nu+1-j}{2}\right)}. \tag{1.16}$$

1.3 Vectơ ngẫu nhiên liên tục

Hàm hợp lí:

$$p(x_1, \ldots, x_n|\theta_1, \ldots, \theta_J) = \Pi_{i=1}^{n} p(x_i|\theta_1, \ldots, \theta_J). \tag{1.17}$$

Hậu nghiệm:

$$p(\theta_1, \ldots, \theta_J|x_1, \ldots, x_n) = \frac{p(\theta_1, \ldots, \theta_J)p(x_1, \ldots, x_n|\theta_1, \ldots, \theta_J)}{p(x_1, \ldots, x_n)}, \tag{1.18}$$

1.4 Ma trận ngẫu nhiên liên tục

Tiên nghiệm:

$$p(\theta_1, \ldots, \theta_J), \tag{1.19}$$

ở đây θ có thể nhận giá trị là ma trận và các tham số không cần độc lập.

Hàm hợp lí:

$$p(X_1, \ldots, X_n|\theta_1, \ldots, \theta_J) = \Pi_{i=1}^n p(X_i|\theta_1, \ldots, \theta_J). \tag{1.20}$$

Hậu nghiệm:

$$p(\theta_1, \ldots, \theta_J|X_1, \ldots, X_n) = \frac{p(\theta_1, \ldots, \theta_J)p(X_1, \ldots, X_n|\theta_1, \ldots, \theta_J)}{p(x_1, \ldots, x_n)}, \tag{1.21}$$

Chương 2

Mở đầu về thống kê Bayes nhiều chiều

2.1 Phân phối tiên nghiệm

2.1.1 Phân phối tiên nghiệm mơ hồ

Phân phối tiên nghiệm mơ hồ là phân phối tiên nghiệm không có thông tin có thể dựa trên bất kì một tham số là bị chặn (có một miền giá trị hữu hạn) hoặc không bị chặn (có một miền giá trị vô hạn).

2.1.2 Phân phối tiên nghiệm liên hợp

Phân phối tiên nghiệm liên hợp là phân phối tiên nghiệm có thông tin.

2.1.3 Phân phối tiên nghiệm tổng quát

Phân phối tiên nghiệm liên hợp tổng quát được tìm ra bằng cách viết dưới hàm hợp lí, bằng cách tráo đổi vai trò của biến ngẫu nhiên và tham số, chúng làm tốt hơn phân phối vì vậy nó không phụ thuộc vào dữ liệu, và giả sử rằng phân phối tiên nghiệm trên mỗi tham số là độc lập.

2.1.4 Vectơ ngẫu nhiên

2.1.4.1 Phân phối chuẩn

Bảng 2.1: Phân phối tiên nghiệm liên hợp vectơ tổng quát

Hàm hợp lí	Các tham số	Họ tiên nghiệm
biết Σ	μ	GMN
biết μ	Σ	Whishart nghịch đảo
Chuẩn nhiều chiều	(μ, Σ)	GMN-IW

2.1.4.2 Ma trận ngẫu nhiên

Bảng 2.2: Ma trận tiên nghiệm liên hợp tổng quát

Hàm hợp lí	Các tham số	Họ tiên nghiệm
biết (Φ, Σ)	M	GMN
biết (M, Φ)	Σ	Whishart nghịch đảo
biết (M, Σ)	Φ	Whishart nghịch đảo
Chuẩn ma trận	(M, Φ, Σ)	GMN-IW-IW

2.1.5 Phân phối tiên nghiệm tương quan

Trong mục này, phân phối tiên nghiệm liên hợp là được sinh ra từ hệ số tương quan giữa các vectơ quan sát. Trong phạm vi nhân tố giải tích Bayes, một phân phối Beta tổng quát được sử dụng cho hệ số tương quan ρ trong ma trận giữa vectơ tương quan Φ. Điều này được thực hiện khi Φ hoặc là tương quan trong lớp ma trận

$$\Phi = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ & 1 & \rho & \cdots & \rho \\ & & \ddots & & \vdots \\ & & & & \rho \\ & & & & 1 \end{pmatrix} = (1-\rho)I_n + \rho e_n e_n', \quad (2.1)$$

ở đây e_n là một vectơ cột của những thứ và $-\dfrac{1}{n-1} < \rho < 1$ hoặc thứ tự đầu tiên ma trận tương quan Markov

$$\Phi = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & & \cdots & 1 \end{pmatrix} \quad (2.2)$$

ở đây $0 < |\rho| < 1$.

2.2 Đánh giá siêu tham số

2.2.1 Hàm hợp lí phân phối chuẩn nhiều chiều

Trước khi biểu diễn một phép thử mà đạt được các biến ngẫu nhiên phân phối chuẩn nhiều chiều, bằng sự hiểu biết chúng ta nhìn thấy rằng một thí nghiệm tương tự đã được thực hiện và dữ liệu tồn tại trong dạng của n_0 quan sát $x_1, x_2, \ldots, x_{n_0}$. Hàm hợp lí của n_0 biến ngẫu nhiên này là

$$p(x_1, \ldots, x_{n_0}|\mu, \Sigma) \propto |\Sigma|^{-\frac{n_0}{2}} e^{-\frac{1}{2}\sum\limits_{i=1}^{n_0}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)}. \quad (2.3)$$

2.2.2 Hàm hợp lí phân phối chuẩn ma trận

Hàm hợp lí của n_0 biến ngẫu nhiên $x_1, x_2, \ldots, x_{n_0}$ là

$$p(X_1, \ldots, X_{n_0}|M, \Sigma, \Phi) \propto |\Sigma|^{-\frac{n_0 n_1}{2}} |\Phi|^{-\frac{n_0 p_1}{2}} e^{-\frac{1}{2} tr\Phi^{-1}\sum\limits_{i=1}^{n_0}(X_i-M)'\Sigma^{-1}(X_i-M)} \quad (2.4)$$

2.3 Phương pháp ước lượng Bayes

2.3.1 Trung bình biên duyên hậu nghiệm

Thông thường chúng ta có một tập các tham số, $\theta = (\theta_1, \ldots, \theta_J)$ trong phân phối hậu nghiệm của chúng ta $p(\theta|X)$, ở đây X biểu diễn cho dữ liệu, đó có thể là tập hợp các quan sát vô hướng, vectơ hoặc ma trận. Phân phối biên duyên hậu nghiệm của bất kì các tham số θ_j có thể đạt được bằng cách lấy tích phân $p(\theta|X)$

đối với tất cả các tham số trừ θ_j. Có nghĩa là, phân phối biên duyên hậu nghiệm của θ_j là

$$p(\theta_j|X) = \int p(\theta_1, \ldots, \theta_J|X)d\theta_1 \ldots d\theta_{j-1}d\theta_{j+1} \ldots d\theta_J \tag{2.5}$$

ở dây tích phân được lấy trên miền thích hợp của tập các tham số.

2.3.2 Tối đa hóa hậu nghiệm

Để xác định một ước lượng tối đa hóa hậu nghiệm kết hợp, lấy vi phân hàm phân phối hậu nghiệm

$$p(\theta_1, \ldots, \theta_J|X) \tag{2.6}$$

đối với mỗi tham số, cho kết quả bằng không

$$\frac{\partial}{\partial\theta_1}p(\theta_1, \ldots, \theta_J|X)\Big|_{\theta_1=\hat{\theta}_1,\ldots,\theta_J=\hat{\theta}_J} = \cdots = \frac{\partial}{\partial\theta_J}p(\theta_1, \ldots, \theta_J|X)\Big|_{\theta_1=\hat{\theta}_1,\ldots} \tag{2.7}$$

giải hệ J phương trình trên với J ẩn chúng ta tìm được ước lượng tối đa hóa hậu nghiệm kết hợp là

$$\hat{\theta}_j = \underset{\theta_j}{\text{Arg Max}} \quad p(\theta_j|\hat{\theta}_1, \ldots, \hat{\theta}_J, X) \tag{2.8}$$

$$= \hat{\theta}_j(\hat{\theta}_1, \ldots, \hat{\theta}_J, X) \tag{2.9}$$

Chương 3

Hồi quy Bayes và áp dụng

3.1 Mô hình hồi quy tuyến tính đa biến

Trong mô hình hồi quy tuyến tính bội, cặp (x_i, u_i) là các quan sát, $i = 1, \ldots, n$, ở đây x_i là một vô hướng nhưng u_i là một vectơ $(q+1)$ chiều chứa 1 và các quan sát có thể $u_{i1}, \ldots, u_{iq}$. Chúng ta áp dụng mô hình

$$\underset{1\times 1}{x_i} = \underset{1\times(q+1)}{\beta'}\underset{(q+1)\times 1}{u_i} + \underset{1\times 1}{\epsilon} \tag{3.1}$$

ở đây

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_q \end{pmatrix}, \; u_i = \begin{pmatrix} 1 \\ u_{i1} \\ \vdots \\ u_{iq} \end{pmatrix}, \tag{3.2}$$

và sau đó phù hợp với một đường tới dữ liệu mẫu.

3.2 Hồi quy Bayes nhiều biến

Mô hình hồi quy Bayes có dạng

$$\underset{(n\times p)}{(X|B,U)} = \underset{[n\times(q+1)]}{U}\underset{[(q+1)\times p]}{B'} + \underset{(n\times p)}{E}, \tag{3.3}$$

ở đây ma trận của các biến độc lập X, các biến phụ thuộc U, và ma trận sai số E là được định nghĩa

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}, \; U = \begin{pmatrix} u_1' \\ \vdots \\ u_n' \end{pmatrix}, \; E = \begin{pmatrix} \epsilon_1' \\ \vdots \\ \epsilon_n' \end{pmatrix}, \tag{3.4}$$

trong khi ma trận B xác định như trên.

3.3 Áp dụng

3.3.1 Xét nghiệm Insulin

Bảng 3.1: Phân phối tiên nghiệm ma trận liên hợp

Các biến X		Các biến U
X_1 Nồng độ tại thời điểm 0	U_1	Loại Insulin
X_2 Nồng độ tại thời điểm 1	U_2	Mức liều lượng
X_3 Nồng độ tại thời điểm 2	U_3	U_1*U_2
X_4 Nồng độ tại thời điểm 3		
X_5 Nồng độ tại thời điểm 4		
X_6 Nồng độ tại thời điểm 5		

3.3.2 Bữa tiệc Cocktail

Tại một bữa tiệc cocktail, có p micro để ghi lại hoặc quan sát m người tham gia hoặc các diễn giả trong khoảng thời gian n. Kí hiệu này là phù hợp với thống kê nhiều chiều. Các cuộc đàm thoại được quan sát bao gồm của sự pha trộn các cuộc đàm thoại không thực sự quan sát được. Một micro không phải được đặt vào miệng của diễn giả và cũng không bị che từ những diễn giải khác. Các micro không quan sát cuộc trò chuyện của các diễn giả một cách tách biệt. Các cuộc hội thoại được ghi lại một cách hỗn hợp. Vấn đề là để không hỗn hợp hoặc ghi lại các cuộc hội thoại gốc từ các cuộc hội thoại hỗn hợp được ghi lại.

3.3.3 Mô hình tách nguồn

Mô hình tách nguồn cho tất cả thời gian i là

$$\underset{(p\times1)}{(x_i|s_i)} = \underset{(p\times1)}{f(s_i)} + \underset{(p\times1)}{\epsilon_i} \tag{3.5}$$

Kết luận

Trong quá trình nghiên cứu, luận văn đã trình bày một số kiến thức cơ bản về thống kê Bayes nhiều chiều và mô hình hồi quy Bayes, đồng thời đưa ra hai ví dụ minh họa cho mô hình hồi quy Bayes.

Ví dụ về xét nghiệm Insulin minh chứng cho ưu điểm của phương pháp Bayes đó là dựa trên những thông tin về dữ liệu đã biết trước về vấn đề quan sát để suy luận cho những thống kê hiện tại.

Mô hình tách nguồn sẽ được giải thích bằng kĩ thuật sử dụng ví dụ "bữa tiệc cocktail". Bữa tiệc cocktail là một ví dụ dễ hiểu, ở đó mô hình tách nguồn có thể được áp dụng.

Tài liệu tham khảo

[1] Aage Volund (1980), *Multivariate bioassay*, Biometrics, 36:225–236

[2] Bernardo, J.M and Smith A.M (1994), *Bayesian Theory*, John Wiley, London

[3] Daniel B.Rowe (2003), *Multivariate Bayes Statistics*, Chapman & Hall/CRC.

[4] Morris, H. D. (1996), *Probability and statistics*, Addison-Wesley, United

[5] Peter M.Lee (2003) *Bayesian Statistics An Introduction*, Oxford University Press Inc., New York.

[6] Peter Congdon (2005), *Bayesian Statistical Modelling*, Wile