Chương 2: SUY DIỄN BAYES

2.1. Các phân phối liên hợp

Định nghĩa

Một họ các phân phối tiên nghiệm π được gọi là liên hợp với mô hình f (x | θ) nếu phân phối hậu nghiệm cũng thuộc họ đó.

Có ba họ liên hợp cơ bản như sau:

2.1.1. Họ Gamma liên hợp với mô hình Poisson

Gọi $(X_1,...,X_n)$ là một mẫu lấy từ phân phối Poisson(θ) với phân phối Gamma(α, λ) là phân phối tiên nghiệm của θ.

Khi đó:

$$f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \sim e^{-n\theta}\theta^{\sum x_i}$$

Phân phối tiên nghiệm Gamma của tham số θ có dạng:

$$\pi(\theta) \sim \theta^{\alpha-1}e^{-\lambda\theta}$$

Đây là một hàm của θ, mật độ tiên nghiệm có dạng tương tự như mô hình $f(x \mid \theta)$. Đây là ý tưởng tổng quát của họ các liên hợp.

Khi đó, với điều kiện X = x, phân phối hậu nghiệm của θ là:

$$\begin{aligned}\pi(\theta \mid x) &\sim f(x \mid \theta)\pi(\theta) \\ &\sim \left(e^{n\theta}\theta^{\sum x_i}\right)\left(\theta^{\alpha-1}e^{-\lambda\theta}\right) \\ &\sim \theta^{\alpha+\sum x_i-1}e^{-(\lambda+n)\theta}\end{aligned}$$

So sánh với dạng tổng quát của mật độ Gamma, ta thấy rằng $\pi(\theta \mid x)$ là phân phối Gamma với tham số mới,

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i; \lambda_x = \lambda + n$$

Nhận xét:

- Họ Gamma của phân phối tiên nghiệm liên hợp với mô hình Poisson

- Sau khi quan sát một mẫu Poisson X = x, phân phối tiên nghiệm Gamma(α,λ) của θ được cập nhật bằng phân phối hậu nghiệm Gamma ($\alpha + \sum x_i$, λ + n).

Ví dụ. Số lần mất mạng mỗi tuần có phân phối Poisson (θ). Tỷ lệ mất mạng hàng tuần không được biết chính xác, nhưng theo kinh nghiệm trước đây với các mạng tương tự cho thấy hàng tuần có trung bình 4 lần mất mạng với độ lệch chuẩn là 2.

Tồn tại một phân phối Gamma với trung bình $\mu = \frac{\alpha}{\lambda} = 4$ và độ lệch chuẩn $\sigma = \frac{\sqrt{\alpha}}{\lambda} = 2$. Do đó các tham số α và λ là nghiệm của hệ phương trình:

$$\begin{cases} \frac{\alpha}{\lambda} = 4 \\ \frac{\sqrt{\alpha}}{\lambda} = 2 \end{cases} \Rightarrow \begin{cases} \alpha = 4 \\ \lambda = 1 \end{cases}$$

Vì vậy, ta giả sử phân phối Gamma(4,1) là phân phối tiên nghiệm của tham số θ. Đây là một tiên nghiệm liên hợp bởi vì phân phối hậu nghiệm cũng thuộc họ Gamma.

Giả sử có $X_1 = 2$ lần mất mạng trong tuần này. Khi đó, phân phối hậu nghiệm của θ là phân phối Gamma với các tham số $\alpha_x = \alpha + 2 = 6; \lambda_x = \lambda + 1 = 2$.

Nếu không có lần nào mất mạng xảy ra trong suốt tuần này, tham số của phân phối hậu nghiệm được cập nhật lại thành: $\alpha_x = \alpha + 2 + 0 = 6; \lambda_x = \lambda + 2 = 3$.

Phân phối hậu nghiệm có tỉ lệ mất mạng trung bình hàng tuần là 6/3 = 2 lần. Như vậy, với hai tuần có số lần mất mạng rất ít đã làm cho ước lượng của tỉ lệ trung bình giảm từ 4 xuống thành 2.

2.1.2. Họ Bêta liên hợp với mô hình nhị thức

Một mẫu từ phân phối nhị thức B(k,θ) (giả sử k đã biết) có:

$$f(x|\theta) = \prod_{i=1}^{n} C_k^{x_i} \theta^{x_i} (1-\theta)^{k-x_i} \sim \theta^{\sum x_i} . (1-\theta)^{nk-\sum x_i}$$

Mật độ của phân phối tiên nghiệm Bê ta(α, β) là một hàm của θ có dạng tương tự như sau:

$$\pi(\theta) \sim \theta^{\alpha-1} (1-\theta)^{\beta-1}; 0 \le \theta \le 1$$

Do đó, mật độ hậu nghiệm của θ là:

$$\pi(\theta|x) \sim f(x|\theta).\pi(\theta) \sim \theta^{\alpha+\sum_{i=1}^{n} x_i - 1} . (1-\theta)^{\beta+nk-\sum_{i=1}^{n} x_i - 1},$$

Và đây là phân phối Beta với tham số mới:

$$\alpha_x = \alpha + \sum_{i=1}^{n} x_i; \beta_x = \beta + nk - \sum_{i=1}^{n} x_i$$

Nhận xét

- Họ các phân phối tiên nghiệm Beta liên hợp với mô hình nhị thức.
- Các tham số hậu nghiệm là: $\alpha_x = \alpha + \sum_{i=1}^{n} x_i; \beta_x = \beta + nk - \sum_{i=1}^{n} x_i$

2.1.3. Họ Chuẩn liên hợp với mô hình chuẩn

Xét một mẫu từ phân phối chuẩn với trung bình μ chưa biết và phương sai σ^2 đã biết:

$$f(x \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \left\{ -\frac{(x_i - \theta)^2}{2\sigma^2} \right\} \sim \exp\left\{ -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} \right\}$$

$$\sim \exp\left\{ \theta \frac{\sum x_i}{\sigma^2} - \theta^2 \frac{n}{2\sigma^2} \right\} = \exp\left\{ \left(\theta \bar{X} - \frac{\theta^2}{2} \right) \frac{n}{\sigma^2} \right\}$$

Nếu phân phối tiên nghiệm của θ cũng là phân phối chuẩn, với trung bình tiên nghiệm là μ và phương sai tiên nghiệm là τ^2, khi đó:

$$\pi(\theta) \sim \exp\left\{ -\frac{(\theta - \mu)^2}{2\tau^2} \right\} \sim \exp\left\{ \left(\theta\mu - \frac{\theta^2}{2} \right) \frac{1}{\tau^2} \right\},$$

Như vậy, phân phối tiên nghiệm của θ cũng có dạng tương tự như $f(x \mid \theta)$.

Phân phối hậu nghiệm của θ là:

$$\pi(\theta \mid x) \sim f(x \mid \theta).\pi(\theta) \sim \exp\left\{ \theta \left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{\tau^2} \right) - \frac{\theta}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \right\}$$

$$\sim \exp\left\{ -\frac{(\theta - \mu_x)^2}{2\tau^2} \right\},$$

Với $\mu_x = \dfrac{\dfrac{n\bar{X}}{\sigma^2} + \dfrac{\mu}{\sigma^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\tau^2}}; \tau_x^2 = \dfrac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\tau^2}}$

Phân phối hậu nghiệm này cũng là phân phối Chuẩn với các tham số μ_x, τ_x.

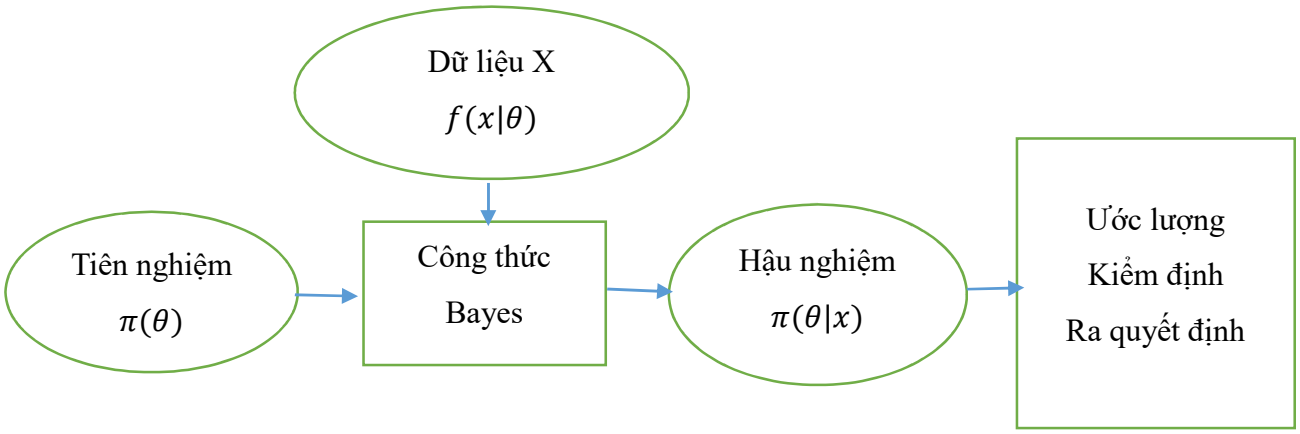
2.2. Các tiên nghiệm không liên hợp (non-conjugate priors)

2.3. Ước lượng Bayes

Để ước lượng θ, ta tính trung bình hậu nghiệm một cách đơn giản như sau:

- θ rời rạc thì: $\hat{\theta}_B = E\{\theta \mid X = x\} = \sum_{\theta} \theta\pi(\theta \mid x) = \dfrac{\sum \theta f(x \mid \theta)\pi(\theta)}{\sum f(x \mid \theta)\pi(\theta)}$
- θ liên tục thì: $\hat{\theta}_B = E\{\theta \mid X = x\} = \int_{\theta} \theta\pi(\theta \mid x)d\theta = \dfrac{\int \theta f(x \mid \theta)\pi(\theta)d\theta}{\int f(x \mid \theta)\pi(\theta)d\theta}$

Kết quả là một kỳ vọng có điều kiện của θ với dữ liệu X cho trước. Nói một cách trừu tượng, ước lượng Bayes $\hat{\theta}_B$ là giá trị ước lượng cho θ sau khi đã quan sát một mẫu.



Trong số tất cả các ước lượng, $\hat{\theta}_B = E\{\theta \mid x\}$ có bình phương sai số rủi ro hậu nghiệm nhỏ nhất

$$\rho(\hat{\theta}) = E\left\{(\hat{\theta} - \theta)^2 \mid X = x\right\}$$

Với ước lượng Bayes $\hat{\theta}_B = E\{\theta \mid x\}$, rủi ro hậu nghiệm bằng với phương sai hậu nghiệm:

$$\rho(\hat{\theta}) = E\left\{(E\{\theta \mid x\} - \theta)^2 \mid x\right\} = E\left\{(\theta - E\{\theta \mid x\}) \mid x\right\} = Var\{\theta \mid x\},$$

Ví dụ. Ước lượng Bayes cho trung bình θ của phân phối chuẩn $N(\theta, \sigma^2)$ với tiên nghiệm $N(\mu, \tau^2)$ là:

$$\hat{\theta}_B = \mu_x = \frac{\frac{n\bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$ và rủi ro hậu nghiệm là: $\rho(\hat{\theta}_B) = \tau_x^2 = \dfrac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$

Ví dụ. Sau dữ liệu hai tuần về tỉ lệ mất mạng trong ví dụ trước, ta đã có phân phối hậu nghiệm với các tham số $\alpha_x = 6$ và $\lambda_x = 3$.

Ước lượng Bayes của tỉ lệ θ là:

$$\hat{\theta}_B = E\{\theta \mid x\} = \frac{\alpha_x}{\lambda_x} = 2 \text{ với rủi ro hậu nghiệm } \rho\left(\hat{\theta}_B\right) = Var\{\theta \mid x\} = \frac{\alpha_x}{\lambda_x^2} = \frac{2}{3}.$$

2.4. Khoảng tin cậy Bayes

Khoảng tin cậy có ý nghĩa hoàn toàn khác nhau trong phân tích Bayes. Với phân phối hậu nghiệm của θ, ta không giải thích độ tin cậy (1 − α) theo hướng dài hạn của các mẫu. Thay vào đó, chúng ta có thể đưa ra một khoảng [a, b] hoặc một tập hợp C có xác suất hậu nghiệm (1 − α) và nói rằng tham số θ thuộc về tập hợp này với xác suất (1 - α). Tập hợp này được gọi là tập tin cậy (1 − α)100%.

Định nghĩa

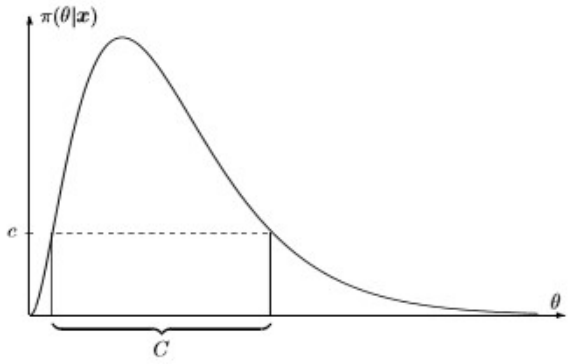
Tập hợp C là tập tin cậy $(1 - \alpha)100\%$ cho tham số θ nếu xác suất hậu nghiệm để θ thuộc tập hợp C bằng đúng $(1 - \alpha)$. Nghĩa là:

$$P\{\theta \in C \mid X = x\} = \int_C \pi(\theta \mid x) d\theta = 1 - \alpha$$

Cực tiểu độ dài của tập C trong tất cả các tập tin cậy $(1 - \alpha)100\%$, ta chỉ cần bao gồm tất cả các giá trị của tham số θ với mật độ hậu nghiệm $\pi(\theta|x)$ cao (HPD), nghĩa là:

$$C = \{\theta : \pi(\theta \mid x) \geq c\}$$

Tập tin cậy mật độ cao hậu nghiệm cao nhất $(1 - \alpha)100\%$ được minh họa như trong hình sau:



2.5. Suy diễn tin cậy cho tỉ lệ p

2.5.1. Sử dụng phân phối tiên nghiệm beta(a, b)

Khi đó phân phối hậu nghiệm của p|y là phân phối $beta(a', b')$. Khoảng tin cậy Bayes 95% với diện tích bằng nhau ở hai đuôi cho p có thể được tìm thấy bằng cách lấy chênh lệch giữa phân vị

thứ 97,5 và phân vị thứ 2,5. Với $a' \geq 10$ và $b' \geq 10$, ta có thể xấp xỉ phân phối hậu nghiệm $beta(a', b')$ bằng phân phối bình chuẩn $N(m', (s')^2)$ với:

$$m' = \frac{a'}{a'+b'} \text{ và } (s')^2 = \frac{a'b'}{(a'+b')^2(a'+b'+1)}$$

Khi đó, tập tin cậy $(1-\alpha)100\%$ của tham số p là: $m' \pm z_{\frac{\alpha}{2}} \times s'$.

Ví dụ. Một nghiên cứu về tỷ lệ sinh viên của một trường đại học có thời gian ngủ ít nhất tám giờ mỗi đêm. Khảo sát một mẫu ngẫu nhiên gồm 27 sinh viên từ trường đại học đó thì có 11 sinh viên cho biết họ ngủ ít nhất tám giờ mỗi đêm. Gọi θ là tham số tổng thể đang được quan tâm. Giả sử phân phối tiên nghiệm của θ là Beta (3.3,7.2). Hãy xây dựng một tập tin cậy HPD 90% cho θ.

Ta có, phân phối hậu nghiệm của $\theta|y = 11$ là:

$$\theta|11 \sim Beta(11+3.3; 27-11+7.2)$$
$$\sim Beta(14.3; 23.2)$$

Với $a' \geq 10$ và $b' \geq 10$, ta có thể xấp xỉ phân phối hậu nghiệm $beta(a', b')$ bằng phân phối bình chuẩn $N(m', (s')^2)$ với:

$$m' = \frac{a'}{a'+b'} = \frac{14.3}{14.3+23.2} = 0.381 \text{ và } (s')^2 = \frac{a'b'}{(a'+b')^2(a'+b'+1)} = \frac{14.3\times23.2}{(14.3+23.2)^2(14.3+23.2+1)} = 0.00613$$

Do đó, tập tin cậy 90% của tham số θ là:

$$m' \pm z_{\frac{\alpha}{2}} \times s' = 0.381 \pm 1.645 \times 0.078 = (0.253; 0.509).$$

Nếu không dùng cách xấp xỉ phân phối Beta bằng phân phối chuẩn, ta có thể tìm tập tin cậy 90% cho θ bằng định nghĩa kết hợp với sự hỗ trợ của phần mềm R như sau:

Theo định nghĩa, chúng ta cần tính các xác suất:

$P(\theta < a|x) = 0.05$ và $P(\theta > b|x) = 0.05$ để tìm hai số a và b.

Như vậy, để tìm a và b, chúng ta cần tính các tích phân sau:

$$\int_0^a Beta(14.3, 23.2)\,d\theta = 0.05 \text{ và } \int_b^1 Beta(14.3; 23.2)\,d\theta = 0.05$$

Với phân phối Beta(14.3,23.2) được cho như sau:

$$f(\theta) = \frac{\Gamma(37.5)}{\Gamma(14.3)\Gamma(23.2)}\theta^{14.3-1}(1-\theta)^{23.2-1}$$

Phần mềm R sẽ hỗ trợ cho việc tính các tích phân này bằng code sau:

Code R:

```
a = 3.3
b = 7.2
n = 27
x = 11
a.star = x+a
b.star = n-x+b
a = qbeta(0.05,a.star,b.star)
b = qbeta(1-0.05,a.star,b.star)
round(cbind(a,b),3)
```

Kết quả thu được như sau:

```
        a     b
[1,] 0.256 0.514
```

2.5.2. Sử dụng phân phối tiên nghiệm Gamma(α, λ)

Ví dụ. Một công ty điện thoại mới dự đoán sẽ xử lý trung bình 1000 cuộc gọi mỗi giờ. Trong 10 giờ hoạt động được chọn ngẫu nhiên, nó đã xử lý 7265 cuộc gọi.

a) Ước lượng ban đầu về tần suất của các cuộc gọi điện thoại?

b) Xây dựng một tập tin cậy HPD 95% , biết rằng số cuộc gọi điện thoại được xử lý mỗi giờ tuân theo quy luật Poisson. Tỷ lệ cuộc gọi hàng giờ có phân phối tiên nghiệm là phân phối mũ.

Gọi θ là tần suất của các cuộc gọi điện thoại, ta cần tìm ước lượng Bayes cho θ.

Ta có, số cuộc gọi trong một giờ có phân phối Poisson $P(\theta)$ với tham số θ chưa biết và phân phối tiên nghiệm là $Exp(\lambda) = Gamma(1, \lambda)$.

Theo đề bài, ta có: $E(\theta) = \frac{1}{\lambda} = 1000 \Rightarrow \lambda = 0.001$.

Và với cỡ mẫu n = 10, ta có: $\sum_{i=1}^{n} X_i = n\overline{X} = 7265$.

Khi đó, phân phối hậu nghiệm là $Gamma(\alpha_x, \lambda_x)$ với;

$$\alpha_x = \alpha + n\overline{X} = 7266$$
$$\lambda_x = \lambda + n = 10.001$$

Phân phối này có trung bình: $\mu_x = \frac{\alpha_x}{\lambda_x} = 726.53$

Và độ lệch chuẩn: $\tau_x = \frac{\alpha_x}{\lambda_x^2} = 72.65$

a) Suy ra ước lượng Bayes của θ là: $E(\theta \mid X) = \mu_x = 726.53$ cuộc gọi mỗi giờ.

b) Vì α_x đủ lớn nên phân phối hậu nghiệm Gamma xấp xỉ với phân phối chuẩn $N(\mu_x, \tau_x^2)$.

Do đó, tập tin cậy HPD 95% được cho θ như sau:

$$\mu_x \pm z_{\frac{0.05}{2}} \times \tau_x = 726.53 \pm 1.96 \times 72.53 = [584.14; 868.92]$$