Một số xác suất tiên nghiệm

1. Xác suất tiên nghiệm trong trường hợp tổng quát:

- Như bài trước ta đã biết:

❖Ý tưởng tiếp cận của trường phái Bayes.

❖Ví dụ về bài toán tung đồng xu (phân phối nhị thức).

❖Sử dụng phương pháp tiếp cận của Bayes, ta đã thấy phân phối xác suất tiên nghiệm trong bài toán tung đồng xu nói riêng hay trong những bài toán liên quan đến phân phối nhị thức nói chung.

- Ta sẽ tiếp tục tìm hiểu về phương pháp tiếp cận Bayes trong các bài toán có phân phối xác suất khác.
- Các trường hợp cần chú ý:

❖Phân phối chuẩn.

❖Phân phối Poisson.

❖Xác suất tiên nghiệm có phân phối không xác định.

Bài toán tổng quát:

- Nhắc lại phương pháp tiếp cận xác suất theo Bayes:

❖Cho θ là một tham số của mô hình, theo ý tưởng của phương pháp Bayes, ta coi θ như một biến ngẫu nhiên.

❖Khi đó ta có công thức:

$$\mathrm{P}(\theta|\mathrm{x}) \sim \mathrm{P}(\mathrm{x}|\theta).\mathrm{P}(\theta)$$

❖Trong đó: $\mathrm{P}(\theta)$ là xác suất tiên nghiệm, $\mathrm{P}(\theta|\mathrm{x})$ là xác suất hậu nghiệm.

- Như vậy $P(\theta)$ là một xác suất và sự thay đổi của theta sẽ dẫn đến sự thay đổi của $P(\theta)$.
- Tham số theta sẽ có phân phối xác suất riêng, bên cạnh phân phối xác suất $\mathrm{P}(\theta|\mathrm{D})$.
- Ta sẽ phải ước lượng tham số cho θ.

2. Suy luận Bayes với các phân phối chuẩn:

- Xét phân phối chuẩn:

$$\mathrm{X} \sim \mathrm{N}(\phi, \mu^2)$$

- Ta có $\theta = (\phi, \mu)$ chính là cặp tham số của bài toán phân phối chuẩn.
- Ta sẽ phải tìm hiểu và ước lượng cặp tham số này.
- Tuy nhiên hiện tại ta chưa nói về bước ước lượng.

- Để bài toán dễ tiếp cận hơn, ta giả sử phương sai cho phân phối của một biến ngẫu nhiên X cũng đã được tìm ra (cố định μ).
- Như vậy trong trường hợp này ta hoàn toàn có thể viết lại thành $\theta = \phi$.
- Tham số trong bài toán sẽ là θ trong $\mathrm{X} \sim \mathrm{N}(\theta, \mu^2)$.

- Theo phương pháp tiếp cận của trường phái Bayes (suy luận Bayes), ta có một "niềm tin" rằng θ sẽ có phân phối chuẩn.
- Khi đó, ta được:

$$\theta \sim \mathrm{N}(\theta_0, \mu_0^2)$$

- Trong đó θ_0 và μ_0 đều đã được biết và cố định.

- Theo phân phối dành cho $\mathrm{X} \sim \mathrm{N}(\theta, \mu^2)$ và $\theta \sim \mathrm{N}(\theta_0, \mu_0^2)$, ta có các công thức sau:

$$\mathrm{p}(\mathrm{x}|\theta) = \frac{1}{\mu\sqrt{2\pi}} \mathrm{e}^{-\frac{(\mathrm{x}-\theta)^2}{2\mu^2}}$$

$$\mathrm{p}(\theta) = \frac{1}{\mu_0\sqrt{2\pi}} \mathrm{e}^{-\frac{(\theta-\theta_0)^2}{2\mu_0^2}}$$

- Theo công thức tổng quát về suy luận Bayes, ta có:

$$\mathrm{P}(\theta|\mathrm{x}) \sim \mathrm{P}(\mathrm{x}|\theta).\mathrm{P}(\theta)$$

$$\Leftrightarrow \mathrm{P}(\theta|\mathrm{x}) \sim \frac{1}{\mu\sqrt{2\pi}}\mathrm{e}^{-\frac{(\mathrm{x}-\theta)^2}{2\mu^2}}.\frac{1}{\mu_0\sqrt{2\pi}}\mathrm{e}^{-\frac{(\theta-\theta_0)^2}{2\mu_0^2}}$$

$$\Leftrightarrow \mathrm{P}(\theta|\mathrm{x}) \sim \mathrm{e}^{-\frac{1}{2}\theta^2(\mu^{-1}+\mu_0^{-1})+\theta\left(\frac{\theta_0}{\mu_0}+\frac{\mathrm{x}}{\mu}\right)}$$

- Như vậy, ta có thể viết lại thành:

$$\mu_1 = \frac{1}{\mu^{-1} + \mu_0^{-1}}$$

$$\frac{\theta_0}{\mu_0} + \frac{\mathrm{x}}{\mu} = \frac{\theta_1}{\mu_1}$$

- Từ đẳng thức ở trên, ta được:

$$p(\theta|\mathrm{x}) \sim \mathrm{e}^{-\frac{\frac{1}{2}\theta^2}{\mu_1}+\theta\frac{\theta_1}{\mu_1}}$$

- Vì ta có $\frac{\theta_1^2}{\mu_1}$ được coi như là một hằng số vì trong trường hợp này ta cố định x và cho θ chạy, nên ta hoàn toàn có thể nhân thêm $\mathrm{e}^{-\frac{\frac{1}{2}\theta_1^2}{\mu_1}}$

- Như vậy, ta được:

$$p(\theta|\mathrm{x}) \sim \mathrm{e}^{-\frac{\frac{1}{2}\theta^2}{\mu_1}+\theta\frac{\theta_1}{\mu_1}+-\frac{\frac{1}{2}\theta_1^2}{\mu_1}} = \mathrm{e}^{-\frac{\frac{1}{2}(\theta-\theta_1)^2}{\mu_1}}$$

- Vì hiện tại $p(\theta|\mathrm{x})$ đang được xấp xỉ với $\mathrm{e}^{-\frac{\frac{1}{2}(\theta-\theta_1)^2}{\mu_1}}$ nên ta hoàn toàn có thể nhân thêm hằng số $\frac{1}{\mu_1\sqrt{2\pi}}$.

- Cuối cùng, ta sẽ được kết quả:

$$\mathrm{p}(\theta|\mathrm{x}) = \frac{1}{\mu_1\sqrt{2\pi}}\mathrm{e}^{-\frac{\frac{1}{2}(\theta-\theta_1)^2}{\mu_1}}$$

- Như vậy ta có thể thấy: $\theta|\mathrm{x} \sim \mathrm{N}(\theta_1, \mu_1)$

Trường hợp n mẫu:

- Bài toán ở trên ta chỉ xét với việc thử nghiệm trên 1 mẫu.
- Bây giờ ta sẽ xét trường hợp n mẫu độc lập với từng mẫu có phân phối chuẩn.
- Khi đó từ công thức của suy luận Bayes, ta có:

$$\mathrm{P}(\theta|\mathrm{x}) \sim \mathrm{P}(\mathrm{x}|\theta).\,\mathrm{P}(\theta) \sim \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}(\mathrm{x_i}|\theta).\,\mathrm{p}(\theta)$$

Trường hợp n mẫu:

- Bằng một vài tính toán, ta thu được:

$$\mathrm{P}(\theta|\mathrm{x}) \sim \mathrm{e}^{-\frac{1}{2}\theta^2\left(\frac{1}{\mu_0}+\frac{\mathrm{n}}{\mu}\right)+\theta\left(\frac{\theta_0}{\mu_0}+\frac{\sum \mathrm{x_i}}{\mu}\right)}$$

- Như bài toán 1 phép thử, ta đặt

$$\theta|\mathrm{x} \sim \mathrm{N}(\theta_1, \mu_1)$$

- Như vậy ta sẽ có các đại lượng:

$$\mu_1 = \frac{1}{\mu^{-1} + \mathrm{n}.\mu_0^{-1}}$$

$$\frac{\theta_0}{\mu_0} + \frac{\sum \mathrm{x}}{\mu} = \frac{\theta_1}{\mu_1}$$

- Nếu ta lấy trung bình $\overline{\mathrm{x}} = \frac{\sum \mathrm{x_i}}{\mathrm{n}}$, ta sẽ được:

$$\theta_1 = \mu_1 \left(\frac{\theta_0}{\mu_0} + \frac{\overline{\mathrm{x}}}{\frac{\mu}{n}} \right)$$

- Như vậy trong trường hợp n mẫu, thay vì ta sử dụng phân phối x như bài toán 1 phép thử, ta sẽ sử dụng:

$$\overline{\mathrm{x}} \sim \mathrm{N}\left(\theta, \frac{\mu}{\mathrm{n}}\right)$$

- Như vậy trong trường hợp bài toán ta gặp là bài toán thử nghiệm với n mẫu, ta sẽ sử dụng phân phối trung bình mẫu:

$$\overline{\mathrm{x}} \sim \mathrm{N}\left(\theta, \frac{\mu}{\mathrm{n}}\right)$$

Cũng như là các đại lượng mang tính trung bình (chia n) khác để tiện cho việc tính toán.

- Ta có thể kết luận rằng phân phối xác suất hậu nghiệm trong trường hợp này cũng sẽ có kết quả tương tự như phân phối xác suất tiên nghiệm.
- Lưu ý rằng điều này có được vì ta xấp xĩ các kết quả liên tục bằng việc nhân cho các hằng số. Nếu trong quá trình vận dụng kĩ thuật lại yêu cầu ta phải nhân cho một hàm số có biến thì không thể làm được.

- Như vậy ta đã có công thức cho bài toán suy luận Bayes với phân phối chuẩn.
- Trên thực tế, ta hoàn toàn có thể thay thế phân phối chuẩn của theta bằng những phân phối khác.
- Tuy nhiên, trong thực tế thì các phân phối liên tục thường được cố gắng xấp xĩ thành phân phối chuẩn nên ta có thể giả định xác suất tiên nghiệm có phân phối chuẩn.

Ví dụ:

- Ví dụ kinh điển của Robinson và Whittaker (1940).
- Ông Robinson tham gia việc đo kích cỡ ngực của đàn ông, tính theo đơn vị Inch.
- Ông ta đo đạc dựa trên một bộ dữ liệu gồm kích cỡ ngực của 10000 người đàn ông khác nhau.
- Khi giải quyết vấn đề này, ông Robinson và ông Whittaker đã thiết lập phương sai cố định là 2 và mẫu ngực trung bình của 10000 người là 39,8.

Ví dụ:

- Như vậy bài toán ta đang xét sẽ là một bài toán có 10.000 mẫu (10000 phép thử).
- Nếu gọi $\mathrm{D_i} \sim \mathrm{N}(\theta, \mu)$ với $1 \leq \mathrm{i} \leq 10000$ là phân phối chuẩn về kích cỡ ngực của đàn ông và $\overline{\mathrm{D}} = \frac{\sum \mathrm{D_i}}{10000}$, thì khi đó, ta có:

$$\overline{\mathrm{D}} \sim \mathrm{N}(39{,}8\ ;\ 2)$$

Ví dụ:

- Theo kinh nghiệm về đo đạc kích cỡ thường thấy của những nhóm người đàn ông khác nhau, người giải quyết bài toán cảm thấy rằng :

$$\theta \sim \mathrm{N}(38, 9)$$

- Như vậy để tính được xác suất hậu nghiệm, ta sẽ phải áp dụng kết quả từ phần lý thuyết.

Ví dụ:

- Ta có các kết quả sau:

$$\mu_1 = \frac{1}{\mu^{-1} + \mathrm{n}.\,\mu_0^{-1}} = \frac{1}{3^{-2} + 1000.2^{-2}} = \frac{1}{2500}$$

$$\theta_1 = \mu_1 \left(\frac{\theta_0}{\mu_0} + \frac{\overline{\mathrm{x}}}{\frac{\mu}{n}} \right) = \frac{1}{2500} \left(\frac{38}{9} + \frac{39.8}{\frac{2^2}{10000}} \right) = 39.8$$

Ví dụ:

- Như vậy sau một số tính toán nhất định, ta có:

$$\theta|\overline{\mathrm{D}} \sim \mathrm{N}\left(39.8, \frac{1}{2500}\right) \sim \mathrm{N}\left(39.8, \frac{2^2}{10000}\right)$$

- Để ý rằng $\theta|\overline{\mathrm{D}} = \frac{\sum \theta|\mathrm{D_i}}{10000}$.

- Trên thực tế nếu ta cho phân phối xác suất tiên nghiệm tương đối gần với phân phối xác suất của từng trường hợp riêng biệt độc lập trong n mẫu, ta sẽ đạt được phân phối của xác suất hậu nghiệm với giá trị tham số gần như là bằng với giá trị tham số đã đo đạc được trên data.
- Tuy nhiên không phải lúc nào ta cũng sẽ đi đến những tình huống rất đẹp như vậy.