NGHIÊN CỨU HÀM MÔ-MEN SINH - MỘT CÔNG CỤ HIỆU QUẢ TRONG PHÂN TÍCH XÁC SUẤT VÀ THỐNG KÊ

Lê Bích Phượng[1,*]
[1]*Trường Đại học Mỏ - Địa chất Hà Nội*
[*]*Email: lebichphuong@humg.edu.vn*

TÓM TẮT

Khoa học dữ liệu (Data Science) là một lĩnh vực liên ngành sử dụng các phương pháp, quy trình, thuật toán và hệ thống khoa học để trích xuất kiến thức và thông tin từ dữ liệu. Nó kết hợp nhiều lĩnh vực khác nhau như thống kê, học máy, khai phá dữ liệu, phân tích dữ liệu và tin học, nhằm phân tích và hiểu sâu hơn về dữ liệu. Khoa học dữ liệu được ứng dụng rộng rãi trong nhiều ngành công nghiệp, bao gồm y tế, tài chính, marketing, sản xuất và dịch vụ công cộng. Xác suất và thống kê đóng vai trò nền tảng trong khoa học dữ liệu. Chúng cung cấp các công cụ và phương pháp cần thiết để thu thập, phân tích, giải thích và trình bày dữ liệu một cách hiệu quả. Hàm mô-men sinh là một công cụ mạnh mẽ và linh hoạt trong lí thuyết xác suất và thống kê vì nó không chỉ giúp xác định các mô-men của biến ngẫu nhiên mà còn hỗ trợ trong việc phân tích và xác định phân phối của các biến ngẫu nhiên.

Từ khóa: *hàm mô-men sinh, xác suất thống kê, kì vọng, phương sai, phân phối, độ xiên.*

1. ĐẶT VẤN ĐỀ

Hàm sinh của biến ngẫu nhiên là một công cụ toán học được sử dụng để mô tả và phân tích các tính chất của phân phối xác suất của biến đó. Nói một cách đơn giản, hàm sinh là giá trị kỳ vọng của một phép biến đổi cụ thể áp dụng lên biến ngẫu nhiên. Có nhiều loại hàm sinh khác nhau, như hàm mô-men sinh (Moment Generating Function - MGF), hàm sinh xác suất (Probability Generating Function - PGF), hàm sinh đặc trưng (Characteristic Function) và hàm sinh tích lũy (Cumulant Generating Function). Mỗi loại hàm sinh có một công thức cụ thể và được sử dụng cho các mục đích khác nhau trong lý thuyết xác suất và thống kê [1-2].

Hàm mô-men sinh của một biến ngẫu nhiên X được định nghĩa là: $M_X(t) = E(e^{tX})$. Trong đó, $E(e^{tX})$ là giá trị kỳ vọng của e^{tX} .Hàm này có thể được sử dụng để tìm các mô-men (như trung bình và phương sai) của biến ngẫu nhiên, và cũng có thể giúp xác định phân phối xác suất của biến đó trong những điều kiện nhất định. Một biến ngẫu nhiên có một phân phối xác suất nhất định nếu hàm sinh của nó xác định. Có một quá trình khôi phục phân phối từ một hàm sinh, và quá trình này được gọi là phép đảo ngược. Tính chất quan trọng là các mô-men của biến ngẫu nhiên có thể được xác định từ các đạo hàm của hàm sinh. Tính chất này vô cùng hữu ích vì việc thu được các mô-men từ hàm sinh thường dễ dàng hơn so với việc tính trực tiếp các mô-men từ định nghĩa của chúng.

Một thuộc tính quan trọng khác là hàm sinh của tổng các biến ngẫu nhiên độc lập là tích của các hàm sinh tương ứng. Thuộc tính này rất hữu ích vì hàm mật độ xác suất của tổng các biến độc lập là tích chập của các hàm mật độ riêng lẻ, và phép toán này phức tạp hơn nhiều. Thuộc tính quan trọng cuối cùng được gọi là định lý liên tục, khẳng định rằng sự hội tụ của dãy các hàm sinh tương ứng với sự hội tụ của các phân phối tương ứng. Thường thì việc chứng minh sự hội tụ của các hàm sinh dễ dàng hơn so với chứng minh sự hội tụ của các phân phối trực tiếp [2-5].

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Phân phối rời rạc và liên tục

Một biến ngẫu nhiên là một hàm số X có thể nhận giá trị một cách ngẫu nhiên và phụ thuộc vào một sự kiện ngẫu nhiên nào đó. Không gian

hoặc miền giá trị của X là tập hợp S các giá trị có thể có của X. Một biến ngẫu nhiên X được gọi là rời rạc nếu tập hợp này có một số lượng hữu hạn hoặc vô hạn đếm được các giá trị khác biệt (tức là có thể liệt kê thành một dãy). Biến ngẫu nhiên X được gọi là có phân phối liên tục nếu nó có thể nhận giá trị bất kì trong một khoảng hoặc một đoạn là một tập con của tập hợp số thực [1, 7].

Thông thường thì có các hàm số gán xác suất cho tất cả các sự kiện trong một không gian mẫu. Những hàm số này được gọi là hàm khối xác suất (probability mass functions) nếu biến ngẫu nhiên có phân phối rời rạc, hoặc hàm mật độ xác suất (probability density functions) nếu biến ngẫu nhiên có phân phối liên tục. Tất cả các giá trị có thể có của một biến ngẫu nhiên và các giá trị xác suất tương ứng của chúng tạo thành phân phối xác suất của biến ngẫu nhiên đó.

Phân phối của một biến ngẫu nhiên X có thể được mô tả bằng hàm phân phối tích lũy:

$$F_X(x) = P(X < x) \quad (1)$$

Cũng có những cách khác để đặc trưng hóa các phân phối xác suất. Do đó, các phân phối xác suất cũng có thể được xác định bằng nhiều phép biến đổi khác nhau, tức là bằng các hàm số nào đó mà mã hóa các thuộc tính của phân phối thành một dạng thuận tiện hơn cho các loại tính toán xác suất nhất định. Đối với một biến ngẫu nhiên rời rạc X, với hàm khối xác suất

$$p(x) = P(X = x) \quad (2)$$

ta có $0 \le p(x) \le 1 \; \forall x$ và $\sum_x p(x) = 1$.

Hàm khối xác suất hoặc hàm mật độ xác suất của một biến ngẫu nhiên X chứa tất cả thông tin mà ta cần về biến này.

2.2. Dãy các mô-men của một biến ngẫu nhiên

Ta biết rằng trung bình $\mu = EX$ và phương sai $\sigma^2 = E((X-EX)^2) = E(X^2) - (EX)^2$ của một biến ngẫu nhiên đóng vai trò quan trọng trong các định lí cơ bản của xác suất, cũng như trong nhiều loại tính toán thực tế khác nhau. Những thuộc tính quan trọng này của một biến ngẫu nhiên chứa đựng những thông tin về hàm phân phối của biến đó. Tuy nhiên, trung bình và phương sai không chứa đựng tất cả thông tin về hàm mật độ của một biến ngẫu nhiên [2].

Ngoài hai đại lượng μ và σ, định vị trung tâm và mô tả độ phân tán của các giá trị của một biến ngẫu nhiên, chúng ta còn định nghĩa một tập hợp các đại lượng khác, gọi là các mô-men, những đại lượng này xác định duy nhất phân phối xác suất của một biến ngẫu nhiên. Đối với một biến ngẫu nhiên rời rạc hoặc liên tục X, mô-men bậc k của X là một số được định nghĩa là $\mu_k = E(X^k)$ với k=1, 2, 3,... với điều kiện các giá trị là tính được. Ta có một dãy các mô-men gắn liền với một biến ngẫu nhiên X. Trong nhiều trường hợp, dãy này xác định phân phối xác suất của X. Tuy nhiên, các mô-men của X có thể không tồn tại. Dựa trên các mô-men này, trung bình và phương sai của X được tính đơn giản bằng $\mu_1 = EX$ và

$$\sigma^2 = E((X-EX)^2) = E(X^2) - (EX)^2 = \mu_2 - (\mu_1)^2 \quad (3)$$

Khi bậc k tăng lên, thì các mô-men bậc cao hơn có ý nghĩa và trở nên phức tạp hơn. Các mô-men cung cấp nhiều thông tin hữu ích về phân phối của X. Kiến thức về hai mô-men đầu tiên của X cho chúng ta biết trung bình và phương sai của nó, nhưng kiến thức về tất cả các mô-men của X xác định hoàn toàn hàm phân phối xác suất của nó. Các phân phối khác nhau không thể có các mô-men giống hệt nhau. Đây chính là điểm then chốt, là lý do tại sao các mô-men lại quan trọng [7].

2.3. Hàm sinh

Nói một cách đơn giản, hàm sinh chuyển đổi các bài toán về chuỗi số thành các bài toán về hàm số. Bằng cách này, chúng ta có thể sử dụng hàm sinh để giải quyết các bài toán đếm số lượng khác nhau. Giả sử rằng $a_0, a_1, a_2 \ldots$ là một dãy số thực hữu hạn hoặc vô hạn. Hàm sinh thông thường của dãy này là chuỗi lũy thừa:

$$G(z) = a_0 + a_1 z + a_2 z^2 + \ldots = \sum_{k=0}^{\infty} a_k z^k \quad (4)$$

Để khôi phục lại dãy ban đầu từ một hàm sinh thông thường đã cho, công thức sau được sử dụng:

$$a_k = \frac{1}{k!}\left[\frac{d^k G(z)}{dz^k}\right]_{z=0}, k = 0,1,2... \quad (5)$$

Giả sử rằng $a_0, a_1, a_2...$ là một dãy số thực hữu hạn hoặc vô hạn. Hàm sinh lũy thừa của dãy này là chuỗi lũy thừa:

$$G(z) = a_0 + \frac{a_1 z}{1!} + \frac{a_2 z^2}{2!} + ... = \sum_{k=0}^{\infty} \frac{a_k z^k}{k!} \quad (6)$$

Để khôi phục lại chuỗi số thực ban đầu từ hàm sinh lũy thừa đã cho, công thức sau được sử dụng:

$$a_k = \left.\frac{d^k G(z)}{dz^k}\right|_{z=0}, k = 0,1,2... \quad (7)$$

Đối với một biến ngẫu nhiên X chỉ nhận các giá trị nguyên không âm k, với xác suất $p_k = P(X = k)$, hàm sinh xác suất được định nghĩa là:

$$G(z) = E(z^X) = \sum_{k=0}^{\infty} p_k z^k, 0 \le z \le 1. \quad (8)$$

Bởi công thức:

$$E(X^k) = \left[\frac{d^k G(z)}{dz^k}\right]_{z=1}, k = 0,1,2,... \quad (9)$$

ta khôi phục các mô-men của X. Một hàm sinh xác suất chính xác sẽ xác định duy nhất một phân phối, và một hàm sinh xác suất xấp xỉ sẽ xác định xấp xỉ một phân phối xác suất.

2.4. Hàm mô-men sinh

Hàm mô-men sinh mang lại nhiều kết quả một cách dễ dàng. Các chứng minh sử dụng hàm mô-men sinh thường dễ dàng hơn nhiều so với việc chứng minh (cùng một kết quả) bằng cách sử dụng các hàm mật độ xác suất (hoặc các phương pháp khác). Hàm mô-men sinh (MGF) được định nghĩa bởi công thức sau:

$$M_X(t) = E(e^{tX}) \quad (10)$$

trong công thức trên, kỳ vọng tồn tại xung quanh một lân cận của 0.

Khi X là biến ngẫu nhiên rời rạc thì mô-men sinh là:

$$M_X(t) = \sum e^{tx} p(x) \quad (11)$$

Khi X là biến ngẫu nhiên liên tục thì mô-men sinh là:

$$M_X(t) = \int e^{tx} f(x)\,dx \quad (12)$$

Ở đây, điều quan trọng là kì vọng phải hữu hạn đối với mọi giá trị t trong một khoảng nào đó của t_0 (với t_0> nào đó). Nếu kỳ vọng không tồn tại trong một lân cận nào đó thì hàm mô-men sinh không tồn tại. Vì hàm mũ luôn dương, $E(e^{tX})$ luôn tồn tại (bằng một số thực hoặc bằng dương vô cùng) [1-2].

Các hàm mô-men sinh có thể không được xác định đối với tất cả các giá trị của t, và một số phân phối nổi tiếng không có hàm mô-men sinh (ví dụ như phân phối Cauchy). Phân phối Cauchy (hay còn được gọi là phân phối Lorentz trong vật lý) có hàm mật độ xác suất như sau:

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]} \quad (13)$$

trong đó x_0 là thông số vị trí (median hoặc mode) mô tả vị trí trung tâm của phân phối; $\gamma > 0$ là thông số thang đo (scale parameter), mô tả độ rộng của phân phối; x là biến. Tích phân $M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x; x_0, \gamma)dx$ không hội tụ nên không tồn tại hàm mô-men sinh.

Hàm mô-men sinh là một hàm biến t, không phải của X. Hàm mô-men sinh của một biến ngẫu nhiên gói gọn tất cả các mô-men của biến ngẫu nhiên đó vào một biểu thức đơn giản. Về mặt hình thức, hàm mô-men sinh được tạo ra bằng cách thay e^t vào hàm sinh xác suất

3. KẾT QUẢ VÀ THẢO LUẬN

Giả sử rằng hàm mô-men sinh tồn tại trong một lân cận của gốc tọa độ. Ta có một số kết quả sau:

3.1. Tính chất 1

Nếu $g_X(t)$ là hàm mô-men sinh của một biến ngẫu nhiên X, thì: $g_X(0)=1$.

Chứng minh:

Thật vậy ta có, $g_X(0)=1=E(e^{0.X})=E(1)=1$.

3.2. Tính chất 2

Các mô-men của biến ngẫu nhiên X có thể được tìm bằng cách khai triển chuỗi lũy thừa. Hàm mô-men sinh của một biến ngẫu nhiên X là hàm sinh lũy thừa của chuỗi mô-men của nó:

$$g_X(t)=\sum_{k=0}^{\infty}\frac{\mu_k t^k}{k!} \quad (14)$$

Hàm mũ có khai triển lũy thừa:

$$e^t=\sum_{k=0}^{\infty}\frac{t^k}{k!}, \quad (15)$$

nên nên bằng cách khai triển chuỗi của hàm e^{tX}, ta có:

$$e^{tX}=\sum_{k=0}^{\infty}\frac{(tX)^k}{k!}. \quad (16)$$

Lấy kì vọng hai vế ta có:

$$E\left(e^{tX}\right)=E\left(\sum_{k=0}^{\infty}\frac{(tX)^k}{k!}\right)=\sum_{k=0}^{\infty}E\left(\frac{X^k t^k}{k!}\right)=\sum_{k=0}^{\infty}\frac{t^k}{k!}E\left(X^k\right) \quad (17)$$

3.3. Tính toán các mô-men

Ta gọi hàm $g_X(t)$ là hàm mô-men sinh của biến ngẫu nhiên X do tất cả các mô-men của X có thể thu được bằng cách lấy đạo hàm rồi thay t=0. Đạo hàm bậc k của $g_X(t)$ tại điểm t=0 là mô-men bậc k (μ_k) của X, cụ thể là:

$$\mu_k=g^{(k)}(0) \quad (18)$$

trong đó

$$g^{(k)}(t)=\left.\frac{d^k g(t)}{dt^k}\right|_{t=0}. \quad (19)$$

Bằng cách này, các mô-men của X cũng có thể được tìm thấy thông qua việc lấy đạo hàm.

$$\frac{d^k}{dt^k}g_X(t)=\frac{d^k}{dt^k}E(e^{tX})=E(\frac{d^k}{dt^k}e^{tX})=E(X^k e^{tX}) \quad (20)$$

Do vậy ta có:

$$\left.\frac{d^k}{dt^k}g_X(t)\right|_{t=0}=\mu_k. \quad (21)$$

Như vậy, hàm mô-men sinh sinh ra tất cả các mô-men của X thông qua việc lấy đạo hàm. Ta có thể tìm các mô-men của X bằng cách tính hàm mô-men sinh và sau đó lấy đạo hàm. Đôi khi, việc làm này dễ dàng thực hiện hơn so với cách tính trực tiếp. Tất cả các mô-men của một phân phối hầu như xác định phân phối đó. Ngoài việc tạo ra các mô-men của X, hàm mô-men sinh còn hữu ích trong việc xác định phân phối của X.

3.4. Tính xác định

Nếu $g_X(t)$ tồn tại trong một lân cận của t=0, thì việc biết hàm mô-men sinh của một biến ngẫu nhiên tương đương với việc biết hàm mật độ xác suất của nó. Điều này có nghĩa là ***hàm mô-men sinh xác định duy nhất hàm mật độ xác suất.*** Trong trường hợp tổng quát, chuỗi $g_X(t)$ sẽ không hội tụ với mọi t. Nhưng trong trường hợp đặc biệt quan trọng khi X bị chặn (tức là khi miền giá trị của X nằm trong một khoảng hữu hạn), ta có thể chứng minh rằng chuỗi này hội tụ với mọi giá trị của t. Tức là hàm phân phối được xác định hoàn toàn bởi các mô-men của nó.

Định lý 1: Giả sử X à một biến ngẫu nhiên liên tục với phạm vi nằm trong khoảng thực $[-M,M]$. Khi đó, chuỗi mô-men $g_X(t)=\sum_{k=0}^{\infty}\frac{\mu_k t^k}{k!}$ hội tụ với mọi giá trị của t thành một hàm khả vi vô hạn $g_X(t)$ và $g_X^{(k)}(0)=\mu_k$.

Chứng minh. Ta biết rằng $\mu_k=\int_{-M}^{M}x^k f_X(x)dx$. Do vậy, với mọi n ta có:

$$\sum_{k=0}^{n}\left|\frac{\mu_k t^k}{k!}\right|\le\sum_{k=0}^{n}\frac{\left(M|t|\right)^k}{k!}\le e^{M|t|}. \quad (22)$$

Bất đẳng thức này cho thấy chuỗi mô-men hội tụ với mọi giá trị của t và tổng của nó là một hàm khả vi vô hạn. Bằng cách này, chúng ta đã chứng minh rằng chuỗi mô-men μ_k xác định hàm $g_X(t)$. Ngược lại, $\mu_k=g_X^{(k)}(0)$, ta thấy $g_X(t)$ xác định các mô-men μ_k.

Nếu X là một biến ngẫu nhiên bị chặn, thì ta có thể chứng minh rằng hàm mô-men sinh $g_X(t)$

của X xác định duy nhất hàm mật độ xác suất $f_X(t)$ của X. Điều này quan trọng vì việc tính toán với các hàm mô-men sinh dễ dàng hơn so với việc tính toán với các hàm mật độ xác suất.

3.5. Tính duy nhất

Hai biến ngẫu nhiên có cùng hàm mô-men sinh thì sẽ có cùng phân phối.

Định lí 2: Giả sử X và Y là hai biến ngẫu nhiên với các hàm mô-men sinh tương ứng là $g_X(t)$ và $g_Y(t)$ và các hàm phân phối xác suất lần lượt là $F_X(x)$ và $F_Y(y)$. Nếu $g_X(t) = g_Y(t)$, thì $F_X(x) = F_Y(y)$.

Điều này đảm bảo rằng phân phối của một biến ngẫu nhiên có thể được xác định bởi hàm mô-men sinh của nó. Hệ quả của định lý trên là nếu tất cả các mô-men của một biến ngẫu nhiên X tồn tại, chúng sẽ hoàn toàn xác định hàm mô-men sinh (vì các mô-men là các đạo hàm của hàm mô-men sinh trong khai triển Taylor của nó) và các mô-men này cũng hoàn toàn xác định phân phối, cũng như hàm phân phối tích lũy, hàm mật độ xác suất và hàm khối xác suất.

Khi hàm mô-men sinh tồn tại, sẽ có một phân phối duy nhất tương ứng với hàm mô-men sinh đó. Do đó, có một đơn ánh giữa các hàm mô-men sinh và các phân phối xác suất. Điều này cho phép ta sử dụng các hàm mô-men sinh để tìm các phân phối của các biến ngẫu nhiên biến đổi trong một số trường hợp. Kỹ thuật này thường được sử dụng cho các tổ hợp tuyến tính của các biến ngẫu nhiên độc lập.

3.6. Tính xác định vô hạn mô-men

Khi hàm mô-men sinh tồn tại, nó xác định một tập hợp vô hạn các mô-men. Câu hỏi hiển nhiên đặt ra là liệu hai phân phối khác nhau có thể có cùng một tập hợp mô-men vô hạn hay không. Câu trả lời là, khi hàm mô-men sinh tồn tại trong một lân cận của 0, dãy mô-men vô hạn sẽ xác định duy nhất phân phối. Điều này cho phép chúng ta xác định phân phối của một dãy các biến ngẫu nhiên bằng cách xem xét các hàm mô-men sinh liên quan.

3.7. Tính toán mô-men của tổng hai biến ngẫu nhiên

Đối với hai biến ngẫu nhiên độc lập X và Y, hàm sinh mô men của tổng $X+Y$ là tích của các hàm mô-men sinh riêng rẽ:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \quad (23)$$

3.8. Hàm mô-men sinh của một số phân phối xác suất phổ biến

Phân phối đều rời rạc

Biến ngẫu nhiên X có phân phối đều rời rạc, X nhận các giá trị trên tập $\{x_1; x_2; ...; x_n\}$ với xác suất mỗi giá trị như nhau $P(X = x_i) = \frac{1}{n}$. Hàm mô-men sinh là:

$$M_X(t) = E[e^{tX}] = \frac{1}{n}\sum_{i=1}^{n} e^{tx_i} \quad (24)$$

Phân phối Nhị thức (Binomial Distribution)

Biến ngẫu nhiên X có phân phối nhị thức với số lần thử n và xác suất thành công p, $X \sim \text{Binomial}(n, p)$. Hàm mô-men sinh là:

$$M_X(t) = \left(1 - p + pe^t\right)^n \quad (25)$$

Phân phối Poisson

Phân phối Poisson được sử dụng để mô tả số sự kiện xảy ra trong một khoảng thời gian với tần suất trung bình λ. Biến ngẫu nhiên X có phân phối Poisson với tham số $\lambda, X \sim \text{Poisson}(\lambda)$. Hàm mô-men sinh là:

$$M_X(t) = \exp\left(\lambda\left(e^t - 1\right)\right) \quad (26)$$

Phân phối siêu bội

Phân phối siêu bội, mô tả số lượng phần tử loại A được chọn từ một tập hợp có N phần tử, trong đó có K phần tử loại A, qua n lần chọn không

hoàn lại. Hàm mô-men sinh của phân phối này là:

$$M_X(t) = E[e^{tX}] = \sum_{k=1}^{K} \frac{C_K^k . C_{N-K}^{n-k}}{C_N^n} e^{tk} \quad (27)$$

Phân phối đều liên tục

Phân phối đều liên tục trên đoạn [a, b]. Hàm mô-men sinh của nó là:

$$M_X(t) = E[e^{tX}] = \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)}, t \neq 0 \\ 1, t = 0 \end{cases} \quad (28)$$

Phân phối mũ

Phân phối mũ với hệ số tỉ lệ λ. Hàm mô-men sinh của nó là:

$$M_X(t) = E[e^{tX}] = \begin{cases} \frac{\lambda}{\lambda - t}, t < \lambda \\ \infty, t \geq \lambda \end{cases} \quad (29)$$

Phân phối Chuẩn (Normal Distribution)

Biến ngẫu nhiên X có phân phối chuẩn với trung bình μ và phương sai σ^2, $X \sim N(\mu, \sigma^2)$. Hàm mô-men sinh của nó là:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \quad (30)$$

Phân phối Chuẩn tắc (Exponential Distribution)

Biến ngẫu nhiên X có phân phối chuẩn tắc với tham số λ, $X \sim \text{Exponential}(\lambda)$.

$$M_X(t) = E[e^{tX}] = \exp\left(\frac{t^2}{2}\right) \quad (31)$$

3.9. Ứng dụng của hàm mô-men sinh

Trong thống kê mô tả các giá trị như: kỳ vọng, phương sai, độ xiên, đọ nhọn là vô cùng quan trọng; nó đặc trưng cho phân phối, bởi vì:

Kỳ vọng (hay trung bình) là giá trị trung bình lý thuyết của một biến ngẫu nhiên. Nó phản ánh trung tâm hoặc xu hướng chung của phân phối xác suất. Kỳ vọng thường được sử dụng để dự đoán giá trị trung bình dài hạn của các biến ngẫu nhiên, như lợi nhuận kỳ vọng trong tài chính.

Phương sai đo lường mức độ phân tán của các giá trị quanh kỳ vọng. Nó cho biết dữ liệu lan rộng hay tập trung. Phương sai thường được sử dụng để đánh giá độ rủi ro hoặc biến động, ví dụ như biến động của giá cổ phiếu.

Độ xiên giúp hiểu rõ hơn về đặc điểm hình dạng của dữ liệu, đặc biệt trong việc phát hiện sự chệch lệch.

Độ nhọn rất hữu ích trong tài chính để đánh giá rủi ro các sự kiện bất thường, như sụt giảm giá trị cổ phiếu.

Hàm mô-men sinh giúp xác định các giá trị này.

Xác định Kỳ vọng

Kỳ vọng của một biến ngẫu nhiên X, ký hiệu là $\mathbb{E}[X]$, có thể được tính toán từ hàm sinh mô men $M_X(t)$ bằng cách lấy đạo hàm thứ nhất của $M_X(t)$ tại $t = 0$:

$$\mathbb{E}[X] = M_X'(0) \quad (32)$$

Xác định Phương sai

Phương sai của một biến ngẫu nhiên X, ký hiệu là $\text{Var}(X)$, có thể được tính toán từ hàm sinh mô men bằng cách sử dụng kỳ vọng và đạo hàm thứ hai của $M_X(t)$ tại $t = 0$:

$$\mathrm{Var}(X) = M_X^{''}(0) - \left(M_X^{'}(0)\right)^2 \quad (33)$$

Xác định độ xiên

Độ xiên của một phân phối là một thước đo cho sự bất đối xứng của phân phối đó. Nó có thể được tính từ hàm sinh mô men bằng cách sử dụng đạo hàm bậc ba của $M_X(t)$:

$$\mathrm{Skewness}(X) = \frac{M_X^{'''}(0) - 3M_X^{'}(0)M_X^{''}(0) + 2\left(M_X^{'}(0)\right)^3}{(\mathrm{Var}(X))^{3/2}} \quad (34)$$

Xác định độ nhọn

Độ nhọn của một phân phối là một thước đo cho sự tập trung của các giá trị xung quanh trung bình. Nó có thể được tính từ hàm sinh mô men bằng cách sử dụng đạo hàm bậc bốn của $M_X(t)$:

$$\mathrm{Kurtosis}(X) = \frac{M_X^{''''}(0) - 4M_X^{'}(0)M_X^{'''}(0) + 6\left(M_X^{''}(0)\right)^2 - 3\left(M_X^{'}(0)\right)^4}{\left(\mathrm{Var}(X)\right)^2} \quad (35)$$

Ví dụ 1: Xác định kỳ vọng và phương sai của biến ngẫu nhiên X có phân phối chuẩn $N(\mu, \sigma^2)$.

Hàm sinh mô men của X là:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

Đạo hàm thứ nhất tại $t = 0$:

$$M_X^{'}(t) = \left(\mu + \sigma^2 t\right)\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

$$M_X^{'}(0) = \mu.$$

Vậy, kỳ vọng của X là $\mathbb{E}[X] = \mu$.

Đạo hàm thứ hai tại $t = 0$:

$$M_X^{''}(t) = \left(\sigma^2 + \left(\mu + \sigma^2 t\right)^2\right)\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

$$M_X^{''}(0) = \sigma^2 + \mu^2.$$

Vậy phương sai của X là $\mathrm{Var}(X) = \sigma^2$.

Ví dụ 2: Xác định kỳ vọng và phương sai của biến ngẫu nhiên X có phân phối Poisson với tham số λ.

Hàm sinh mô men của X là:

$$M_X(t) = \exp\left(\lambda\left(e^t - 1\right)\right)$$

Đạo hàm thứ nhất tại $t = 0$:

$$M_X^{'}(t) = \lambda e^t \exp\left(\lambda\left(e^t - 1\right)\right)$$

$M_X^{'}(0) = \lambda$. Vậy, kỳ vọng của X là

$$\mathbb{E}[X] = \lambda.$$

Đạo hàm thứ hai tại $t = 0$:

$$M_X^{''}(t) = \lambda e^t\left(\lambda e^t + 1\right)\exp\left(\lambda\left(e^t - 1\right)\right)$$

$$M_X^{''}(0) = \lambda(\lambda + 1).$$

Vậy, phương sai của X là $\mathrm{Var}(X) = \lambda$.

3.10. Một số vấn đề thực tế của hàm mô-men sinh

Trong khoa học dữ liệu và các dự án phân tích dữ liệu, việc sử dụng hàm mô-men sinh có thể giúp cải thiện kết quả và độ chính xác của các mô hình, đặc biệt là khi làm việc với các phân phối không tiêu chuẩn hoặc khi cần xác định đặc điểm của các phân phối.

Phân phối các giá trị bất thường trong dữ liệu giao thông: ***Bối cảnh:*** Trong một dự án khoa học dữ liệu liên quan đến dự đoán lưu lượng giao thông, nhóm phân tích phát hiện ra rằng dữ liệu có chứa nhiều giá trị ngoại lệ (outliers), ví dụ như tắc nghẽn giao thông bất thường do tai nạn hoặc thời tiết. ***Ứng dụng hàm mô-men sinh:*** Để mô tả chính xác hơn các ngoại lệ này, hàm sinh mô-men có thể được sử dụng để mô hình hóa các phân phối với đuôi dài (long-tail distributions, tức là xác suất xảy ra các giá trị cực đoan (rất lớn hoặc rất nhỏ) không như trong phân phối chuẩn) như phân phối Cauchy hoặc phân phối Pareto. Bằng cách sử dụng MGF để phân tích đặc tính các ngoại lệ, nhóm có thể tạo ra mô hình dự đoán chính xác hơn về lưu lượng giao thông trong các tình huống bất thường. ***Kết quả:*** So với các phương pháp truyền thống, việc sử dụng hàm mô-men sinh giúp nhận diện chính xác hơn các sự kiện hiếm, cải thiện khả năng dự đoán lưu lượng trong các trường hợp đặc biệt.

Mô hình hóa sự biến động của giá chứng khoán. ***Bối cảnh:*** Một nhóm nghiên cứu phân tích thị trường chứng khoán để dự đoán sự biến động giá cả (volatility). Dữ liệu thị trường thường không tuân theo phân phối chuẩn và có tính chất phức tạp như kurtosis cao (độ nhọn của phân phối). ***Ứng dụng hàm mô-men sinh:*** Trong trường hợp này, nhóm có thể sử dụng MGF để tính toán các mô-men của các phân phối khác nhau và phân tích các phân phối với kurtosis cao hơn so với phân phối chuẩn. Điều này giúp họ hiểu rõ hơn sự thay đổi lớn trong giá chứng khoán, từ đó cải thiện mô hình dự đoán giá. ***Kết quả:*** Việc sử dụng MGF để mô tả tốt hơn các đặc tính phân phối giúp cải thiện độ chính xác trong việc dự đoán biến động giá chứng khoán.

Phân tích hành vi người tiêu dùng. ***Bối cảnh:*** Một công ty bán lẻ lớn muốn phân tích hành vi tiêu dùng để tối ưu hóa các chiến lược tiếp thị cá nhân hóa. Dữ liệu người tiêu dùng có nhiều đặc điểm khác nhau như số lần mua hàng, giá trị đơn hàng trung bình, và sự thay đổi trong hành vi mua sắm theo thời gian. ***Ứng dụng hàm mô-men sinh:*** Sử dụng MGF, công ty có thể mô hình hóa sự biến động trong hành vi người tiêu dùng và tạo ra các mô hình phân phối mô tả tốt hơn các sự kiện hiếm hoặc bất thường, chẳng hạn như sự tăng vọt đột ngột trong chi tiêu. Điều này cho phép công ty phát triển các chiến lược tiếp thị hiệu quả hơn và dự đoán chính xác hơn sự thay đổi trong hành vi mua sắm. ***Kết quả:*** Kết hợp hàm sinh mô-men giúp công ty cải thiện chiến lược dự đoán hành vi mua hàng, từ đó nâng cao hiệu quả tiếp thị và giữ chân khách hàng.

4. KẾT LUẬN VÀ KIẾN NGHỊ

4.1. Kết luận

Hàm mô-men sinh đóng vai trò quan trọng trong khoa học dữ liệu vì nó phân tích và mô tả các đặc tính phân phối của dữ liệu một cách toàn diện. Hàm mô-men sinh giúp tính toán các mô-men (trung bình, phương sai, độ xiên, độ nhọn) và cung cấp thông tin chi tiết về các phân phối không chuẩn hoặc dữ liệu có tính chất

phức tạp, có nhiều biến động. Trong các dự án khoa học dữ liệu, hàm mô-men sinh giúp: Mô hình hóa chính xác hơn các phân phối phức tạp. Cải thiện khả năng dự đoán khi làm việc với các dữ liệu có nhiều biến động hoặc ngoại lệ.

4.2. Kiến nghị

Cần nắm vững lý thuyết về hàm sinh mô-men: Việc hiểu cách các hàm mô-men sinh hoạt động, cách tính toán mô-men và vai trò của nó trong việc mô tả các phân phối khác nhau là rất hữu ích khi xử lý dữ liệu không tuân theo phân phối chuẩn.

Ứng dụng hàm mô-men sinh vào các bài toán thực tế: Các dự án liên quan đến dự đoán, phân tích rủi ro hoặc nhận diện ngoại lệ đều có thể làm tốt hơn từ việc sử dụng hàm sinh mô-men. Các mô hình liên quan đến thị trường tài chính, hành vi người tiêu dùng, hoặc phân tích y tế đều là những lĩnh vực quan trọng để áp dụng chúng.

Sử dụng MGF để so sánh và phân tích dữ liệu: Khi làm việc với nhiều tập dữ liệu khác nhau, MGF có thể được sử dụng để phân tích sự khác biệt giữa các phân phối hoặc để so sánh tính chất của các tập dữ liệu.

5. LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Mỏ-Địa chất, trong đề tài mã số T25-20.

TÀI LIỆU THAM KHẢO

1. Hogg, R.V.,Tanis, E.A (2009). Probability and statistical inference. Pearson Education.
2. Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2008). Mathematical statistics with applications (7th ed.). Cengage Learning.
3. Johnson, R. A., & Wichern, D. W. (2018). Applied multivariate statistical analysis (6th ed.). Pearson.
4. Panik, M. J. (2012). Statistical inference: A short course. John Wiley & Sons.
5. Cox, D. R., & Hinkley, D. V. (1974). Theoretical statistics. Chapman and Hall.
6. Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
7. Nguyễn Thị Hằng Lê Bích Phượng (chủ biên), Phạm Ngọc Anh, Nguyễn Thế Lâm, Nguyễn Thu Hằng (2024). Giáo trình lý thuyết xác suất và thống kê toán học. NXB Giao thông vận tải.

Thông tin của tác giả:

TS. Lê Bích Phượng
Bộ môn Toán, khoa Khoa học Cơ bản, nhóm nghiên cứu BSASD, trường Đại học Mỏ - Địa chất
Điện thoại: +(84) 988782112 - Email: lebichphuong@humg.edu.vn

STUDY ON MOMENT GENERATING FUNCTIONS - AN EFFECTIVE TOOL IN PROBABILITY AND STATISTICAL ANALYSIS

Information about authors:

Le Bich Phuong, Ph.D., Department of Mathematics, Faculty of Basic Sciences, BSASD research group, Hanoi University of Mining and Geology
Email: lebichphuong@humg.edu.vn

ABSTRACT:

Data Science is an interdisciplinary field that employs scientific methods, processes, algorithms, and systems to extract knowledge and insights from data. It integrates various domains such as statistics, machine learning, data mining, data analysis, and informatics to analyze and gain deeper understanding of data. Data Science is widely applied across numerous industries, including healthcare, finance, marketing, manufacturing, and public services. Probability and statistics serve as foundational pillars of data science, providing essential tools and methods for collecting, analyzing, interpreting, and effectively presenting data. The moment generating function (MGF) is a powerful and versatile tool in probability and statistics, as it not only helps determine the moments of random variables but also aids in analyzing and identifying the distributions of random variables.

Keywords: *Moment generating function, probability and statistics, expectation, variance, distribution, skewness, kurtosis.*

REFERENCES

1. Hogg, R.V.,Tanis, E.A (2009). Probability and statistical inference. Pearson Education.
2. Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2008). Mathematical statistics with applications (7th ed.). Cengage Learning.
3. Johnson, R. A., & Wichern, D. W. (2018). Applied multivariate statistical analysis (6th ed.). Pearson.
4. Panik, M. J. (2012). Statistical inference: A short course. John Wiley & Sons.
5. Cox, D. R., & Hinkley, D. V. (1974). Theoretical statistics. Chapman and Hall.
6. Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
7. Nguyễn Thị Hằng Lê Bích Phượng (chủ biên), Phạm Ngọc Anh, Nguyễn Thế Lâm, Nguyễn Thu Hằng (2024). Giáo trình lý thuyết xác suất và thống kê toán học. NXB Giao thông vận tải.

Ngày nhận bài: 12/12/2024;
Ngày gửi phản biện: 13/12/2024;
Ngày nhận phản biện: 06/01/2024;
Ngày chấp nhận đăng: 06/01/2024.