Biến ngẫu nhiên và phân phối xác suất

Biến ngẫu nhiên

Biến ngẫu nhiên (*random variables*) là các biến nhận 1 giá trị ngẫu nhiên đại diện cho kết quả của phép thử. Mỗi giá trị nhận được x của biến ngẫu nhiên X được gọi là một thể hiện của X, đây cũng là kết quả của phép thử hay còn được hiểu là một sự kiện.

Gọi tên là một biến có vẻ hơi kì kì một chút bởi biến ngẫu nhiên thực chất là một hàm ánh xạ từ không gian sự kiện đầy đủ tới 1 số thực: $X : \Omega \mapsto \mathbb{R}$

Biến ngẫu nhiên có 2 dạng:

- Rời rạc (*discrete*): tập giá trị nó là rời rạc, tức là đếm được. Ví dụ như mặt chấm của con xúc xắc.
- Liên tục (*continous*): tập giá trị là liên tục tức là lấp đầy 1 khoảng trục số. Ví dụ như giá thuê nhà ở Hà Nội.

Phân phối xác suất

Là phương pháp xác định xác suất của biến ngẫu nhiên được phân phối ra sao. Có 2 cách để xác định phân bố này là dựa vào bảng phân bố xác xuất và hàm phân phối xác suất. Ở đây, tôi chỉ đề cập tới phương pháp hàm phân bố xác suất. Hàm phân phối xác suất của biến ngẫu nhiên X được xác định như sau:

$F_X(x) = P(Xx), xR$

Hàm phân phối xác suất còn có tên là hàm phân phối tích luỹ (*CDF - Cumulative Distribution Function*) do đặc trưng là lấy xác suất của các biến ngẫu nhiên bên trái của một giá trị x bất kì nào đó. Hàm này có đặc điểm là một hàm không giảm, tức là nếu $a < b$ thì $F_X(a) \leq F_X(b)$ vì sự kiện b đã bao gồm cả sự kiện a rồi.

Hàm khối xác suất của biến rời rạc

Với các biến ngẫu nhiên ta còn quan tâm xem xác suất tại mỗi tại 1 giá trị x nào đó trong miền giá trị của nó là bao nhiêu, hàm xác suất như vậy đối với biến ngẫu nhiên rời rạc được gọi là hàm khối xác suất (*PMF - Probability Mass Function*). Giả sử miền xác định của X là D, tức $X : \Omega \mapsto \mathsf{D}$ thì hàm khối xác suất được xác định như sau:

$$p(x) = p_X(x) = \begin{cases} P(X = x) & \text{if } x \in \mathsf{D} \\ 0 & \text{if } x \notin \mathsf{D} \end{cases}$$

Như vậy ta có thể thấy rằng hàm khối xác suất thực chất cũng là một xác suất nên nó mang đầy đủ tất cả các tính chất của xác suất như:

$0 \leq p(x) \leq 1$

$$\sum_{x_i \in \mathsf{D}} p(x_i) = 1$$

Hàm mật độ xác suất của biến liên tục

Với các biến ngẫu nhiên liên tục ta có khái niệm hàm mật độ xác suất (*PDF - Probability Density Function*) để ước lượng độ tập trung xác suất tại lân cận điểm nào đó. Hàm mật độ xác suất $f(x)$ tại điểm x được xác định bằng cách lấy đạo hàm của hàm phân phối tích luỹ $F(x)$ tại điểm đó:

$f(x) = F'(x)$

Như vậy thì nơi nào $f(x)$ càng lớn thì ở đó mức độ tập xác suất càng cao. Từ đây ta cũng có thể biểu diễn hàm phân phối tích luỹ như sau:

$F(x) = \int_{-\infty}^{x} f(t)dt$

Xác suất trong 1 khoảng (α, β) cũng có thể được tính bằng hàm mật độ xác suất:

$P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx$

Hàm mật độ xác suất cũng có 2 tính chất như xác suất như sau:

- Không âm: $f(x) \geq 0 \quad , \forall x \in \mathbb{R}$
- Tổng toàn miền bằng 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

Các đặc trưng

Kỳ vọng

Kỳ vọng (Expectation) của biến ngẫu nhiên là trung bình của biến ngẫu nhiên. Kỳ vọng của biến ngẫu nhiên X được kí hiệu là $E[X]$:

$$E[X] = \begin{cases} \sum_{\forall i} x_i p_i & \text{if x is discrete} \\ \int_{-\infty}^{\infty} x f(x)dx & \text{if x is continous} \end{cases}$$

`Lưu ý là trung bình của biến ngẫu nhiên ở đây là trung bình với trọng lượng chứ không phải là trung bình cộng của xác suất biến ngẫu nhiên.`

Kỳ vọng còn được biết tới với những tên gọi khác như giá trị trung bình (Mean), giá trị trung bình có trọng lượng (Weighted Average),giá mong đợi (Expected Value) hay moment bậc một (first moment).

Kỳ vọng có một số tính chất như sau:

- $E(c) = c$ với c là hằng số
- $E(cX) = cE(X)$ với c là hằng số
- $E[aX + b] = aE[X] + b$ với a, b là các hằng số
- $E[X + Y] = E[X] + E[Y]$
- $E[XY] = E[X]E[Y]$ với X, Y là độc lập
- $E[g(X)] = \begin{cases} \sum_{\forall i} g(x_i)p_X(x_i) & \text{if x is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if x is continous} \end{cases}$

Phương sai

Dựa vào kì vọng ta sẽ có được trung bình của biến ngẫu nhiên, tuy nhiên nó lại không cho ta thông tin về mức độ phân tán xác suất nên ta cần 1 phương pháp để đo được độ phân tán đó. Một trong những phương pháp đó là phương sai (variance).

Phương sai $Var(X)$ là trung bình của bình phương khoảng cách từ biến ngẫu nhiên X tới giá trị trung bình:

$Var(X) = E[(X - E[X])^2]$

Việc tính toán dựa vào công thức này khá phức tạp, nên trong thực tế người ta thường sử dụng công thức tương đương sau:

$Var(X) = E[X^2] - E^2[X]$

Chứng minh:

$$\begin{aligned} Var(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E^2[X]] \\ &= E[X^2] - E[2XE[X]] + E[E^2[X]] \quad \text{, E[X] is constant} \\ &= E[X^2] - 2E[X]E[X] + E^2[X] \\ &= E[X^2] - E^2[X] \end{aligned}$$

Như vậy ta có thể thấy rằng phương sai luôn là một giá trị không âm và phương sai càng lớn thì nó thể hiện mức độ phân tán dữ liệu càng rộng hay nói cách khác mức độ ổn định càng nhỏ.

Phương sai có một số tính chất sau:

- $Var(c) = 0$với c là hằng số
- $Var(cX) = c^2Var(X)$ với c là hằng số

- $Var(aX + b) = a^2 Var(X)$ với a, b là các hằng số
- $Var(X + Y) = Var(X) + Var(Y)$ với X, Y là độc lập