1. Phân phối xác suất

1.1. Phân phối đồng thời

Hàm phân phối xác suất đồng thời hay hàm phân phối tích luỹ xác suất đồng thời (*Joint CDF - Joint Cumulative Probability Distribution Function*) của 2 biến ngẫu nhiên X, Y được định nghĩa như sau:

$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \quad , x, y \in \mathbb{R}$

Như vậy đây thực chất là hàm hợp xác suất của 2 biến ngẫu nhiên X, Y và tích luỹ xác suất được lấy là phần giao tích luỹ bên trái của X và bên trái của Y. Tương tự như với 1 biến ngẫu nhiên, hàm phân phối của 2 biến cũng là một hàm không giảm theo từng đối số 1 và ta còn có thể tính được tất cả các kiểu xác suất hợp của 2 biến X, Y thông qua hàm xác suất đồng thời.

2 biến ngẫu nhiên rời rạc

Hàm khối xác suất đồng thời (Joint PMF) của 2 biến ngẫu nhiên X, Y cùng rời rạc sẽ có dạng:

$p_{X,Y}(x, y) = P(X = x, Y = y)$

Khi đó với mỗi $p(x_i, y_j)$ là hàm khối xác suất đồng thời, ta có:

- $0 \leq p(x_i, y_j) \leq 1$
- $\sum\limits_{\forall i} \sum\limits_{\forall j} p(x_i, y_j) = 1$
- $F(x, y) = \sum\limits_{\forall x_i \leq x} \sum\limits_{\forall y_j \leq y} p(x_i, y_j)$

2 biến ngẫu nhiên liên tục

Còn hàm mật độ xác suất đồng thời (Joint PDF) của 2 biến ngẫu nhiên X, Y cùng liên tục có dạng:

$f(x, y) = \dfrac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x, y)$

Hay dưới dạng tích phân:

$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v) du dv$

Tương tự như trường hợp 1 biến ta có:

- $f(x, y) > 0$
- $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

- $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x,y)dxdy$
- $P(X = x, Y = y) = \int_{y}^{y} \int_{x}^{x} f(x,y)dxdy = 0$

Như vậy nếu để ý thì ta có thể nhớ 1 cách rằng trường hợp biến rời rạc ta lấy tổng còn biến là liên tục ta lấy tích phân. Đương nhiên là với biến rời rạc ta phải sử dụng hàm khối xác suất còn biến liên tục là hàm mật độ xác suất.

1.2. Phân phối biên

Phân phối biên (Marginal Probability) là phân phối của riêng từng biến một.

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(X \leq x, Y < +\infty) \\ &= F_{X,Y}(x, +\infty) \\ F_Y(y) &= P(Y \leq y) \\ &= P(+\infty, Y \leq y) \\ &= F_{X,Y}(+\infty, y) \end{aligned}$$

Đối với các biến rời rạc, ta có hàm khối xác suất biên (Marginal PMF):

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= \sum_{\forall j} p(x, y_j) \\ p_Y(y) &= P(Y = y) \\ &= \sum_{\forall i} p(x_i, y) \end{aligned}$$

Đối với các biến liên tục, ta có hàm mật độ xác suất biên (Marginal PDF):

$$\begin{aligned} f_X(x) &= P(X = x) \\ &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \frac{\partial}{\partial x} F_X(x) \\ f_Y(y) &= P(Y = y) \\ &= \int_{-\infty}^{\infty} f(x,y)dx \\ &= \frac{\partial}{\partial y} F_Y(y) \end{aligned}$$

Nếu bạn để ý sẽ thấy rằng công thức này khá giống với công thức xác suất đầy đủ khi tính xác suất của 1 sự kiện theo toàn bộ 1 sự kiện khác.

1.3. Biến độc lập

2 biến X,YX,Y độc lập khi xác suất của chúng không phụ thuộc vào nhau. Như ta đã biết 2 sự kiện A,BA,B độc lập khi và chỉ khi $P(AB) = P(A)P(B)$, tương tự với biến ngẫu nhiên chúng độc lập khi và chỉ khi

$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad ,\forall x,y \in \mathbb{R}$

Với trường hợp các biến ngẫu nhiên rời rạc:

$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad ,\forall x,y \in \mathbb{R}$

Với trường hợp các biến ngẫu nhiên liên tục:

$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad ,\forall x,y \in \mathbb{R}$

Như vậy từ đây ta có thể thấy rằng nếu các biến ngẫu nhiên là độc lập thì xác suất đồng thời của chúng có thể tính qua các xác suất biên của chúng bằng cách lấy tích chúng lại với nhau.

Ngoài ra ta còn có thể chứng minh được rằng nếu X, Y là độc lập với nhau thì hợp xác suất của chúng có thể được tính bằng tích của 2 hàm riêng từng biến một độc lập.

$$\begin{cases} p_{X,Y}(x,y) = u(x)v(y) & \text{for X, Y is discrete} \\ f_{X,Y}(x,y) = u(x)v(y) & \text{for X, Y is continous} \end{cases}$$

1.4. Phân phối có điều kiện

Tương tự như xác suất có điều kiện của các sự kiện ta cũng có thể biểu diễn các phân phối có điều kiện của các biến ngẫu nhiên.

Với X, Y là rời rạc:

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x|Y = y) \\ &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= \frac{p_{X,Y}(x,y)}{\sum_{\forall i} p(x_i,y)} \end{aligned}$$

Tương tự với X, Y là liên tục, ta có:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f(x,y)dy} \end{aligned}$$

Qua các phép biến đổi trên ta thấy rằng hợp phân phối của các biến ngẫu nhiên có thể tính toán tương tự như các phép kết hợp của các sự kiện mà ta đã làm quen ở bài đầu tiên.

2. Các đặc trưng

2.1. Kỳ vọng

Kỳ vọng của từng biến ngẫu nhiên vẫn được tính tương tự như trường hợp 1 biến ngẫu nhiên:

Với X, Y là biến ngẫu nhiên rời rạc:

$$\begin{aligned} E[X] &= \sum_i x_i p(x_i) \\ &= \sum_{\forall i} x_i \sum_{\forall j} p(x_i, y_j) \\ &= \sum_{\forall i} \sum_{\forall j} x_i p(x_i, y_j) \end{aligned}$$

Còn X, Y là liên tục:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy \end{aligned}$$

Các tính chất của kỳ vọng vẫn được giữ nguyên như trường hợp 1 biến. Ví dụ: giả sử rằng ta có hàm g(X,Y) định nghĩa 1 biến ngẫu nhiên mới thì khi đó:

$$E[g(X,Y)] = \begin{cases} \sum_{\forall i} \sum_{\forall j} g(x_i, y_j) p(x_i, y_j) & \text{for X, Y is discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & \text{for X, Y is continous} \end{cases}$$

2.2. Phương sai

Tương tự như trường hợp 1 biến ngẫu nhiên, phương sai của X là:

$$\begin{aligned} Var(X) &= E[(X - E[X])^2] \\ &= E[X^2] - E^2[X] \end{aligned}$$

Cũng như kỳ vọng, các tính chất của phương sai vẫn được bảo tồn như vậy.

2.3. Hiệp phương sai

Hiệp phương sai (Covariance) của 2 biến ngẫu nhiên X, Y kí hiệu là $Cov(X, Y)$ được định nghĩa rằng:

$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$

Tương tự như cách khai triển của phương sai, ta cũng sẽ thu được công thức tương đương sau:

$Cov(X, Y) = E[XY] - E[X]E[Y]$

Nếu X, Y là độc lập thì $E[XY] = E[X]E[Y]$ nên lúc này ta có $Cov(X, Y) = 0$, nhưng điều ngược lại chưa chắc đã đúng!

Hiệp phương sai có một số tính chất sau:

- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, X) = Var(X)$
- $Cov(aX, bY) = abCov(X, Y)$ với a, b là hằng số
- $Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$
- $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i}\sum_{j>i} Cov(X_i, X_j)$

Hiệp phương sai thường được tập hợp lại thành 1 ma trận đối xứng gọi là ma trận hiệp phương sai:

$$\begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{bmatrix}$$

2.4. Hệ số tương quan

Như ta đã biết khi hiệp phương sai bằng 0 thì vẫn chưa chắc được rằng chúng là độc lập mà khi đó ta sẽ đưa ra khái niệm là chúng không tương quan nhau. Còn trường hợp hiệp phương sai khác 0 thì ta nói rằng chúng tương quan với nhau. Với lý do tương tự khi đưa ra khái niệm độ lệch chuẩn (*Standard Deviation*, ta cũng sẽ đưa ra khái niệm **hệ số tương quan** (*Correlation*) được kí hiệu là $\rho(X, Y)$ và định nghĩa như sau:

$$\begin{aligned} \rho(X, Y) &= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \\ &= \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)} \end{aligned}$$

Ta có thể chứng minh được rằng $-1 \leq \rho(X, Y) \leq 1$.

Khi $|\rho(X, Y)| = 1$ thì giữa X, Y có quan hệ tuyến tính tức là: $Y = a + bX$.

Nếu $\rho(X,Y)=1$ thì $b=\dfrac{\sigma(Y)}{\sigma(X)}>0$ còn nếu $\rho(X,Y)=-1$ thì $b=-\dfrac{\sigma(Y)}{\sigma(X)}<0$

Hệ số tương quan cho ta thấy được các biến ngẫu nhiên có quan hệ tuyến tính chặt tới đâu tức là khi 1 biến biến thiên thì biến còn lại cũng sẽ biến thiên tương ứng. $|\rho(X,Y)|$ càng lớn thì ta nói rằng 2 biến có quan hệ tuyến tính càng chặt chẽ. $\rho(X,Y)>0$ ám chỉ rằng 2 biến là thuận biến với nhau, còn $\rho(X,Y)<0$ ám chỉ rằng 2 biến là nghịch biến với nhau. Khi $\rho(X,Y)=0$ ta nói rằng chúng không tương quan với nhau. Lưu ý rằng khi 2 biến độc lập thì chúng không tương quan nhưng điều ngược lại thì không đúng.

2.5. Đặc trưng có điều kiện

Các hàm khối xác suất của biến rời rạc và hàm mật độ xác suất của biến liên tục có điều kiện cũng có các đặc trưng như kỳ vọng và phương sai tương tự như các hàm khác chỉ khác 1 điều là thêm điều kiện tương ứng.

Kỳ vọng được định nghĩa như sau:

$$E[X|Y=y]=\begin{cases}\sum\limits_{i}^{\infty} x_i p_{X|Y}(x_i|y) & \text{for X,Y is discrete}\\ \int\limits_{-\infty}^{\infty} x f_{X|Y}(x|y)dx & \text{for X,Y is continous}\end{cases}$$

Kỳ vọng có điều kiện cũng có các tính chất tương tượng như kỳ vọng thông thường: - $E[g(Y)XY]=g(Y)E[XY]$ với $g(Y)$ là 1 hàm liên tục - $E[X_1+X_2|Y]=E[X_1|Y]+E[X_2|Y]$ - $E[X|Y]=E[X]$ nếu X,Y là độc lập

Một tính chất quan trọng nữa là $E[X]=E[E[X|Y]]$ tức là kì vọng của 1 biến có thể lấy bằng cách tương tự như xác suất của nó. Ta có thể biểu diễn tương tự như vậy:

$$E[X]=\begin{cases}\sum\limits_{j}^{\infty} E(X|Y=y_j)p_Y(y_j) & \text{for X,Y is discrete}\\ \int\limits_{-\infty}^{\infty} E(X|Y=y)f_Y(y)dy & \text{for X,Y is continous}\end{cases}$$

Một cách tương tự ta cũng định nghĩa phương sai có điều kiện như sau:

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

Ta cũng có $Var(X|Y)=E[X^2|Y]-E^2[X|Y]$, từ đây ta cũng có thể chứng minh rằng: $Var(X)=E[Var(X|Y)]+Var(E[X|Y])$

Do $Var(X|Y)=E[X^2|Y]-E^2[X|Y]$Var(X Y)=E[X, nên:

$$\begin{aligned}E[Var(X|Y)] &= E[E[X^2|Y]]-E[E^2[X|Y]]\\ &= E[X^2]-E[E^2[X|Y]]\end{aligned}$$

Ngoài ra,

$$\begin{aligned} Var(E[X|Y]) &= E[E^2[X|Y]] - E^2[E[X|Y]] \\ &= E[E^2[X|Y]] - E^2[X] \end{aligned}$$

Vậy nên:

$$Var(X) = E[Var(X|Y)] + Var(E[X|Y])$$

Kỳ vọng có điều kiện là nền tảng để có tạo được mối quan hệ giữa các biến ngẫu nhiên tức là ta có thể vẽ 1 đường dự đoán giá trị của YY khi biết XX bằng 1 hàm hồi quy của YY đối với X: $g(X) = E[Y|X]$. Thực chất để ước lượng được YY ta cần tìm hàm hồi quy $g(X)$ sao cho kỳ vọng khoảng cách của chúng là nhỏ nhất có thể $\text{argmin} E[(Y - g(X))^2] argmin$. Và giá trị nhỏ nhất này chính là: $g(X) = E[Y|X]$.