Véc tơ ngẫu nhiên

Nguyen Thi Ngoc Anh

AI Academy Vietnam

September 19, 2020

Nội dung

1. Véc tơ ngẫu nhiên và phân bố của véc tơ ngẫu nhiên
   - Phân phối xác suất của biến ngẫu nhiên hai chiều rời rạc
   - Phân phối xác suất của biến ngẫu nhiên hai chiều liên tục
2. Ma trận hiệp phương sai
   - Kỳ vọng và phương sai của các thành phần
   - Hiệp phương sai và hệ số tương quan
3. Hệ số tương quan
4. Ví dụ minh hoạ

Ví dụ xuất phát từ thực tế

- Trong thực tế nhiều khi ta phải xét đồng thời nhiều biến khác nhau có quan hệ tương quan.
- Một vấn đề xuất phát từ thực tế là Ủy ban an toàn giao thông Mỹ quan tâm tới việc sự dụng thắt đai an toàn của trẻ em trên xe ô tô liên quan tới mức độ an toàn tính mạng. Họ quan tâm tới đai an toàn cho các cháu dưới 5 tuổi. Một thống kê các vụ an tai nạn từ năm 1985 tới 1989, kết quả chỉ ra rằng:

Trạng thái có đai an toàn	Sống sót	Bị chết	Tổng số
Không có đai an toàn	1129	509	1638
Có đai an toàn người lớn	432	73	505
Có đai an toàn trẻ em	733	139	872
Tổng số	2294	721	3015

Ví dụ dẫn nhập biến ngẫu nhiên nhiều chiều

- Chúng ta định nghĩa

$$X = \begin{cases} 0, & \text{nếu em bé sống sót} \\ 1, & \text{nếu em bé bị chết.} \end{cases}$$

  Biến này sẽ mô tả con số sống sót của trẻ.

- Thông thường trên xe ô tô chỉ có đai an toàn của người lớn. Nếu có em bé trên xe, đai an toàn cho bé có thể được được sử dụng.

$$Y = \begin{cases} 0, & \text{không có đai an toàn} \\ 1, & \text{có đai an toàn của người lớn} \\ 2, & \text{nếu đai an toàn cho bé được sử dụng.} \end{cases}$$

  Biến này sẽ mô tả cho việc sử dụng đai an toàn.

Biến ngẫu nhiên nhiều chiều

- Phân phối xác suất đồng thời của hai biến ngẫu nhiên (X, Y)

Y / X	0	1	$\sum$
0	0.38	0.17	0.55
1	0.14	0.02	0.16
2	0.24	0.05	0.29
$\sum$	0.76	0.24	1

- $P(X = x, Y = y)$ là xác suất đồng thời của hai biến ngẫu nhiên (X, Y) nhận giá trị thể hiện tại (x, y).
- Ví dụ $P(X = 0, Y = 2) = \frac{733}{3015} = 0.24$ đây là xác suất chọn ngẫu nhiên một đứa trẻ từ một vụ tai nạn mà đứa trẻ này sống xót và sử dụng đai an toàn cho trẻ em.

Các khái niệm cơ sở

- Vec tơ ngẫu nhiên gồm na thành phần là một vec tớ gồm n biến ngẫu nhiên một chiều có dạng $(X_1, \cdots, X_n)$
- Để cho đơn giản, ta nghiên cứu biến ngẫu nhiên hai chiều (X, Y), trong đó X, Y là các biến ngẫu nhiên một chiều.
- Biến ngẫu nhiên hai chiều được gọi là rời rạc (liên tục) nếu các thành phần của nó là các biến ngẫu nhiên rời rạc (liên tục).

Phân phối xác suất của biến ngẫu nhiên hai chiều rời rạc

Bảng phân phối xác suất của biến ngẫu nhiên hai chiều (X, Y) rời rạc được xác định như sau

XY	y_1	$\cdots$	y_j	$\cdots$	y_n	$\sum_j$
x_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1n}	$P(X = x_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{in}	$P(X = x_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	p_{m1}	$\cdots$	p_{mj}	$\cdots$	p_{mn}	$P(X = x_m)$
$\sum_i$	$P(Y = y_1)$	$\cdots$	$P(Y = y_j)$	$\cdots$	$P(Y = y_n)$	1

Phân phối xác suất của biến ngẫu nhiên hai chiều rời rạc

Trong đó $p_{ij} = P\{X = x_i, Y = y_j\} \forall i = \overline{1, m}, j = \overline{1, n}$. Kích thước bảng này có thể chạy ra vô hạn khi m, n chạy ra vô hạn.

Tính chất

- $p_{ij} \geq 0 \forall i, j$;
- $\sum_{i,j} p_{ij} = 1$;
- Các phân phối biên được xác định như sau:

$$\begin{aligned} P(X = x_i) &= \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} \\ P(Y = y_j) &= \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij}. \end{aligned}$$

Ví dụ biến ngẫu nhiên nhiều chiều rời rạc

- Bảng phối xác suất đồng thời của hai biến ngẫu nhiên (X, Y)

Y / X	0	1	$\sum$
0	0.38	0.17	0.55
1	0.14	0.02	0.16
2	0.24	0.05	0.29
$\sum$	0.76	0.24	1

- $P(X = 1) = P((X = 1, Y = 0)+(X = 1, Y = 1)+(X = 1, Y = 2)) = P(X = 1, Y = 0) + P(X = 1, Y = 1) + P(X = 1, Y = 2) = 0.76.$
- Tương tự ta có các xác suất $P(X = 0), P(Y = 0), P(Y = 1), P(Y = 2).$

Bảng phân phối biên marginal distribution

- Bảng phân phối của biến ngẫu nhiên X

X	0	1
P(X=x)	0.76	0.24

- Bảng phân phối của biến ngẫu nhiên Y

Y	0	1	3
P(Y=y)	0.55	0.16	0.29

Hàm phân phối xác suất đồng thời của biến ngẫu nhiên nhiều chiều joint CDF

Definition

Hàm phân phối xác suất của biến ngẫu nhiên hai chiều (X, Y) được xác định như sau

$$F(x, y) = P(X < x, Y < y), x, y \in R. \quad (1)$$

Nhiều tài liệu gọi hàm trên là hàm phân phối xác suất đồng thời của hai biến X và Y. Đối với biến ngẫu nhiên (X, Y) rời rạc ta có: Hàm phân phối xác suất được xác định theo công thức $F(x, y) = \sum_{i,j:x_i<x,y_j<y} p_{ij}$;

Hàm phân phối xác suất đồng thời của biến ngẫu nhiên nhiều chiều

Tính chất

- $0 \leq F(x, y) \leq 1, \forall x, y \in R$;
- $F(x, y)$ là hàm không giảm theo từng đối số;
- $F(-\infty, y) = F(x, -\infty) = 0, \forall x, y \in R$ và $F(+\infty, +\infty) = 1$;
- Với $x_1 < x_2, y_1 < y_2$ ta luôn có $P(x_1 \leq X \leq x_2, y_1 \leq y \leq y_2) = F(x_2, y_2) + F(x_1, y_1) - F(x_1, y_2) - F(x_2, y_1)$.

Các khái niệm cơ sở

Tính chất (tiếp)

- Các hàm

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

là các hàm phân phối riêng của các biến ngẫu nhiên X và Y và còn được gọi là các *phân phối biên* của biến ngẫu nhiên hai chiều (X, Y).

Các khái niệm cơ sở

Definition

Hai biến ngẫu nhiên X, Y được gọi là *độc lập* nếu

$$F(x, y) = F_X(x).F_Y(y), \forall x, y \in R.$$

- Hai biến ngẫu nhiên X, Y được gọi là độc lập với nhau nếu ta có

$$P(X = x_i, Y = y_j) = P(X = x_i).P(Y = y_j), \forall i = \overline{1, m}, j = \overline{1, n}$$

Phân phối xác suất của biến ngẫu nhiên hai chiều liên tục

Definition

Hàm hai biến không âm, liên tục $f(x, y)$ được gọi là *hàm mật độ xác suất đồng thời* của biến ngẫu nhiên hai chiều liên tục $(X < Y)$ nếu nó thỏa mãn

$$P\{(X, Y) \in D\} = \int\int_D f(x, y)dxdy \forall D \subset R^2. \quad (2)$$

Tính chất

- $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dudv;$
- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dxdy = 1.$

Phân phối xác suất của biến ngẫu nhiên hai chiều liên tục

Tính chất (tiếp)

- $f(x,y) = \dfrac{\partial^2 F(x,y)}{\partial x \partial y}$;
- Các hàm mật độ biên
  - theo x : $f_X(x) = \int\limits_{-\infty}^{+\infty} f(x,y)dy$;
  - theo y : $f_Y(y) = \int\limits_{-\infty}^{+\infty} f(x,y)dx$.
- Hai biến ngẫu nhiên X và Y được gọi là độc lập nếu $f(x,y) = f_X(x).f_Y(y) \forall x, y$.

Kỳ vọng và phương sai của các thành phần

Trường hợp (X, Y) rời rạc

$$EX = \sum_i P(X = x_i) = \sum_i \sum_j x_i p_{ij};$$

$$EY = \sum_j y_j P(Y = y_j) = \sum_i \sum_j y_j p_{ij}$$

$$VX = \sum_i \sum_j x_i^2 p_{ij} - \{EX\}^2;$$

$$VY = \sum_i \sum_j y_j^2 p_{ij} - \{EY\}^2.$$

Kỳ vọng và phương sai của các thành phần

Trường hợp (X, Y) liên tục

$$EX = \int\int_{R^2} x.f(x,y)dxdy;$$

$$EY = \int\int_{R^2} y.f(x,y)dxdy$$

$$VX = \int\int_{R^2} x^2.f(x,y)dxdy - \{EX\}^2;$$

$$VY = \int\int_{R^2} y^2.f(x,y)dxdy - \{EY\}^2.$$

Kỳ vọng và phương sai của các thành phần

Đối với biến ngẫu nhiên $Z = g(X, Y)$ ta có

$$EZ = E\{g(X, Y)\} = \int\int_{R^2} g(x, y).f(x, y)dxdy$$

Hiệp phương sai Covariance

Definition

Cho biến ngẫu nhiên hai chiều (X, Y), hiệp phương sai của hai thành phần X và Y, kí hiệu là μ_{XY} , được xác định bởi

$$\boxed{cov(X,Y) = E\{(X - EX)(Y - EY)\} = E(XY) - EX.EY,} \quad (3)$$

trong đó $E(XY)$ được xác định theo công thức

$$E(XY) = \begin{cases} \sum\limits_i \sum\limits_j x_i y_j p_{ij}, & \text{đối với biến ngẫu nhiên rời rạc} \\ \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} xy.f(x,y), & \text{đối với biến ngẫu nhiên liên tục} \end{cases}$$

Hiệp phương sai và sự tương quan Covariance, correlation coefficient

Definition

Ta nói rằng X và Y *không tương quan* nếu $cov(X, Y) = 0$.

Nhận xét

- $cov(X, Y) = cov(Y, X)$;
- Phương sai chính là trường hợp riêng của hiệp phương sai ($VX = cov(X, X), VY = cov(Y, Y)$);
- Nếu X, Y độc lập thì ta có $E(XY) = EX.EY$. Khi đó $cov(X, Y) = 0$, tức là X và Y không tương quan. Vậy ta có, nếu hai biến ngẫu nhiên độc lập thì không tương quan. Điều ngược lại chưa chắc đã đúng.

Hiệp phương sai Covariance matrix

Definition

Ma trận hiệp phương sai và hệ số tương quan của biến ngẫu nhiên hai chiều (X, Y) được xác định bởi

$$\Gamma = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix}$$

Definition

Hệ số tương quan correlation coefficient của hai biến ngẫu nhiên X và Y, ký hiệu là ρ_{XY} và được xác định theo công thức

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}. \tag{4}$$

Hệ số tương quan correlation coefficient

Nhận xét:

- Có thể chứng minh được $|\rho_{XY}| \leq 1$. Nếu $\rho_{XY} = \pm 1$ ta nói hai biến ngẫu nhiên X và Y có tương quan tuyến tính;
- Nếu $\rho_{XY} = 0$ ta nói hai biến ngẫu nhiên X và Y là không tương quan.

Ví dụ về phân phối có điều kiện

Bảng phân phối xác suất đồng thời của ví dụ về bài toán thắt đai an toàn của trẻ em
$Y = \{0, 1, 2\}$ tương ứng với việc không sử dung thắt đai an toàn, sử dụng thắt đai an toàn người lớn, sử dụng thắt đai an toàn trẻ em. $X = \{0, 1\}$ tương ứng với trẻ sống sót, trẻ bị chết.

Y / X	0	1	$\sum$
0	0.38	0.17	0.55
1	0.14	0.02	0.16
2	0.24	0.05	0.29
$\sum$	0.76	0.24	1

Ma trận hiệp phương sai và hệ số tương quan

- $EX = 0.24, EY = 0.74$
- $E(XY) = 0.12$
- $Var(X) = 0.1824, Var(Y) = 0.7724, Cov(X, Y) = -0.0576$
- Hệ số tương quan $\rho_{XY} = -0.1535$
- Ma trận hiệp phương sai

$$\Gamma = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix} = \begin{bmatrix} 0.1824 & -0.0576 \\ -0.0576 & 0.7724 \end{bmatrix}$$

Hiệp phương sai và hệ số tương quan

```
import numpy as np
x=np.array([0,1])# Gía trị X có thể nhận
y=np.array([0,1,2])# Giá trị Y có thể nhận
jpmf=np.array([[0.38,0.17], [0.14,0.02],[0.24,0.05]])# Ma trận trọng số
```

```
pmfx=[]# Hàm trọng số biên của X
for i in x:
    pmfx.append(jpmf[:,i].sum())
```

```
pmfy=[]# Hàm trọng số biên của Y
for j in y:
    pmfy.append(jpmf[j,:].sum())
```

```
EX=np.sum(x*pmfx); EY=np.sum(y*pmfy)# Kỳ vọng của từng biến
```

```
VarX=np.sum((x-EX)**2*pmfx); VarY=np.sum((y-EY)**2*pmfy)# Phương sai
```

```
EXY=0 # Hàm tương quan
for i in x:
    for j in y:
        EXY= EXY+ i*j*jpmf[j,i]
```

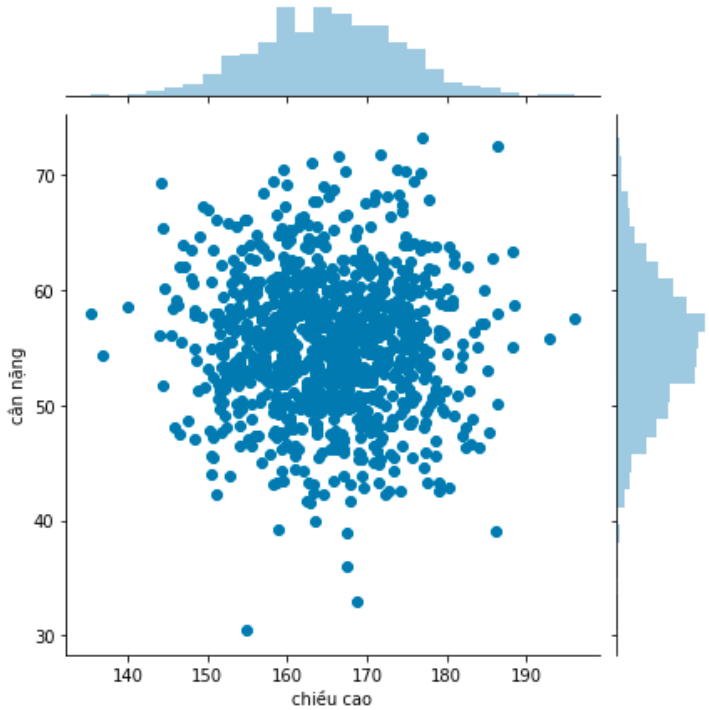
```
cov=EXY-EX*EY # Hiệp phương sai
```

Ví dụ mô phỏng biến ngẫu nhiên tuân theo luật phân phối chuẩn 2 chiều

```
import seaborn as sns
import numpy as np
mean = [165, 55]# véc tơ kỳ vọng 2 chiều
cov = [[81, 0.6], [0.6, 36]]# ma trận hiệp phương sai
```

```
chieu_cao, can_nang = np.random.multivariate_normal(mean, cov, 1000).T
sns.jointplot(chieu_cao, can_nang, stat_func=None).set_axis_labels("chiều cao", "cân nặng")
```

```
<seaborn.axisgrid.JointGrid at 0x1a1ac9bb00>
```



Bài tập thực hành 1

- Tung hai con xúc sắc cân đối, đồng chất gọi X, Y lần lượt là số chấm xuất hiện trên mặt của con xúc sắc thứ thứ 1 và thứ hai.
- Tìm hàm phân phối biên của X, Y
- Tìm kỳ vọng, phương sai của X, Y
- tìm $cov(X, Y)$, hệ số tương quan
- Tìm ma trận hiệp phương sai của (X, Y)

Bài tập thực hành 2

- Cho hai biến ngẫu nhiên tuân theo luật phân phối chuẩn độc lập với nhau có kỳ vọng lần lượt là 1, 2 và phương sai là 25, 9.
- Tìm ma trận hiệp phương sai
- biểu diễn véc tơ kỳ vọng, ma trận hiệp phương sai qua python
- dùng lệnh np.random.multivariate_normal() để tạo ra 5000 giá trị
- Vẽ đồ thị các giá trị đồng thời mà (X, Y) có thể nhận và phân phối biên trong cùng một đồ thị.