

VIP Refresher: Xác suất và thống kê

Afshine AMIDI và Shervine AMIDI

Ngày 17 tháng 5 năm 2020

Dịch bởi Hoàng Minh Tuấn và Hung Nguyễn

Giới thiệu về Xác suất và Tổ hợp

□ **Không gian mẫu** – Một tập hợp các kết cục có thể xảy ra của một phép thử được gọi là không gian mẫu của phép thử và được kí hiệu là S .

□ **Sự kiện (hay còn gọi là biến cố)** – Bất kỳ một tập hợp con E nào của không gian mẫu đều được gọi là một sự kiện. Một sự kiện là một tập các kết cục có thể xảy ra của phép thử. Nếu kết quả của phép thử chứa trong E , chúng ta nói sự kiện E đã xảy ra.

□ **Tiền đề của xác suất** – Với mỗi sự kiện E , chúng ta kí hiệu $P(E)$ là xác suất sự kiện E xảy ra.

$$(1) \quad 0 \leq P(E) \leq 1 \quad (2) \quad P(S) = 1 \quad (3) \quad P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$$

□ **Hoán vị** – Hoán vị là một cách sắp xếp r phần tử từ một nhóm n phần tử, theo một thứ tự nhất định. Số lượng cách sắp xếp như vậy là $P(n, r)$, được định nghĩa như sau:

$$P(n, r) = \frac{n!}{(n-r)!}$$

□ **Tổ hợp** – Một tổ hợp là một cách sắp xếp r phần tử từ n phần tử, không quan trọng thứ tự. Số lượng cách sắp xếp như vậy là $C(n, r)$, được định nghĩa như sau:

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n-r)!}$$

Ghi chú: Chúng ta lưu ý rằng với $0 \leq r \leq n$, ta có $P(n, r) \geq C(n, r)$

Xác suất có điều kiện

□ **Định lý Bayes** – Với các sự kiện A và B sao cho $P(B) > 0$, ta có:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Ghi chú: ta có $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$

□ **Phân vùng** – Cho $\{A_i, i \in [1, n]\}$ sao cho với mỗi i , $A_i \neq \emptyset$. Chúng ta nói rằng $\{A_i\}$ là một phân vùng nếu có:

$$\forall i \neq j, A_i \cap A_j = \emptyset \quad \text{và} \quad \bigcup_{i=1}^n A_i = S$$

Ghi chú: với bất cứ sự kiện B nào trong không gian mẫu, ta có $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$.

□ **Định lý Bayes mở rộng** – Cho $\{A_i, i \in [1, n]\}$ là một phân vùng của không gian mẫu. Ta có:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

□ **Sự kiện độc lập** – Hai sự kiện A và B được coi là độc lập khi và chỉ khi ta có:

$$P(A \cap B) = P(A)P(B)$$

Biến ngẫu nhiên

□ **Biến ngẫu nhiên** – Một biến ngẫu nhiên, thường được kí hiệu là X , là một hàm nối mỗi phần tử trong một không gian mẫu thành một số thực.

□ **Hàm phân phối tích lũy (CDF)** – Hàm phân phối tích lũy F , là một hàm đơn điệu không giảm, sao cho $\lim_{x \rightarrow -\infty} F(x) = 0$ và $\lim_{x \rightarrow +\infty} F(x) = 1$, được định nghĩa là:

$$F(x) = P(X \leq x)$$

Ghi chú: chúng ta có $P(a < X \leq b) = F(b) - F(a)$.

□ **Hàm mật độ xác suất (PDF)** – Hàm mật độ xác suất f là xác suất mà X nhận các giá trị giữa hai giá trị thực liên tiếp của biến ngẫu nhiên.

□ **Mối quan hệ liên quan giữa PDF và CDF** – Dưới đây là các thuộc tính quan trọng cần biết trong trường hợp rời rạc (D) và liên tục (C).

Trường hợp	CDF F	PDF f	Thuộc tính của PDF
(D)	$F(x) = \sum_{x_i \leq x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$0 \leq f(x_j) \leq 1$ và $\sum_j f(x_j) = 1$
(C)	$F(x) = \int_{-\infty}^x f(y)dy$	$f(x) = \frac{dF}{dx}$	$f(x) \geq 0$ và $\int_{-\infty}^{+\infty} f(x)dx = 1$

□ **Phương sai** – Phương sai của một biến ngẫu nhiên, thường được kí hiệu là $\text{Var}(X)$ hoặc σ^2 , là một độ đo mức độ phân tán của hàm phân phối. Nó được xác định như sau:

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

□ **Độ lệch chuẩn** – Độ lệch chuẩn của một biến ngẫu nhiên, thường được kí hiệu σ , là thước đo mức độ phân tán của hàm phân phối của nó so với các đơn vị của biến ngẫu nhiên thực tế. Nó được xác định như sau:

$$\sigma = \sqrt{\text{Var}(X)}$$

□ **Kỳ vọng và moment của phân phối** – Dưới đây là các biểu thức của giá trị kỳ vọng $E[X]$, giá trị kỳ vọng tổng quát $E[g(X)]$, moment bậc k $E[X^k]$ và hàm đặc trưng $\psi(\omega)$ cho các trường hợp rời rạc và liên tục:

Trường hợp	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\sum_{i=1}^n x_i f(x_i)$	$\sum_{i=1}^n g(x_i) f(x_i)$	$\sum_{i=1}^n x_i^k f(x_i)$	$\sum_{i=1}^n f(x_i) e^{i\omega x_i}$
(C)	$\int_{-\infty}^{+\infty} x f(x) dx$	$\int_{-\infty}^{+\infty} g(x) f(x) dx$	$\int_{-\infty}^{+\infty} x^k f(x) dx$	$\int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx$

□ **Biến đổi các biến ngẫu nhiên** – Đặt các biến X và Y được liên kết với nhau bởi một hàm. Kí hiệu f_X và f_Y lần lượt là các phân phối của X và Y , ta có:

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

□ **Quy tắc tích phân Leibniz** – Gọi g là hàm của x và có khả năng c , và a, b là các ranh giới có thể phụ thuộc vào c . Chúng ta có:

$$\frac{\partial}{\partial c} \left(\int_a^b g(x) dx \right) = \frac{\partial b}{\partial c} \cdot g(b) - \frac{\partial a}{\partial c} \cdot g(a) + \int_a^b \frac{\partial g}{\partial c}(x) dx$$

□ **Bất đẳng thức Chebyshev** – Gọi X là biến ngẫu nhiên có giá trị kỳ vọng μ . Với $k, \sigma > 0$, chúng ta có bất đẳng thức sau:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Phân phối đồng thời biến ngẫu nhiên

□ **Mật độ có điều kiện** – Mật độ có điều kiện của X với Y , thường được kí hiệu là $f_{X|Y}$, được định nghĩa như sau:

$$f_{X|Y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

□ **Tính chất độc lập** – Hai biến ngẫu nhiên X và Y độc lập nếu ta có:

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

□ **Mật độ biên và phân phối tích lũy** – Từ hàm phân phối mật độ đồng thời f_{XY} , ta có

Trường hợp	Mật độ biên	Hàm tích lũy
(D)	$f_X(x_i) = \sum_j f_{XY}(x_i, y_j)$	$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f_{XY}(x_i, y_j)$
(C)	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$	$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(x', y') dx' dy'$

□ **Tính chất độc lập** – Hai biến ngẫu nhiên X và Y độc lập nếu ta có:

$$\psi_{X+Y}(\omega) = \psi_X(\omega) \times \psi_Y(\omega)$$

□ **Hiệp phương sai** – Chúng ta xác định hiệp phương sai của hai biến ngẫu nhiên X và Y , thường được kí hiệu σ_{XY}^2 hay $\text{Cov}(X, Y)$, như sau:

$$\text{Cov}(X, Y) \triangleq \sigma_{XY}^2 = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

□ **Hệ số tương quan** – Kí hiệu σ_X, σ_Y là độ lệch chuẩn của X và Y , chúng ta xác định hệ số tương quan giữa X và Y , kí hiệu ρ_{XY} , như sau:

$$\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}$$

Ghi chú 1: chúng ta lưu ý rằng với bất cứ biến ngẫu nhiên X, Y nào, ta luôn có $\rho_{XY} \in [-1, 1]$.

Ghi chú 2: Nếu X và Y độc lập với nhau thì $\rho_{XY} = 0$.

□ **Các phân phối chính** – Dưới là các phân phối chính cần ghi nhớ:

Loại	Phân phối	PDF	$\psi(\omega)$	$E[X]$	$\text{Var}(X)$
(D)	$X \sim \mathcal{B}(n, p)$ Binomial	$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $x \in \llbracket 0, n \rrbracket$	$(pe^{i\omega} + q)^n$	np	npq
	$X \sim \text{Po}(\mu)$ Poisson	$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega} - 1)}$	μ	μ
(C)	$X \sim \mathcal{U}(a, b)$ Uniform	$f(x) = \frac{1}{b-a}$ $x \in [a, b]$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X \sim \mathcal{N}(\mu, \sigma)$ Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
	$X \sim \text{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1 - \frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

$$\bar{X} \underset{n \rightarrow +\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Ước lượng tham số

□ **Mẫu ngẫu nhiên** – Mẫu ngẫu nhiên là tập hợp của n biến ngẫu nhiên X_1, \dots, X_n độc lập và được phân phối giống hệt với X .

□ **Công cụ ước tính** – Công cụ ước tính (estimator) là một hàm của dữ liệu được sử dụng để suy ra giá trị của một tham số chưa biết trong mô hình thống kê.

□ **Thiên vị** – Thiên vị (bias) của Estimator $\hat{\theta}$ được định nghĩa là chênh lệch giữa giá trị kì vọng của phân phối $\hat{\theta}$ và giá trị thực, tức là

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Ghi chú: một công cụ ước tính được cho là không thiên vị (unbiased) khi chúng ta có $E[\hat{\theta}] = \theta$.

□ **Giá trị trung bình mẫu** – Giá trị trung bình mẫu của mẫu ngẫu nhiên được sử dụng để ước tính giá trị trung bình thực μ của phân phối, thường được kí hiệu \bar{X} và được định nghĩa như sau:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Ghi chú: trung bình mẫu là không thiên vị (unbiased), nghĩa là $E[\bar{X}] = \mu$.

□ **Phương sai mẫu** – Phương sai mẫu của mẫu ngẫu nhiên được sử dụng để ước lượng phương sai thực sự σ^2 của phân phối, thường được kí hiệu là s^2 hoặc $\hat{\sigma}^2$ và được định nghĩa như sau:

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Ghi chú: phương sai mẫu không thiên vị (unbiased), nghĩa là $E[s^2] = \sigma^2$.

□ **Định lý giới hạn trung tâm** – Giả sử chúng ta có một mẫu ngẫu nhiên X_1, \dots, X_n theo một phân phối nhất định với trung bình μ và phương sai σ^2 , sau đó chúng ta có: