TRƯỜNG ĐẠI HỌC BÁCH KHOA TP. HỒ CHÍ MINH
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

CƠ SỞ TOÁN CHO KHOA HỌC MÁY TÍNH (CO5263)

XÁC SUẤT VÀ THỐNG KÊ

Giảng viên hướng dẫn & Học viên

GVHD:
TS. Nguyễn An Khương
TS. Trần Tuấn Anh

Danh sách học viên nhóm 7:
Trần Đăng Hùng – 2470750
Trần Hoài Tâm – 2470743
Ngô Minh Đại – 2470722
Vương Minh Toàn – 2491057
Nguyễn Đình Nhật Minh – 2370736
Nguyễn Xuân Hiền – 2470749

Nội dung

1. Các khái niệm cơ bản về xác suất
2. Biến ngẫu nhiên
3. Ứng dụng
4. Ví dụ về bài toán HIV/AIDS
5. Hàm mật độ xác suất, hàm phân phối tích lũy
6. Kỳ vọng, phương sai, độ lệch chuẩn
7. Bài tập

Không gian mẫu, biến cố, các phép toán trên biến cố

Không gian mẫu, biến cố

Thí nghiệm (experiment): Trong lý thuyết xác suất, một *thí nghiệm* (experiment) là một quy trình hay một phép đo thu được một quan sát hoặc thu được một kết quả đầu ra (outcome) không được dự đoán chắc chắn (certainty).

Ví dụ: Tung đồng xu, Đo lượng mưa, Phỏng vấn lấy ý kiến.

Sự kiện đơn giản (simple event): Khi thực hiện một thí nghiệm, kết quả thu được gọi là *sự kiện đơn giản*, ký hiệu E_i.

Ví dụ: Gọi E_i là sự kiện xuất hiện mặt i khi tung xí ngầu 6 mặt, với $i = 1, \ldots, 6$.

Không gian mẫu, biến cố, các phép toán trên biến cố

Không gian mẫu, biến cố

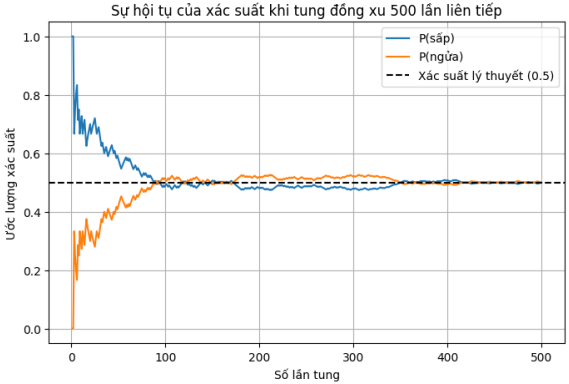
Biến cố (event): Một *biến cố* (event) là một tập hợp các sự kiện đơn giản.

Biến cố loại trừ lẫn nhau: Hai biến cố là *loại trừ lẫn nhau* (mutually exclusive) nếu không thể xảy ra cùng lúc.

Không gian mẫu (sample space): Là tập hợp tất cả các sự kiện đơn giản (simple events), ký hiệu S. Mỗi biến cố là một tập con của S.

Tổng quan về xác suất

Thí nghiệm tung đồng xu 500 lần



Hình: Thí nghiệm tung đồng xu 500 lần liên tiếp

Định nghĩa về xác suất

Theo quan điểm tần suất

Giả định rằng một thí nghiệm được lặp đi lặp lại nhiều lần dưới điều kiện giống hệt nhau và tần suất tương đối (relative frequency) của một biến cố là tỷ lệ số lần biến cố đó xảy ra.

Định nghĩa về xác suất

Ba tiên đề xác suất cơ bản (Kolmogorov)

- Mỗi xác suất của một sự kiện đơn giản nằm giữa 0 và 1 (bao gồm 0 và 1):

$$0 \leq P(E) \leq 1$$

- Tổng xác suất của tất cả các sự kiện đơn giản trong không gian mẫu bằng 1:

$$P(S) = 1$$

- Với bất kỳ dãy biến cố loại trừ $E_1, E_2, \ldots$ sao cho $E_i \cap E_j = \emptyset$ với $i \neq j$, ta có:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \tag{1}$$

Mối quan hệ biến cố và Quy luật xác suất

- **Hợp (union)** của hai biến cố $\mathcal{A}$ và $\mathcal{B}$, ký hiệu $\mathcal{A} \cup \mathcal{B}$ là biến cố "$\mathcal{A}$ hoặc $\mathcal{B}$ hoặc cả hai" xảy ra. Xác suất (công thức cộng):

$$P\Big(\mathcal{A} \cup \mathcal{B}\Big) \;=\; P(\mathcal{A}) + P(\mathcal{B}) - P\Big(\mathcal{A} \cap \mathcal{B}\Big) \tag{2}$$

  Đặc biệt, nếu $\mathcal{A}$ và $\mathcal{B}$ loại trừ nhau (mutually exclusive hay disjoint), thì:

$$P\Big(\mathcal{A} \cup \mathcal{B}\Big) \;=\; P(\mathcal{A}) + P(\mathcal{B}) \quad \text{vì} \quad P\Big(\mathcal{A} \cap \mathcal{B}\Big) = 0 \tag{3}$$

- **Phần bù (complement)** của biến cố $\mathcal{A}$, ký hiệu $\mathcal{A}^c$, là biến cố "$\mathcal{A}$ không xảy ra".

$$P\Big(\mathcal{A}^c\Big) \;=\; 1 - P(\mathcal{A}) \tag{4}$$

- **Giao (intersection)** của hai biến cố $\mathcal{A}$ và $\mathcal{B}$, ký hiệu $\mathcal{A} \cap \mathcal{B}$, là biến cố "cả $\mathcal{A}$ và $\mathcal{B}$ cùng xảy ra".

Các biến cố độc lập

Hai biến cố $\mathcal{A}$ và $\mathcal{B}$ được coi là **độc lập** (independent events) nếu việc xảy ra của biến cố $\mathcal{A}$ không ảnh hưởng đến xác suất xảy ra của biến cố $\mathcal{B}$. Nếu hai biến cố không độc lập, chúng được gọi là **phụ thuộc** (dependent). Khái niệm biến cố độc lập có liên quan chặt chẽ với xác suất có điều kiện.

Xác suất có điều kiện và công thức nhân xác suất

Xác suất có điều kiện

Xác suất biến cố $\mathcal{A}$ tìm được khi biến cố $\mathcal{B}$ xảy ra được gọi là xác suất có điều kiện (conditional probability) của $\mathcal{A}$, ký hiệu

$$P\Big(\mathcal{A} \mid \mathcal{B}\Big)$$

với điều kiện $P(\mathcal{B}) > 0$.

Khái niệm xác suất có điều kiện rất quan trọng khi chúng ta quan tâm đến việc tính toán xác suất khi có sẵn một số thông tin bộ phận liên quan đến kết quả của một thí nghiệm. Ngay cả khi không có thông tin bộ phận nào, xác suất có điều kiện vẫn có thể giúp tính toán các xác suất mong muốn dễ dàng hơn.

Xác suất có điều kiện và công thức nhân xác suất

Công thức nhân xác suất tổng quát

Công thức nhân xác suất (Multiplication rule) liên quan đến xác suất đồng thời của hai hoặc nhiều biến cố và thường được sử dụng cùng với xác suất có điều kiện, được xác định như sau:

$$P\Big(\mathcal{A} \cap \mathcal{B}\Big) \;=\; P(\mathcal{A})\, P\Big(\mathcal{B} \mid \mathcal{A}\Big) \;=\; P(\mathcal{B})\, P\Big(\mathcal{A} \mid \mathcal{B}\Big) \tag{5}$$

Từ công thức nhân xác suất tổng quát, ta suy ra công thức tính xác suất có điều kiện:

$$P\Big(\mathcal{A} \mid \mathcal{B}\Big) \;=\; \frac{P\Big(\mathcal{A} \cap \mathcal{B}\Big)}{P(\mathcal{B})} \quad \text{với} \quad P(\mathcal{B}) > 0 \tag{6}$$

hoặc

$$P\Big(\mathcal{B} \mid \mathcal{A}\Big) \;=\; \frac{P\Big(\mathcal{A} \cap \mathcal{B}\Big)}{P(\mathcal{A})} \quad \text{với} \quad P(\mathcal{A}) > 0$$

Xác suất có điều kiện và công thức nhân xác suất

Công thức nhân xác suất cho các biến cố độc lập

Nếu hai biến cố $\mathcal{A}$ và $\mathcal{B}$ độc lập thì:

$$P(\mathcal{B} \mid \mathcal{A}) = P(\mathcal{B}) \quad \text{hay} \quad P(\mathcal{A} \mid \mathcal{B}) = P(\mathcal{A}) \tag{7}$$

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A})\, P(\mathcal{B}) \tag{8}$$

Nếu ba biến cố $\mathcal{A}, \mathcal{B}, \mathcal{C}$ độc lập (hoặc độc lập từng đôi một) thì:

$$P(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}) = P(\mathcal{A})\, P(\mathcal{B})\, P(\mathcal{C}) \tag{9}$$

Công thức xác suất đầy đủ và công thức Bayes

Công thức xác suất đầy đủ (Law of total probability)

Cho tập các sự kiện đơn giản $S_1, S_2, \ldots, S_k$ loại trừ lẫn nhau và đầy đủ, xác suất của một biến cố $\mathcal{A}$ được xác định như sau:

$$P(\mathcal{A}) = P(S_1)\,P(\mathcal{A} \mid S_1) \;+\; P(S_2)\,P(\mathcal{A} \mid S_2) \;+\; \ldots \;+\; P(S_k)\,P(\mathcal{A} \mid S_k) \tag{10}$$

Công thức xác suất đầy đủ và công thức Bayes

Công thức Bayes

Công thức Bayes (Bayes's rule hoặc Bayes's formula) là một công thức quan trọng trong xác suất. Công thức này cho phép tính toán xác suất hậu nghiệm (posterior probabilities) dựa trên xác suất tiên nghiệm (prior probabilities) và xác suất có điều kiện. Cụ thể, với k tập sự kiện đơn giản $S_1, S_2, \ldots, S_k$ loại trừ lẫn nhau và đầy đủ với xác suất tiên nghiệm $P(S_1), P(S_2), \ldots, P(S_k)$, nếu biến cố $\mathcal{A}$ xảy ra, xác suất hậu nghiệm có điều kiện của S_i được xác định như sau:

$$P(S_i \mid \mathcal{A}) = \frac{P(S_i)\, P(\mathcal{A} \mid S_i)}{\sum_{j=1}^{k} P(S_j)\, P(\mathcal{A} \mid S_j)} \quad \text{với } i = 1, 2, \ldots, k \tag{11}$$

Biến ngẫu nhiên

Định nghĩa

Định nghĩa 2.1 [6]

Một biến ngẫu nhiên (random variable) với giá trị thực là một hàm số đo được trên một không gian xác suất.

$$X : (\Omega, P) \longrightarrow \mathbb{R}$$

Ví dụ: Tung một con xúc xắc cân đối.

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad P(\{i\}) = \tfrac{1}{6}.$$

Định nghĩa biến ngẫu nhiên

$$Y : \Omega \longrightarrow \mathbb{R}, \quad Y(\omega) = \omega$$

(giá trị mặt xúc xắc). Y là biến ngẫu nhiên vì với mọi tập $\{a < Y < b\}$, tập $\{\omega \mid a < \omega < b\}$ là tập đo được trong Ω.

Biến ngẫu nhiên

Định nghĩa

Định nghĩa 2.2 [6]

Nếu ta có hai biến ngẫu nhiên X, Y (với cùng một mô hình không gian xác suất), thì ta sẽ nói rằng

$$X = Y \text{ theo nghĩa xác suất}$$

(hay $X = Y$ hầu khắp mọi nơi), nếu như sự kiện "$X = Y$" có xác suất bằng 1.

Ví dụ: Giả sử Z là biến ngẫu nhiên đồng nhất trên đoạn $[0, 1]$. Định nghĩa

$$U(\omega) = \begin{cases} Z(\omega), & \text{nếu } Z(\omega) \neq 0.5, \\ 0, & \text{nếu } Z(\omega) = 0.5. \end{cases}$$

Vì $P(Z = 0.5) = 0$, nên $P(U = Z) = 1$. Do đó $U = Z$ hầu khắp mọi nơi.

Biến ngẫu nhiên rời rạc

Định nghĩa

Một **biến ngẫu nhiên rời rạc** (Random discrete variable) chỉ có thể nhận một tập giá trị hữu hạn hoặc đếm được.

Biến ngẫu nhiên rời rạc

Tính chất phân phối xác suất

Phân phối xác suất (probability distribution) đối với một biến ngẫu nhiên rời rạc có thể là:

- Hàm xác suất (probability mass function – PMF).
- Bảng phân phối xác suất (probability distribution table – PMT).
- Đồ thị biểu diễn các giá trị x và xác suất tương ứng $p(x)$.

Tính chất của phân phối xác suất cho biến ngẫu nhiên rời rạc:

- $0 \leq p(x) \leq 1$.
- $\sum_{x} p(x) = 1$.

Biến ngẫu nhiên rời rạc

Các tham số đặc trưng

- **Giá trị kỳ vọng (expected value) / Trung bình (mean)** của x:

$$\mu = E(x) = \sum_x x\, p(x) \tag{12}$$

- **Phương sai (variance)** của x:

$$\sigma^2 = E\left[(x-\mu)^2\right] = \sum_x (x-\mu)^2\, p(x) \tag{13}$$

- **Độ lệch chuẩn (standard deviation)** của x là σ.

Biến ngẫu nhiên rời rạc

Một số phân phối xác suất phổ biến

- **Bernoulli** ($X \in \{0, 1\}$, tham số p):

$$p(0) = 1 - p, \quad p(1) = p \tag{14}$$

- **Binomial** ($X \sim \mathrm{Bin}(n, p)$):

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i} \tag{15}$$

$$E[X] = np, \quad \mathrm{Var}(X) = np(1-p) \tag{16,17}$$

$$P\{X \leq i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k} \tag{18}$$

$$P\{X = k+1\} = \left(\frac{p}{1-p}\right)\left(\frac{n-k}{k+1}\right) P\{X = k\} \tag{19}$$

Biến ngẫu nhiên rời rạc

Một số phân phối xác suất phổ biến

- **Poisson** ($X \sim \text{Poi}(\lambda)$):

$$P\{X = i\} = \frac{e^{-\lambda}\lambda^i}{i!} \tag{19}$$

$$\sum_{i=0}^{\infty} P\{X = i\} = 1 \tag{20}$$

$$P\{X = i\} \approx \frac{e^{-\lambda}\lambda^i}{i!}, \quad \lambda = np \tag{21}$$

$$E[X] = \lambda, \quad \text{Var}(X) = \lambda \tag{22,23}$$

$$P\{X = i+1\} = \frac{\lambda}{i+1} P\{X = i\} \tag{25}$$

- **Geometric**:

$$P\{X = n\} = p(1-p)^{n-1} \tag{26}$$

$$E[X] = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2} \tag{27,28}$$

Biến ngẫu nhiên rời rạc

Một số phân phối xác suất phổ biến

- **Negative Binomial**:

$$P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r} \tag{29}$$

$$E[X] = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1-p)}{p^2} \tag{30, 31}$$

- **Hypergeometric**:

$$P\{X = i\} = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \tag{32}$$

$$E[X] = np, \quad \text{Var}(X) \approx np(1-p), \quad p = \frac{M}{N} \tag{33,34}$$

- **Zeta (Zipf)**:

$$P\{X = k\} = \frac{C}{k^{\alpha+1}}, \quad C = \left[\sum_{k=1}^{\infty} \frac{1}{k^{\alpha+1}}\right]^{-1} \tag{35}$$

Biến ngẫu nhiên liên tục

Khái niệm và Hàm phân phối tích luỹ

Một biến liên tục có thể nhận vô số giá trị tương ứng với các điểm trên một khoảng tuyến tính.

- Hàm phân phối tích lũy (CDF) $F_X(x)$ là hàm liên tục với mọi $x \in \mathbb{R}$.

Định lý 1: Các tính chất của hàm phân phối tích luỹ $F(x)$:

- $F_X(x)$ không giảm: nếu $a < b$ thì $F(a) \leq F(b)$
- $\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to+\infty} F_X(x) = 1$
- $F_X(x)$ liên tục bên phải: $\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$

Định lý 2: Nếu $a < b$ thì

$$P[a < X \leq b] = F_X(b) - F_X(a)$$

Định lý 3: Với $\forall x \in \mathbb{R}$ và $F_X(x^-) = \lim_{z\to x^-} F_X(z)$ thì

$$P[X = x] = F_X(x) - F_X(x^-)$$

Biến ngẫu nhiên liên tục

Tính chất phân phối xác suất

Xác suất X thuộc tập B : $$P\{X \in B\} = \int_B f(x)\,dx \tag{36}$$

Xác suất trên đoạn $[a, b]$: $$P\{a \le X \le b\} = \int_a^b f(x)\,dx \tag{37}$$

Xấp xỉ xác suất gần điểm a : $$P\left\{a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}\right\} \approx \varepsilon f(a) \tag{38}$$

Xác suất tại 1 điểm cụ thể: $$P\{X = a\} = 0 \tag{39}$$

Hàm phân phối tích lũy: $$P\{X \le a\} = F_X(a) = \int_{-\infty}^{a} f(x)\,dx \tag{40}$$

Tổng xác suất toàn trục số: $$\int_{-\infty}^{+\infty} f(x)\,dx = 1 \tag{41}$$

Kỳ vọng (trung bình) (Mean):

$$E[X] = \int_{-\infty}^{+\infty} x\, f(x)\,dx \tag{42}$$

Biến ngẫu nhiên liên tục

Một số phân phối xác suất phổ biến

1. Phân phối đều (Uniform) trên (α, β):

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha}, & \alpha < x < \beta \\ 0, & \text{khác} \end{cases}$$

$$E[X] = \frac{\beta + \alpha}{2}, \quad \mathrm{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

2. Phân phối chuẩn (Normal) $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

3. Phân phối chuẩn tắc (Standard Normal) $N(0, 1)$:

$$z = \frac{x - \mu}{\sigma}$$

Các định lý về giới hạn

Markov: Với $X \geq 0$, $a > 0$

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Chebyshev: Với $k > 0$

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

Hệ quả: Nếu $\mathrm{Var}(X) = 0 \Rightarrow P\{X = E[X]\} = 1$

Các định lý về giới hạn

Luật số lớn yếu:

$$P\left\{\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|\geq\varepsilon\right\}\to 0 \quad (n\to\infty)$$

Luật số lớn mạnh:

$$\frac{X_1+\cdots+X_n}{n}\to\mu \quad \text{(xác suất 1)}$$

Ứng dụng: Với $X_i=\mathbf{1}_{\text{E xảy ra tại lần } i}$ thì:

$$\frac{X_1+\cdots+X_n}{n}\to P(E)$$

Các định lý về giới hạn

Định lý giới hạn trung tâm (CLT):

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

Tức là:

$$P\left\{\left|\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right| \leq a\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx$$

Phiên bản tổng quát:

$$P\left\{\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \leq a\right\} \to \Phi(a)$$

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)

Định nghĩa: Cho hai biến ngẫu nhiên A và B trên cùng một không gian xác suất $(\Omega, \mathbb{P})$. Xác suất đồng thời cho biết xác suất để A và B cùng đạt hai giá trị cụ thể đồng thời.

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)-Trường hợp rời rạc

Giả sử A chỉ nhận giá trị trong tập đếm được $\mathcal{A}$, và B chỉ nhận giá trị trong tập đếm được $\mathcal{B}$.

$$p_{A,B}(a,b) \;=\; P\Big(A=a,\; B=b\Big), \quad a\in\mathcal{A},\; b\in\mathcal{B}.$$

- $p_{A,B}(a,b)\geq 0$ với mọi a,b, và $\displaystyle\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p_{A,B}(a,b)=1$.
- Hàm khối xác suất biên:

$$p_A(a) \;=\; \sum_{b\in\mathcal{B}} p_{A,B}(a,b), \quad a\in\mathcal{A};$$

$$p_B(b) \;=\; \sum_{a\in\mathcal{A}} p_{A,B}(a,b), \quad b\in\mathcal{B}.$$

- Do $\{A=a,\; B=b\}\subseteq\{A=a\}$, luôn có

$$p_{A,B}(a,b) \;=\; P(A=a,\; B=b) \;\leq\; P(A=a)=p_A(a).$$

Tương tự, $p_{A,B}(a,b)\leq p_B(b)$.

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)-Trường hợp rời rạc

Ví dụ: $A \in \{1,2,3,4,5,6\}$: số mặt khi tung xúc xắc. $B = \begin{cases} 0 & \text{nếu } A \text{ lẻ} \\ 1 & \text{nếu } A \text{ chẵn} \end{cases}$

Hàm khối xác suất chung: $p_{A,B}(a,b) = P(A=a, B=b)$

- Nếu b không khớp chẵn/lẻ: $p_{A,B}(a,b) = 0$; Nếu a lẻ: $p_{A,B}(a,0) = \frac{1}{6}$; Nếu a chẵn: $p_{A,B}(a,1) = \frac{1}{6}$

Biên:

$$p_A(a) = \frac{1}{6}, \quad p_B(0) = p_B(1) = 0.5$$

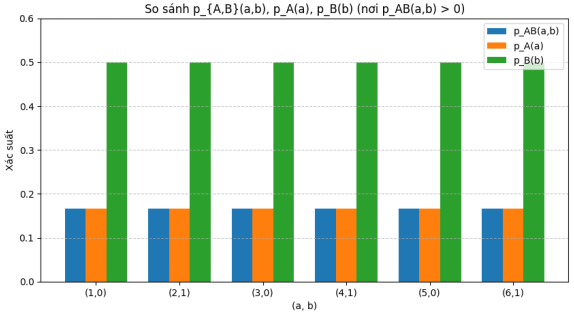
Kiểm tra:

$$\sum_{a=1}^{6} \sum_{b=0}^{1} p_{A,B}(a,b) = 1$$

Bất đẳng thức biên: $p_{A,B}(a,b) \leq p_A(a),\ p_B(b)$

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)-Trường hợp rời rạc



Hình: Biểu đồ so sánh bất đẳng thức biên

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)-Trường hợp liên tục

Giả sử A và B là hai biến ngẫu nhiên liên tục (định nghĩa trên $\mathbb{R}$). Khi đó, tồn tại hàm mật độ chung (joint probability density function) $f_{A,B}(a,b)$ sao cho với mọi miền $D \subset \mathbb{R}^2$,

$$P\Big((A,B) \in D\Big) \;=\; \iint_D f_{A,B}(a,b)\,\mathrm{d}a\,\mathrm{d}b.$$

- $f_{A,B}(a,b) \geq 0$ với mọi $(a,b) \in \mathbb{R}^2$, và $\displaystyle\iint_{\mathbb{R}^2} f_{A,B}(a,b)\,\mathrm{d}a\,\mathrm{d}b = 1$.
- Mật độ biên:

$$f_A(a) \;=\; \int_{-\infty}^{+\infty} f_{A,B}(a,b)\,\mathrm{d}b, \quad a \in \mathbb{R},$$

$$f_B(b) \;=\; \int_{-\infty}^{+\infty} f_{A,B}(a,b)\,\mathrm{d}a, \quad b \in \mathbb{R}.$$

- Mặc dù $f_{A,B}(a,b)$ không phải xác suất mà là mật độ, ta vẫn có quan hệ cận trên:

$$\int_{-\infty}^{+\infty} f_{A,B}(a,b)\,\mathrm{d}b \;=\; f_A(a) \implies f_{A,B}(a,b) \;\leq\; f_A(a), \quad f_{A,B}(a,b) \;\leq\; f_B(b).$$

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)-Trường hợp liên tục

Ví dụ: Chọn ngẫu nhiên điểm (A, B) đều trên hình chữ nhật:

$$D = \{(a, b) \in \mathbb{R}^2 \mid 0 \leq a \leq 2,\ 1 \leq b \leq 3\}$$

1. Mật độ chung:

$$f_{A,B}(a, b) = \begin{cases} \frac{1}{4}, & (a, b) \in D \\ 0, & \text{ngược lại} \end{cases}$$

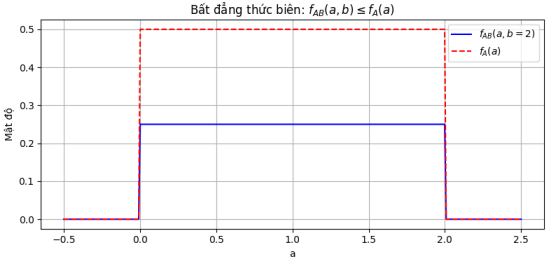
2. Mật độ biên: $f_A(a) = \int_1^3 \frac{1}{4}\, db = 0.5 \quad$ với $\quad a \in [0, 2]$;
$f_B(b) = \int_0^2 \frac{1}{4}\, da = 0.5 \quad$ với $\quad b \in [1, 3]$

3. Bất đẳng thức biên:

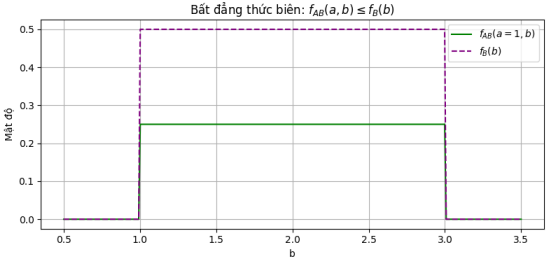
$$f_{A,B}(a, b) = 0.25 \leq f_A(a) = 0.5, \quad f_{A,B}(a, b) \leq f_B(b) = 0.5$$

Phân phối xác suất đa biến

Xác suất đồng thời (Joint Probability)-Trường hợp liên tục



Bất đẳng thức biên cho fA(a)



Bất đẳng thức biên cho fB(b)

Phân phối xác suất đa biến

Xác suất có điều kiện (Conditional Probability)

Định nghĩa:

- Giả sử $P(B) > 0$. Khi đó, xác suất của sự kiện A dưới điều kiện B được định nghĩa là

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1.18}$$

- Từ đó suy ra *công thức nhân xác suất*:

$$P(A \cap B) = P(A \mid B) \cdot P(B). \tag{1.19}$$

Ví dụ: Rút hai viên bi không trả

- Hộp có 5 viên bi: 2 viên đỏ, 3 viên xanh. Rút 2 lần, không trả lại.
  - (a) B: "Viên bi đầu tiên rút được là đỏ."
    $P(B) = \frac{2}{5}$.
  - (b) A: "Viên bi thứ hai rút được là đỏ."
    Khi đã rút đỏ ở lần 1, trong hộp còn 1 đỏ, 3 xanh nên
    $P(A \mid B) = \frac{1}{\ldots}$.

Phân phối xác suất đa biến

Xác suất có điều kiện (Conditional Probability)

Khái niệm Độc lập – Phụ thuộc

- Hai sự kiện A và B gọi là *độc lập* nếu:

$$P(A \mid B) = P(A) \quad \Longleftrightarrow \quad P(A \cap B) = P(A)\,P(B). \tag{1.20}$$

- Nếu không thỏa mãn $P(A \cap B) \neq P(A)\,P(B)$, thì A và B *phụ thuộc*.
- Nếu $P(A \mid B) > P(A)$ thì "B thuận lợi cho A"; nếu $P(A \mid B) < P(A)$ thì "B bất lợi cho A".
- Công thức đối xứng:

$$\frac{P(A \mid B)}{P(A)} = \frac{P(B \mid A)}{P(B)} \tag{1.23}$$

Phân phối xác suất đa biến

Xác suất có điều kiện (Conditional Probability)-Độc lập toàn phần

Định nghĩa 1.7 [6]:

- Tập $\mathcal{M}$ là một *họ các sự kiện độc lập toàn phần* nếu:

$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i) \quad \text{với mọi } k \text{ và mọi } A_1, \ldots, A_k \in \mathcal{M}. \tag{1.22}$$

Phân phối xác suất đa biến

Xác suất toàn phần (Total Probability)

Định nghĩa 1.8 [6]: Một họ các tập con $B_1, \ldots, B_n$ của không gian xác suất Ω là một *phân hoạch* nếu:

$$B_i \cap B_j = \varnothing \quad (i \neq j), \quad \bigcup_{i=1}^{n} B_i = \Omega. \tag{1.24}$$

- Mỗi B_i đôi một không giao nhau với các B_j khác.
- Hợp tất cả các B_i bằng toàn bộ không gian mẫu Ω.

Phân phối xác suất đa biến

Phân phối xác suất của hai biến ngẫu nhiên

Định nghĩa véc tơ ngẫu nhiên Cho một phép thử ngẫu nhiên với không gian mẫu S và hai biến ngẫu nhiên X_1, X_2. Với mỗi phần tử $s \in S$, nếu $X_1(s) = x_1$ và $X_2(s) = x_2$, thì:

$$(X_1, X_2) \text{ là một véc tơ ngẫu nhiên}$$

Không gian của (X_1, X_2):

$$D = \{(x_1, x_2) : x_1 = X_1(s),\ x_2 = X_2(s),\ s \in S\}$$

Phân phối xác suất đa biến

Phân phối xác suất của hai biến ngẫu nhiên

Hàm phân phối tích lũy (CDF):

$$F_{X_1,X_2}(x_1,x_2) = P[X_1 \leq x_1,\ X_2 \leq x_2], \quad \forall (x_1,x_2) \in \mathbb{R}^2$$

Đối với tập $(a_1,b_1] \times (a_2,b_2]$, ta có:

$$\begin{aligned} P[a_1 \leq X_1 \leq b_1,\ a_2 \leq X_2 \leq b_2] &= F_{X_1,X_2}(b_1,b_2) - F_{X_1,X_2}(a_1,b_2) \\ &\quad - F_{X_1,X_2}(b_1,a_2) + F_{X_1,X_2}(a_1,a_2) \end{aligned}$$

Trường hợp véc tơ ngẫu nhiên rời rạc:

$$p_{X_1,X_2}(x_1,x_2) = P[X_1 = x_1,\ X_2 = x_2], \quad \forall (x_1,x_2) \in S$$

Phân phối xác suất đa biến

Phân phối xác suất của hai biến ngẫu nhiên

Tham số - Kỳ vọng Giả sử $Y = g(X_1, X_2)$:

- Rời rạc:

$$E[Y] = \sum_{x_1} \sum_{x_2} g(x_1, x_2)\, p_{X_1,X_2}(x_1, x_2)$$

- Liên tục:

$$E[Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x_1, x_2)\, f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2$$

Phân phối xác suất đa biến

Phân phối xác suất của hai biến ngẫu nhiên

Hàm tạo sinh động lượng Với $X = (X_1, X_2)'$, nếu tồn tại kỳ vọng:

$$M_{X_1,X_2}(t) = E\left[e^{t'X}\right], \quad \text{với } t = (t_1, t_2)'$$

$$E[X] = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix}$$

Phân phối xác suất đa biến

Phân phối xác suất của hai biến ngẫu nhiên

Định lý về kỳ vọng Cho $Y_1 = g_1(X_1, X_2)$, $Y_2 = g_2(X_1, X_2)$ và $k_1, k_2 \in \mathbb{R}$, ta có:

$$E[k_1 Y_1 + k_2 Y_2] = k_1 E(Y_1) + k_2 E(Y_2)$$

Nếu $\mathrm{Var}(X_2)$ hữu hạn:

$$E[E(X_2 \mid X_1)] = E(X_2)$$
$$\mathrm{Var}[E(X_2 \mid X_1)] \leq \mathrm{Var}(X_2)$$

Phân phối xác suất đa biến

Phân phối xác suất của hai biến ngẫu nhiên

Hiệp phương sai, hệ số tương quan và Định lý hồi quy Cho hai biến ngẫu nhiên X và Y có giá trị trung bình μ_1, μ_2 và phương sai σ_1^2, σ_2^2:

$$\mathrm{Cov}(X, Y) = E[(X - \mu_1)(Y - \mu_2)] = E(XY) - \mu_1\mu_2$$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_1\sigma_2}, \quad E(XY) = \mu_1\mu_2 + \mathrm{Cov}(X, Y)$$

Nếu $E(Y \mid X)$ tuyến tính theo X, thì:

$$E(Y \mid X) = \mu_2 + \mathrm{Corr}(X, Y) \cdot \frac{\sigma_2}{\sigma_1}(X - \mu_1)$$

$$E[\mathrm{Var}(Y \mid X)] = \sigma_2^2\Big[1 - (\mathrm{Corr}(X, Y))^2\Big]$$

Ứng dụng

Ngành khoa học máy tính

1. **Thuật toán ngẫu nhiên**:
   - **Randomized QuickSort**: Chọn pivot ngẫu nhiên.
   - **Bloom Filter**: Kiểm tra phần tử với xác suất sai lệch nhỏ.
2. **Machine Learning**:
   - **Naive Bayes**: Phân loại dựa vào định lý Bayes.
   - **Bayesian Networks**: Biểu diễn phụ thuộc giữa biến ngẫu nhiên.
3. **Mô phỏng Monte Carlo**:
   - Gần đúng tích phân/phân tích hệ phức tạp.
   - Ứng dụng trong đồ họa và tài chính.
4. **Lý thuyết thông tin**:
   - **Entropy**: Đo độ bất định.
   - **Huffman Coding**: Mã hóa tối ưu.

Một ví dụ về xét nghiệm HIV AIDS

Giả sử rằng một bác sĩ phụ trách xét nghiệm AIDS cho một bệnh nhân. Việc xét nghiệm này khá chính xác và nó chỉ thất bại với xác suất 1%, khi việc xét nghiệm cho kết quả dương tính dù bệnh nhân khỏe mạnh. Hơn nữa, việc xét nghiệm không bao giờ thất bại trong việc phát hiện HIV nếu bệnh nhân thực sự bị nhiễm bệnh. Ta sử dụng D_1 để biểu diễn kết quả chẩn đoán (1 nếu dương tính và 0 nếu âm tính) và H để biểu thị tình trạng nhiễm HIV (1 nếu dương tính và 0 nếu âm tính).

H: tình trạng nhiễm HIV

H=1 nếu bệnh nhân thực sự nhiễm HIV

H=0 nếu bệnh nhân không nhiễm HIV

D_1: kết quả xét nghiệm

D_1=1 nếu test dương tính

D_1=0 nếu test âm tính

Bảng xác suất có điều kiện

Ta có liệt kê xác suất có điều kiện theo bảng dưới đây: **Xác suất có điều kiện của** $P(D_1 \mid H)$

Xác suất có điều kiện	$H = 1$	$H = 0$
$P(D_1 = 1 \mid H)$	1	0.01
$P(D_1 = 0 \mid H)$	0	0.99

Ta có thể nhận thấy rằng tổng của từng cột đều bằng 1 (nhưng tổng từng hàng thì không), vì xác suất có điều kiện cần có tổng bằng 1.

Xác định biến ngẫu nhiên và các xác suất cho trước

Ta có: $P(D_1 = 1 \mid H = 1) = 1$ (xét nghiệm luôn phát hiện đúng khi có bệnh theo dữ kiện đã cho) $P(D_1 = 1 \mid H = 0) = 0.01$ (tỉ lệ dương tính giả $= 1\%$)
Hãy cùng tìm xác suất bệnh nhân bị AIDS nếu xét nghiệm trả về kết quả dương tính, tức là: $P(H = 1 \mid D = 1)$.
Giả sử rằng dân số khá khỏe mạnh,tỷ lệ nhiễm HIV trong dân số là: $P(H = 1) = 0.0015$

Áp dụng quy tắc biên hóa và công thức nhân xác suất

Để áp dụng định lý Bayes, trước hết ta cần áp dụng quy tắc biên hóa và công thức nhân xác suất có điều kiện để tính:

$$P(D_1 = 1) = P(D_1 = 1, H = 0) + P(D_1 = 1, H = 1)$$

$$= P(D_1 = 1, H = 0)P(H = 0) + P(D_1 = 1, H = 1)P(H = 1)$$

Thay giá trị vào công thức, ta có:
$0.01 * (1 - 0.0015) + 1 * 0.0015 = 0.00985 + 0.0015 = 0.011485$

Áp dụng định lý Bayes

Theo định lý Bayes: $P(H = 1 \mid D_1 = 1) = \frac{P(D_1=1|H=1)P(H=1)}{P(D_1=1)}$
Thay giá trị vào công thức, ta có:

$$\frac{1 * 0.0015}{0.011485} = 0.1306051$$

Nói cách khác, chỉ có xấp xi 13,06% khả năng bệnh nhân thực sự mắc bệnh AIDS, dù ta dùng một bài kiểm tra rất chính xác. Như ta có thể thấy, xác suất có thể trở nên khá phản trực giác do trước khi xét nghiệm, trực giác chỉ ra có 0.15% khả năng nhiễm bệnh.

Xét nghiệm lần thứ hai

Một bệnh nhân phải làm gì nếu nhận được tin dữ như vậy? Nhiều khả năng họ sẽ yêu cầu bác sĩ thực hiện một xét nghiệm khác để làm rõ sự việc. Giả sử rằng bài kiểm tra thứ hai có những đặc điểm khác và không tốt bằng bài thứ nhất, như ta có thể thấy trong bảng sau: **Xác suất có điều kiện của $P(D_2 \mid H)$.**

Xác suất có điều kiện	$H = 1$	$H = 0$
$P(D_2 = 1 \mid H)$	0.98	0.03
$P(D_2 = 0 \mid H)$	0.02	0.97

Tính các xác suất cần thiết

Không may thay, bài kiểm tra thứ hai cũng có kết quả dương tính. Hãy cùng tính các xác suất cần thiết để sử dụng định lý Bayes bằng cách giả định tính độc lập có điều kiện:

$$\begin{aligned} P(D_1 = 1, D_2 = 1 \mid H = 0) &= P(D_1 = 1 \mid H = 0)P(D_2 = 1 \mid H = 0) \\ &= 0.01 \times 0.03 \\ &= 0.0003, \end{aligned}$$

$$\begin{aligned} P(D_1 = 1, D_2 = 1 \mid H = 1) &= P(D_1 = 1 \mid H = 1)P(D_2 = 1 \mid H = 1) \\ &= 1 \times 0.98 \\ &= 0.98. \end{aligned}$$

Các bước tính toán

Bây giờ chúng ta có thể áp dụng phép biên hóa và quy tắc nhân xác suất:

$$\begin{aligned} P(D_1 = 1, D_2 = 1) &= P(D_1 = 1, D_2 = 1, H = 0) + P(D_1 = 1, D_2 = 1, H = 1) \\ &= P(D_1 = 1, D_2 = 1 \mid H = 0)P(H = 0) + P(D_1 = 1, D_2 = 1 \mid H = 1)P(H = \\ &= 0.0003 \times 0.9985 + 0.98 \times 0.0015 \\ &= 0.00176955 \end{aligned}$$

Cuối cùng xác suất bệnh nhân mắc bệnh AIDS qua hai lần dương tính là

$$\begin{aligned} P(H = 1 \mid D_1 = 1, D_2 = 1) = \frac{P(D_1 = 1, D_2 = 1 \mid H = 1)P(H = 1)}{P(D_1 = 1, D_2 = 1)} &= \frac{0.98 \times 0.0015}{0.00176955} \\ \approx 0.8307. \end{aligned}$$

Nhận xét

Xét nghiệm thứ hai đã cho thấy được mức độ tin cậy cao hơn nhiều rằng có điều gì đó không ổn. Mặc dù xét nghiệm thứ hai kém chính xác hơn đáng kể so với xét nghiệm đầu tiên, nhưng nó vẫn cải thiện đáng kể ước lượng của chúng ta.
Giả định rằng hai xét nghiệm độc lập có điều kiện với nhau là yếu tố then chốt giúp đưa ra ước lượng chính xác hơn. Hãy xét một trường hợp cực đoan khi thực hiện cùng một xét nghiệm hai lần. Trong tình huống này, ta kỳ vọng kết quả sẽ giống nhau ở cả hai lần, do đó không có thông tin mới nào được rút ra từ việc lặp lại cùng một xét nghiệm.
Chúng ta có thể nhận ra rằng quá trình chẩn đoán hoạt động giống như một bộ phân loại (classifier) đang "ẩn mình", khả năng của chúng ta trong việc xác định bệnh nhân có khỏe mạnh hay không sẽ tăng lên khi chúng ta thu thập thêm các đặc trưng (kết quả xét nghiệm).

Hàm mật độ xác suất

Hàm mật độ xác suất (Probability Density Function hay PDF) dùng để biểu diễn một phân bố xác suất theo tích phân. Gọi $p(x)$ là một hàm mật độ xác suất, vì xác suất không bao giờ âm, do đó: $p(x) \geq 0$
Hơn nữa, Ta có:

$$P(X \in \mathbb{R}) = 1$$

và

$$P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} p(x)\, dx.$$

Vì biến ngẫu nhiên này phải nhận một giá trị nào đó trong tập số thực, do đó ta có thể kết luận rằng với bất kỳ hàm mật độ nào thì:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

Hàm mật độ xác suất

Đi sâu hơn vào phương trình trên, ta thấy rằng với bất kỳ a và b nào thì:

$$P(X \in (a, b]) = \int_a^b p(x)\, dx. \tag{1.5.1}$$

Công thức này dùng để tính xác suất biến ngẫu nhiên nằm trong một khoảng cụ thể.

Hàm Phân phối Tích lũy

Trong phần trước, nhóm đã trình bày về hàm mật độ xác suất (PDF). Trong thực tế, đây là một phương pháp thường dùng để thảo luận về các biến ngẫu nhiên liên tục, nhưng nó có một nhược điểm khá lớn: bản thân các giá trị của PDF không phải là các giá trị xác suất, mà ta phải tích phân hàm này để có xác suất. Không có gì sai với môt hàm mật độ lớn hơn 10, miễn là nó không lớn hơn 10 trong khoảng có chiều dài lớn hơn 1/10. Điều này có thể hơi phản trực giác, do đó người ta thường dùng hàm phân phối tích lũy - Cumulative Distribution Function hay CDF, mà có giá trị trả về là xác suất.

Hàm Phân phối Tích lũy

Hàm phân phối tích lũy mô tả đầy đủ phân phối xác suất của một biến ngẫu nhiên giá trị thực X. Với mỗi số thực x, hàm phân phối tích lũy được định nghĩa như sau:

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

Với một hàm phân phối tích lũy $F(x)$ tương ứng với một hàm mật độ xác suất $p(x)$ được định nghĩa là:

$$F(x) = \int_{-\infty}^{x} p(x)\, dx.$$

Với việc sử dụng công thức (1.5.1) ở trên, ta định nghĩa CDF cho một biến ngẫu nhiên X với mật độ $p(x)$ như sau:

$$F(x) = \int_{-\infty}^{x} p(x)\, dx = P(X \leq x).$$

Hàm Phân phối Tích lũy

Một vài tính chất của hàm phân phối tích lũy:

- $F(x) \to 0$ khi $x \to -\infty$.
- $F(x) \to 1$ khi $x \to \infty$.
- $F(x)$ không giảm ($y > x \Longrightarrow F(y) \geq F(x)$).
- $F(x)$ liên tục (không có bước nhảy) nếu X là một biến ngẫu nhiên liên tục.

Kỳ vọng là gì

Thông thường, việc ra quyết định không chỉ yêu cầu xem xét các xác suất được gán cho từng sự kiện riêng lẻ mà còn cần tổng hợp chúng thành các đại lượng hữu ích có thể cung cấp cho chúng ta sự chỉ dẫn. Chẳng hạn, khi các biến ngẫu nhiên nhận giá trị liên tục, chúng ta thường quan tâm đến việc biết được giá trị kỳ vọng trung bình là bao nhiêu. Đại lượng này được gọi một cách chính thức là **kỳ vọng (expectation)**. **Kỳ vọng** là giá trị trung bình của một biến ngẫu nhiên

Ví dụ về kỳ vọng

Ví dụ 5.1 về bài toán lợi nhuận kỳ vọng khi đầu tư: Nếu chúng ta đang đầu tư, điều đầu tiên cần quan tâm có thể là lợi nhuận kỳ vọng – trung bình cộng tất cả các kết quả có thể xảy ra (và được cân nhắc theo xác suất tương ứng).
Giả sử rằng với 50% xác suất, một khoản đầu tư có thể thất bại hoàn toàn, với 40% xác suất nó có thể mang lại lợi nhuận gấp 2 lần, và với 10% xác suất nó có thể mang lại lợi nhuận gấp 10 lần. Để tính lợi nhuận kỳ vọng, ta cộng tất cả các mức lợi nhuận lại, mỗi mức được nhân với xác suất xảy ra của nó. Điều này dẫn đến kỳ vọng là: $0.5 * 0 + 0.4 * 2 + 0.1 * 10 = 1.8$ Vậy nên, lợi nhuận kỳ vọng là 1.8 lần.

Công thức kỳ vọng của biến ngẫu nhiên rời rạc X

Kỳ vọng (hay trung bình) của biến ngẫu nhiên rời rạc X được định nghĩa theo công thức:

$$E[X] = E_{x\sim P}[x] = \sum_x xP(X = x)$$

Giải thích: Đây là giá trị trung bình kỳ vọng của biến ngẫu nhiên rời rạc X. Mỗi giá trị có thể xảy ra của X được nhân với xác suất xảy ra của chính nó. Sau đó, các tích này được cộng lại để tính kỳ vọng.
Chẳng hạn: Nếu một đồng xu có 50% ra sấp (giá trị 0) và 50% ra ngửa (giá trị 1), thì:
$E[X] = 0 * 0.5 + 1 * 0.5 = 0.5 =>$ Trung bình kỳ vọng là 0.5.

Công thức kỳ vọng của biến ngẫu nhiên liên tục X

Kỳ vọng (hay trung bình) của biến ngẫu nhiên liên tục X được định nghĩa theo công thức:

$$E[X] = \int x \cdot p(x)\, dx$$

Giải thích: Khi biến ngẫu nhiên X có phân phối liên tục, ta không thể dùng tổng như trên. Thay vào đó, ta dùng tích phân của x nhân với hàm mật độ xác suất $p(x)$.

Công thức kỳ vọng của một hàm số $f(x)$

Khi quan tâm đến kỳ vọng của một hàm số $f(x)$, ta có:

- Công thức tính kỳ vọng theo phân phối rời rạc:

$$E_{x\sim P}[f(x)] = \sum_x f(x)P(x),$$

- Công thức tính kỳ vọng theo phân phối liên tục:

$$E_{x\sim P}[f(x)] = \int f(x)p(x)dx$$

Trở lại ví dụ 5.1 về đầu tư, nếu độ hài lòng với mất trắng là -1, và các độ hài lòng tương ứng với các mức lợi nhuận 1, 2 và 10 lần là 1, 2 và 4 thì lợi nhuận kỳ vọng sẽ là:

$$0.5 \cdot (-1) + 0.4 \cdot 2 + 0.1 \cdot 4 = 0.7$$

Nếu thực sự đây là hàm ích lợi (utility) của ta, tốt nhất ta nên giữ tiền trong ngân hàng, không nên đầu tư.

Phương sai

Trong các quyết định tài chính, ta không chỉ quan tâm đến kỳ vọng mà còn đến mức độ *dao động* của các kết quả quanh kỳ vọng đó. Lưu ý rằng ta không thể chỉ lấy kỳ vọng của hiệu giữa giá trị thực và giá trị kỳ vọng: $[E[X - E[X]]$ Vì:

$$E[X - E[X]] = E[X] - E[E[X]] = 0$$

Tuy nhiên, ta có thể xét kỳ vọng của một hàm không âm bất kỳ của phần chênh lệch này. Phương sai (variance) của một biến ngẫu nhiên được tính bằng cách lấy kỳ vọng của bình phương hiệu:

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

Phương sai

Phương sai của một hàm của một biến ngẫu nhiên được định nghĩa tương tự:

$$\text{Var}_{x\sim P}[f(x)] = E_{x\sim P}[f^2(x)] - E_{x\sim P}[f(x)]^2$$

Ví dụ: quay trở lại ví dụ 5.1 về đầu tư ở trên, ta có thể tính phương sai khoản đầu tư như sau:

$$0.5 \cdot 0 + 0.4 \cdot 2^2 + 0.1 \cdot 10^2 - 1.8^2 = 8.36$$

Theo quy ước, kỳ vọng và phương sai được ký hiệu lần lượt là μ và σ^2.

Phương sai

Một vài tính chất của phương sai:

- Với biến ngẫu nhiên X bất kỳ: $\mathrm{Var}(X) \geq 0$, với $\mathrm{Var}(X) = 0$ khi và chỉ khi X là hằng số.
- Với biến ngẫu nhiên X và hai số a, b bất kỳ: $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- Nếu hai biến ngẫu nhiên X và Y là *độc lập*: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Độ lệch chuẩn

Độ lệch chuẩn là căn bậc hai của phương sai, giúp diễn giải dễ hơn vì cùng đơn vị với biến gốc.

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Các tính chất của phương sai có thể được áp dụng lại cho độ lệch chuẩn:

- Với biến ngẫu nhiên X bất kỳ: $\sigma_X \geq 0$.
- Với biến ngẫu nhiên X và hằng số a, b bất kỳ: $\sigma_{aX+b} = |a|\sigma_X$
- Nếu hai biến ngẫu nhiên X và Y là độc lập: $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

Hiệp phương sai

Khi làm việc với nhiều biến ngẫu nhiên, còn có một thông số thống kê nữa rất có ích: ***hiệp phương sai (covariance)***. Thông số này đo mức độ biến thiên cùng nhau của hai biến ngẫu nhiên. Để bắt đầu, giả sử ta có hai biến ngẫu nhiên rời rạc X và Y, xác suất mang giá trị (x_i, y_j) là p_{ij}. Trong trường hợp này, hiệp phương sai được định nghĩa như sau:

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \sum_{i,j}(x_i - \mu_X)(y_j - \mu_Y)p_{ij} = E[XY] - E[X]E[Y].$$

Hiệp phương sai

Với biến ngẫu nhiên liên tục, khái niệm hiệp phương sai không đổi. Khi đó:

$$\sigma_{XY} = \int_{\mathbb{R}^2} (x - \mu_X)(y - \mu_Y) p(x, y)\, dx\, dy.$$

Tính chất của hiệp phương sai:

- Với biến ngẫu nhiên X bất kỳ: $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- Với hai biến ngẫu nhiên X, Y và hai số a, b bất kỳ: $\mathrm{Cov}(aX + b, Y) = \mathrm{Cov}(X, aY + b) = a\mathrm{Cov}(X, Y)$.
- Nếu X và Y độc lập: $\mathrm{Cov}(X, Y) = 0$.

Bài tập số 1 I

Lấy một ví dụ cho thấy việc quan sát thêm dữ liệu hoặc tăng kích thước tập huấn luyện có thể làm giảm mức độ không chắc chắn (uncertainty) về kết quả, tới mức tùy ý nhỏ.

Lời giải bài tập 1

Hiện tượng bất định hay không chắc chắn trong Machine learning có liên quan nhau "Entropy càng thấp, sự chắc chắn về kết quả càng cao". Hiện tượng bất định liên quan đến khái niệm entropy trong lý thuyết xác suất, entropy dùng để đo lường mức độ không chắc chắn của một biến ngẫu nhiên. Xác suất cao lượng thông tin thu được càng giảm và ngược lại.

Định nghĩa Entropy: ký hiệu $H(X)$) là đại lượng trong lý thuyết thông tin dùng xác suất để tính toán, đo lường mức độ không chắc chắn hoặc lượng thông tin trung bình của một biến ngẫu nhiên X. Entropy càng thấp thì càng dễ đoán trước giá trị của biến ngẫu nhiên đó.

- Công thức tính Entropy (cho biến rời rạc):

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

Trong đó:

Bài tập số 1 II

- $p(x)$: Xác suất xảy ra sự kiện x.
- $\mathcal{X}$: Tập hợp tất cả các kết quả có thể.

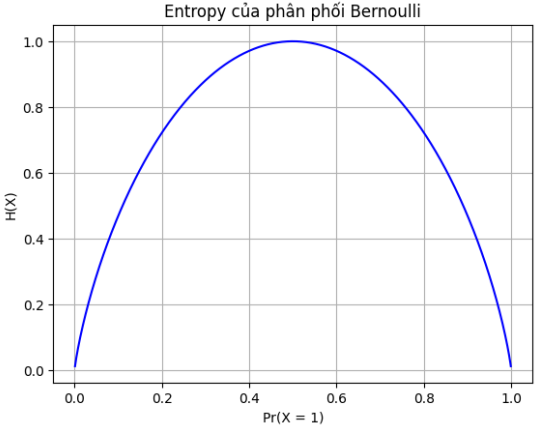
- Entropy đo lường sự không chắc chắn
  - Khi entropy cao:
    - Xác suất của các kết quả gần bằng nhau (ví dụ: đồng xu cân bằng, $p = 0.5$).
    - Khó dự đoán kết quả nhất vì mọi khả năng đều có khả năng xảy ra tương đương.
    - Ví dụ: Tung một đồng xu cân bằng (p = 0.5), entropy là 1 bit nếu dùng log cơ số 2 (không chắc chắn tối đa).
  - Khi entropy thấp (Entropy gần bằng 0):
    - Một kết quả có xác suất rất cao, các kết quả khác rất thấp (ví dụ: đồng xu lệch với $p = 0.99$).
    - Dễ dự đoán kết quả hơn vì một sự kiện gần như chắc chắn xảy ra.
    - Ví dụ: Đồng xu với $p = 0.99$, entropy $H(X) \approx 0.08$ bit (rất không chắc chắn).
  - Trường hợp xấu: Nếu một sự kiện có xác suất $p = 1$ (chắc chắn xảy ra), entropy $H(X) = 0 \Rightarrow$ Không có không chắc chắn nào cả. - Mối quan hệ giữa entropy và dữ liệu
    - Quan sát thêm dữ liệu giúp cải thiện ước lượng xác suất $p(x)$, từ đó giảm entropy.

Bài tập số 1 III

- Ví dụ:
  - Ban đầu: Không biết tỷ lệ mặt ngửa của đồng xu (p có thể là bất kỳ giá trị nào trong $[0, 1]$), entropy cao.
  - Sau khi tung đồng xu 1000 lần và thấy 950 lần ngửa: Ước lượng $p \approx 0.95$, entropy giảm mạnh (gần 0) $\Rightarrow$ Sự không chắc chắn về kết quả của lần tung thứ 1001 gần như biến mất.

- Giải thích bằng hình ảnh Đồ thị entropy của biến Bernoulli, ví dụ đồng xu với xác suất mặt ngửa p:

Bài tập số 1 IV



- Entropy đạt tối đa khi $p = 0.5$ (không chắc chắn nhất).
- Entropy (với log cơ số 2) tiến về 0 khi $p \to 0$ hoặc $p \to 1$ (không chắc chắn giảm).

- Ứng dụng trong giải bài tập 1:

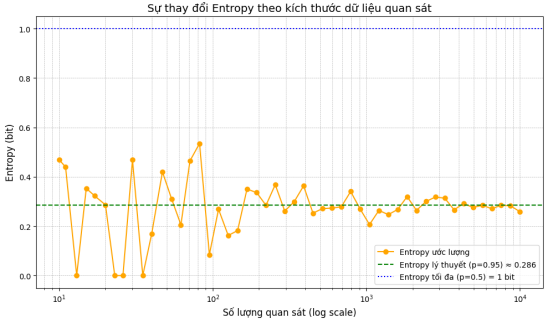
Bài tập số 1 V

- Ví dụ phù hợp: Ước lượng tham số p của phân phối Bernoulli (như tung đồng xu).
- Khi số lần quan sát $n \to \infty$, ước lượng $\hat{p}$ hội tụ về p thực, entropy của biến ngẫu nhiên giảm về 0 (nếu p là 0 hoặc 1) hoặc một giá trị nhỏ (nếu p gần $0/1$) $\Rightarrow$ Không chắc chắn gần như mất hẳn" hoặc "Entropy tiến tới giá trị tối thiểu là 0.
  - Tóm tắt:
    - Entropy thấp = Phân phối xác suất tập trung vào một vài kết quả $\to$ Dễ dự đoán $\to$ Không chắc chắn giảm.
    - Entropy cao = Phân phối đều giữa nhiều kết quả $\to$ Khó dự đoán $\to$ Không chắc chắn tăng.
    - Thêm dữ liệu giúp "làm rõ"phân phối thực $\to$ Giảm entropy.

Các kết quả chạy mô phỏng bải toán

- Mô phỏng và trực quan hóa mối quan hệ giữa kích thước tập dữ liệu quan sát và entropy, theo: Đường entropy ước lượng từ dữ liệu; Đường entropy lý thuyết cho $p = 0.95$; Đường entropy tối đa khi $p = 0.5$ để đối chiếu.

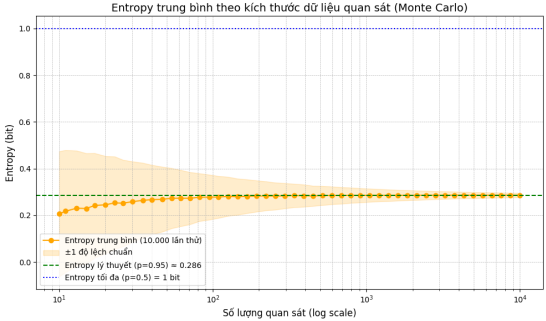
Bài tập số 1 VI



Hình: Trực quan hóa mối quan hệ giữa kích thước tập dữ liệu quan sát và entropy

Bài tập số 1 VII

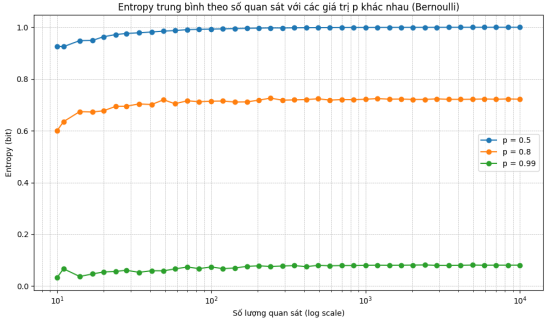
- Trung bình 10.000 lần thử (Monte Carlo Averaging)



Hình: Monte Carlo Averaging với 10.000 lần thử

Bài tập số 1 VIII

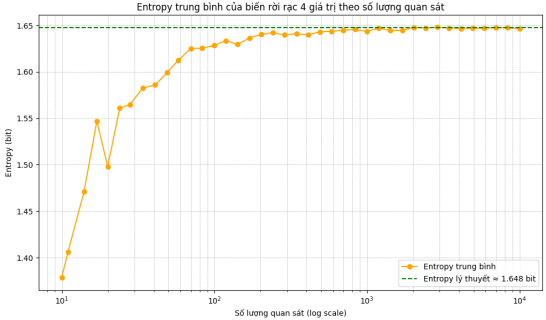
- So sánh nhiều giá trị p trong phân phối Bernoulli



Hình: So sánh nhiều giá trị p trong phân phối Bernoulli

Bài tập số 1 IX

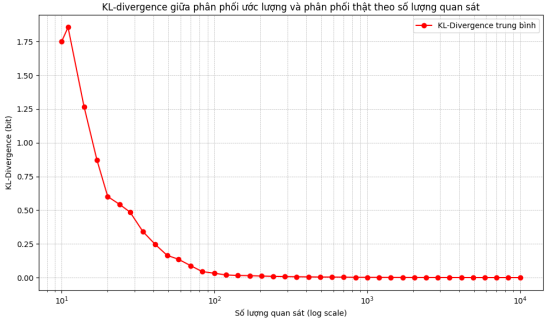
- Mô phỏng entropy cho biến rời rạc nhiều giá trị



Hình: Mô phỏng entropy cho biến rời rạc nhiều giá trị

Bài tập số 1 X

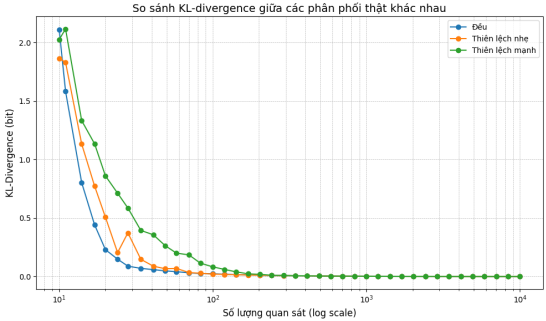
- KL-divergence giữa phân phối ước lượng và thật



Hình: KL-divergence giữa phân phối ước lượng và thật

Bài tập số 1 XI

- So sánh KL-divergence giữa nhiều phân phối thật khác nhau.



Hình: So sánh KL-divergence giữa nhiều phân phối thật khác nhau

Bài tập số 2 I

Đưa ra một ví dụ về việc quan sát thêm dữ liệu sẽ chỉ giảm lượng không chắc chắn đến một điểm nhất định và sau đó không giảm thêm nữa. Giải thích tại sao lại như vậy và bạn mong đợi điểm này sẽ xảy ra ở đâu.

Lời giải bài tập 2

Ví dụ minh họa: Đo chiều dài bằng thước có độ phân giải hữu hạn.

Bài tập số 2 II

Tình huống: Đo chiều dài thanh kim loại bằng thước có vạch chia đến 1mm. Tiến hành 500 phép đo để giảm sai số.

Hiện tượng:

- Ban đầu: Sai số ngẫu nhiên (tay đo, góc nhìn...) giảm khi lấy trung bình.
- Nhưng: Thước chỉ chia đến 1mm nên không phân biệt được các giá trị như 153.3mm hay 153.7mm.
- Do đó, kết quả bị làm tròn $\rightarrow$ Thêm phép đo không giúp giảm bất định nữa.

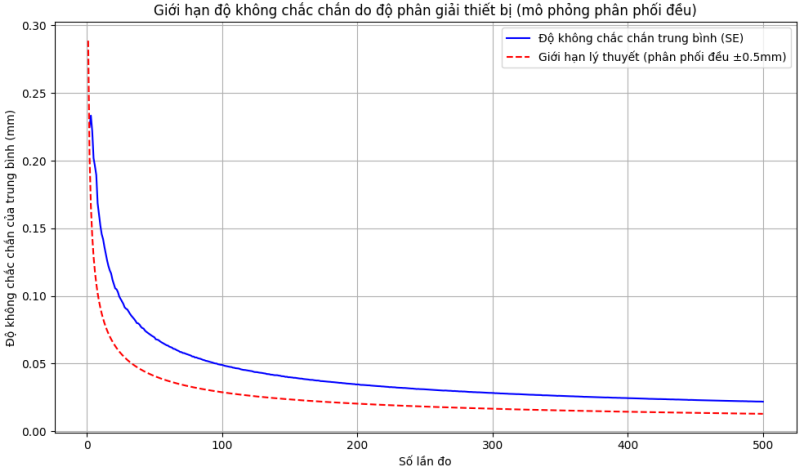
Nguyên nhân:

- **Sai số ngẫu nhiên:** Giảm được nhờ trung bình nhiều phép đo.
- **Sai số hệ thống (độ phân giải thiết bị):** Không thể giảm nếu không cải thiện thiết bị.

Khi nào xảy ra "điểm bão hòa"?

- Khi số phép đo đủ lớn để trung bình hóa hết sai số ngẫu nhiên.
- Sai số còn lại chủ yếu do giới hạn hệ thống $\rightarrow$ Không thể giảm bằng thêm dữ liệu.

Bài tập số 2 III



Hình: Mô phỏng ví dụ về giới hạn không chắc chắn quan sát được.

Bài tập số 2 IV

Kết luận từ biểu đồ:

- Ban đầu, việc tăng số lượng đo giúp giảm độ không chắc chắn rõ rệt.
- Nhưng sau một mức nhất định (khoảng 100 lần đo), đường cong tiệm cận về một giới hạn không thể vượt qua – chính là sai số hệ thống do độ phân giải thước đo.
- Quan sát thêm không làm giảm thêm độ không chắc chắn.

Bài tập số 3 I

Kết quả thực nghiệm cho thấy sự hội tụ về giá trị trung bình khi tung đồng xu. Hãy tính phương sai của ước lượng xác suất xuất hiện mặt ngửa sau khi tung đồng xu n lần.

1. Phương sai thay đổi thế nào khi số lần quan sát n tăng lên?
2. Sử dụng bất đẳng thức Chebyshev để chặn độ lệch so với kỳ vọng.
3. Mối liên hệ với định lý giới hạn trung tâm (Central Limit Theorem - CLT)?

Lời giải bài tập 3

1. Tính phương sai và Phương sai thay đổi thế nào khi số lần quan sát n tăng lên?
   - Ước lượng xác suất (tần suất):
     - Gọi $X_1, X_2, \ldots, X_n$ là đại diện cho kết quả n lần tung đồng xu:
     - $X_i = 1$ nếu ra mặt ngửa (head).
     - $X_i = 0$ nếu ra mặt sấp (tail).
     - Xác suất của mặt ngửa là p, tứclà

$$P(X_i = 1) = p \tag{1}$$

Bài tập số 3 II

- Ước lượng $\tilde{p}_n$ của p sau n lần tung là:

$$\tilde{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

- Phương sai
  - Tính $\mathrm{Var}(X_i)$:

$$\mathrm{Var}(X_i) = E[X_i^2] - (E[X_i])^2 = p - p^2 = p(1-p) \tag{3}$$

  - Do các X_i độc lập và cùng phân phối (mỗi X_i là biến nhị phân Bernoulli) nên:

$$\mathrm{Var}(\tilde{p}_n) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n\cdot \mathrm{Var}(X_i) = \frac{\mathrm{Var}(X_i)}{n} = \frac{p(1-p)}{n} \tag{4}$$

- Phương sai tỷ lệ nghịch với n. Khi n tăng, phương sai giảm với tốc độ $\frac{1}{n}$.
- Ví dụ: Nếu n tăng gấp 5 lần, phương sai giảm còn $\frac{1}{5}$.

2. Bất đẳng thức Chebyshev

Bài tập số 3 III

- Áp dụng bất đẳng thức Chebyshev cho biến ngẫu nhiên $\tilde{p}_n$, với kỳ vọng $E[\tilde{p}_n] = p$:

$$P(|p - \tilde{p}_n)| \geq \epsilon) \leq \frac{\text{Var}(\tilde{p}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \tag{5}$$

- Điều này chứng tỏ xác suất $E(p)$ lệch khỏi p một khoảng ϵ bị chặn bởi $\frac{p(1-p)}{n\epsilon^2}$.
- Ý nghĩa là khi n lớn, xác suất này tiến về 0, $E(p) \approx p$.

3. Liên hệ với định lý giới hạn trung tâm (CLT):
   - Áp dụng được CLT vì X_i độc lập và phân phối Bernoulli giống nhau.
   - Định lý giới hạn trung tâm phát biểu với n đủ lớn:

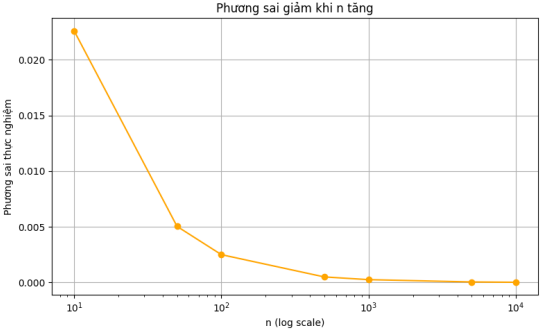
$$\sqrt{n}(\tilde{p}_n - p) \xrightarrow{d} N(0, p(1-p)) \tag{6}$$

   - Tức là $\tilde{p}_n$ có phân phối xấp xỉ chuẩn với kỳ vọng p và phương sai $\frac{p(1-p)}{n}$.
   - CLT cho thấy sự hội tụ phân phối của $\tilde{p}_n$ về phân phối chuẩn, mạnh hơn so với chỉ sử dụng phương sai hoặc Chebyshev.

Các kết quả chạy mô phỏng, tính toán và trực quan hoán kết quả của bài tập 3

Bài tập số 3 IV

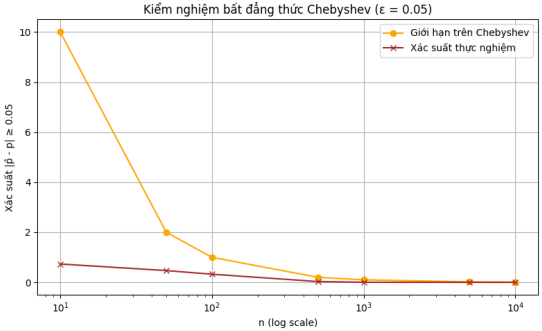
- Phương sai giảm khi n tăng



Hình: Phương sai giảm khi n tăng

- Kiểm nghiệm bất đẳng thức Chebyshev (ε)

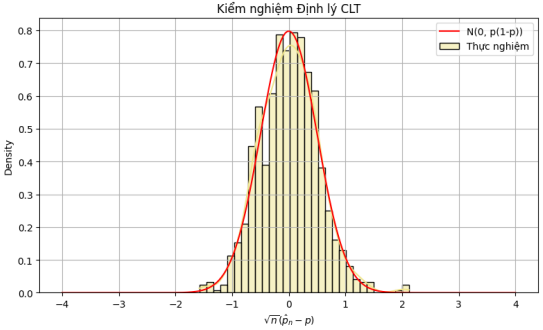
Bài tập số 3 V



Hình: Kiểm nghiệm bất đẳng thức Chebyshev

- Kiểm nghiệm Định lý CLT

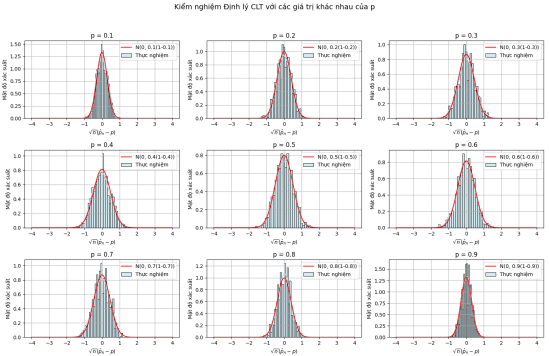
Bài tập số 3 VI



Hình: Kiểm nghiệm Định lý CLT

- Kiểm nghiệm Định lý CLT với các giá trị khác nhau của p

Bài tập số 3 VII



Hình: Kiểm nghiệm Định lý CLT với các giá trị khác nhau của p

Bài tập số 4 I

(phần 2.6.8) Assume that we draw m samples x_i from a probability distribution with zero mean and unit variance. Compute the averages $z_m \stackrel{\text{def}}{=} m^{-1} \sum_{i=1}^{m} x_i$. Can we apply Chebyshev's inequality for every z_m independently? Why not?

Tóm tắt:

- Cho m mẫu $x_1, x_2, \ldots, x_m$ được lấy từ một phân phối xác suất có:
  - Kỳ vọng $E[x_i] = 0$ (trung bình bằng 0).
  - Phương sai $\text{Var}(x_i) = 1$ (phương sai đơn vị).
  - Trung bình mẫu:

$$z_m = \frac{1}{m} \sum_{i=1}^{m} x_i$$

- Question: Có thể áp dụng bất đẳng thức Chebyshev cho từng z_m một cách độc lập không? Tại sao?

Tóm tắt:

- Cho m mẫu $x_1, x_2, \ldots, x_m$ được lấy từ một phân phối xác suất có:

Bài tập số 4 II

- Kỳ vọng $E[x_i] = 0$ (trung bình bằng 0).
- Phương sai $\text{Var}(x_i) = 1$ (phương sai đơn vị).
- Trung bình mẫu:

$$z_m = \frac{1}{m} \sum_{i=1}^{m} x_i$$

- Question: Có thể áp dụng bất đẳng thức Chebyshev cho từng z_m một cách độc lập không? Tại sao?

Tóm tắt:

- Cho m mẫu $x_1, x_2, \ldots, x_m$ được lấy từ một phân phối xác suất có:
  - Kỳ vọng $E[x_i] = 0$ (trung bình bằng 0).
  - Phương sai $\text{Var}(x_i) = 1$ (phương sai đơn vị).
  - Trung bình mẫu:

$$z_m = \frac{1}{m} \sum_{i=1}^{m} x_i$$

Bài tập số 4 III

- Question: Có thể áp dụng bất đẳng thức Chebyshev cho từng z_m một cách độc lập không? Tại sao?

Bài tập số 4 I

Lời giải bài tập 4

Ta cần tính kỳ vọng và phương sai của z_m.

Vì các x_i có kỳ vọng $E[x_i] = 0$ và phương sai $\text{Var}(x_i) = 1$, và các mẫu là độc lập, ta có:

- Kỳ vọng:

Chúng ta áp dụng tính chất tuyến tính của kỳ vọng, đó là:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Cho nên:

$$\mathbb{E}[z_m] = \mathbb{E}\left[\sum_{i=1}^{m} x_i\right] = \sum_{i=1}^{m} \mathbb{E}[x_i] = \frac{1}{m} \cdot (0 + 0 + \cdots + 0) = 0$$

Bài tập số 4 I

- Phương sai:
Phương sai tổng của hai biến ngẫu nhiên:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y)$$

Nếu X và Y là độc lập, thì $\mathrm{Cov}(X,Y) = 0$, và công thức rút gọn thành:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Áp dụng:

$$\text{Nếu } Z = aX, \text{ thì } \mathrm{Var}(Z) = a^2 \cdot \mathrm{Var}(X)$$

$$\mathrm{Var}(z_m) = \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} x_i\right) = \left(\frac{1}{m}\right)^2 \cdot \mathrm{Var}\left(\sum_{i=1}^{m} x_i\right) = \frac{1}{m^2} \cdot m = \frac{1}{m}$$

Bài tập số 4 I

2. Có thể áp dụng bất đẳng thức Chebyshev cho từng z_m không?
Có vì:
Với mỗi giá trị cụ thể của z_m, ta có thể áp dụng công thức Chebyshev:

$$\mathbb{P}(|z_m| \geq \varepsilon) \leq \frac{1}{m\varepsilon^2}$$

$$\text{Nếu } m = 100, \varepsilon = 0.1, \text{ thì: } \mathbb{P}(|z_{100}| \geq 0.1) \leq \frac{1}{100 \cdot 0.01} = 1$$

$$\text{Nếu } m = 1000, \varepsilon = 0.1, \text{ thì: } \mathbb{P}(|z_{1000}| \geq 0.1) \leq \frac{1}{1000 \cdot 0.01} = 0.1$$

Bài tập số 5 l

(phần 2.6.8)

Given two events with probability $P(\mathcal{A})$ and $P(\mathcal{A})$, compute upper and lower bounds on $P(\mathcal{A} \cup \mathcal{B})$ and $P(\mathcal{A} \cap \mathcal{B})$. Hint: graph the situation using a Venn diagram.

Tóm tắt:

- Cho hai biến cố A và B với xác suất $P(A)$ và $P(B)$. Hãy tìm cận trên và cận dưới cho:
  - $P(A \cup B)$
  - $P(A \cap B)$.

Bài tập số 5 I

1. Giải thích bằng sơ đồ Venn
- Vẽ hai hình tròn giao nhau, một đại diện cho A, một cho B. Diện tích mỗi hình tròn tương ứng với xác suất của biến cố đó.
- Phần giao nhau thể hiện $P(A \cap B)$
- Toàn bộ phần nằm trong cả hai hình tròn (không tính phần chồng 2 lần) thể hiện $P(A \cup B)$.

Bài tập số 5 I

Lời giải bài tập 5

Chúng ta sẽ tìm cận trên và cận dưới của $P(A \cup B)$

Ta có công thức tổng quát:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Cận dưới:

Để $P(A \cup B)$ nhỏ nhất, thì $P(A \cap B)$ phải lớn nhất có thể, tức là $P(A \cap B) = min(P(A), P(B))$

Khi đó:

$$P(A \cup B)_{\min} = P(A) + P(B) - min(P(A), P(B)) = \max(P(A), P(B))$$

- Cận trên:

Để $P(A \cup B)$ lớn nhất, thì $P(A \cap B)$ phải nhỏ nhất có thể, tức là $P(A \cap B) = 0$ (hai biến cố rời nhau)

Khi đó:

$$P(A \cup B)_{\max} = P(A) + P(B)$$

Bài tập số 5 II

Mà tổng này không được vượt quá 1. Vậy:

$$P(A \cup B)_{\max} = min(1, P(A) + P(B))$$

Bài tập số 5 I

3. Chúng ta sẽ tìm cận trên và cận dưới của $P(A \cap B)$
- Cận trên:
Giao của A và B không thể lớn hơn biến cố nhỏ hơn, nên:

$$P(A \cap B)_{\max} = min(P(A), P(B))$$

- Cận dưới:
Từ công thức ở trên, để $P(A \cap B)$ nhỏ nhất, thì $P(A \cup B)$ phải lớn nhất, tức là:

$$P(A \cap B)_{\min} = P(A) + P(B) - min(1, P(A) + P(B)) = max(0, P(A) + P(B) - 1)$$

Bài tập số 6 I

(phần 2.6.8) Assume that we have a sequence of random variables, say A, B, and C, where B only depends on A, and C only depends on B, can you simplify the joint $P(A, B, C)$ probability? Hint: this is a Markov chain.
Tạm dịch: Cho A, B, C là các biến ngẫu nhiên. B chỉ phụ thuộc vào A, C chỉ phụ thuộc vào B. Đơn giản hóa $P(A, B, C)$ thế nào?

Bài tập số 6 I

Lời giải bài tập 6

C chỉ phụ thuộc vào B $\Rightarrow$ C độc lập có điều kiện với A khi đã biết B, ký hiệu là:

$$P(C|A,B) = P = (C|B)$$

Theo quy tắc chuỗi (chain rule) trong xác suất:

$$P(A,B,C) = P(A).P(B|A).P(C|A,B)$$

Ta được công thức rút gọn:

$$P(A,B,C) = P(A).P(B|A).P(C|B)$$

Bài tập số 7 I

Trong Mục 2.6.5, giả sử rằng kết quả của hai xét nghiệm không độc lập với nhau. Cụ thể, giả sử rằng mỗi xét nghiệm riêng lẻ có tỉ lệ dương tính giả là 10% và tỉ lệ âm tính giả là 1%. Tức là, giả sử rằng:

- $P(D = 1 \mid H = 0) = 0.1$ (xét nghiệm cho kết quả dương tính khi bệnh nhân không bị bệnh)
- $P(D = 0 \mid H = 1) = 0.01$ (xét nghiệm cho kết quả âm tính khi bệnh nhân thật sự bị bệnh)

Hơn nữa, giả sử rằng đối với trường hợp bệnh nhân bị nhiễm bệnh ($H = 1$), kết quả hai xét nghiệm là độc lập có điều kiện, tức là:

$$P(D_1, D_2 \mid H = 1) = P(D_1 \mid H = 1) \cdot P(D_2 \mid H = 1)$$

Nhưng đối với bệnh nhân khỏe mạnh ($H = 0$), kết quả hai xét nghiệm phụ thuộc theo phân phối:

$$P(D_1 = D_2 = 1 \mid H = 0) = 0.02$$

Bài tập số 7 II

1. Hãy xây dựng bảng xác suất liên hợp (joint probability table) cho D_1 và D_2, biết rằng $H = 0$, dựa trên thông tin đã cho ở trên.
2. Suy ra xác suất bệnh nhân bị bệnh ($H = 1$) sau khi một xét nghiệm cho kết quả dương tính. Bạn có thể giả sử xác suất tiên nghiệm (prior) $P(H = 1) = 0.0015$ như trước.
3. Suy ra xác suất bệnh nhân bị bệnh ($H = 1$) sau khi cả hai xét nghiệm đều dương tính.

Tóm tắt bài toán
Biến cố:

- $H = 1$: Bệnh nhân mắc bệnh (nhiễm bệnh).
- $H = 0$: Bệnh nhân khỏe mạnh.
- D_1, D_2: Kết quả của hai xét nghiệm ($D_i = 1$ nếu dương tính, $D_i = 0$ nếu âm tính).

Giả định:

- Xác suất:
  - Xác suất tiên nghiệm: $P(H = 1) = 0.0015$
    $\Rightarrow P(H = 1) = 0.9985$.

Bài tập số 7 III

- Xét nghiệm dương tính giả (False Positive): $P(D_i = 1 \mid H = 0) = 0.1$.
  $\Rightarrow$ Xét nghiệm âm tính thật (True Positive): $P(D_i = 0 \mid H = 0) = 0.9$.
- Xét nghiệm âm tính giả (False Negative): $P(D_i = 0 \mid H = 1) = 0.01$.
  $\Rightarrow$ Xét nghiệm dương tính thật (True positive): $P(D_i = 1 \mid H = 1) = 0.99$.

- Tính độc lập có điều kiện:
  - Nếu $H = 1$: D_1 và D_2 độc lập, tức:

$$P(D_1, D_2 \mid H = 1) = P(D_1 \mid H = 1)P(D_2 \mid H = 1).$$

  - Nếu $H = 0$: D_1 và D_2 không độc lập, với:

$$P(D_1 = D_2 = 1 \mid H = 0) = 0.02.$$

Câu hỏi:

1. Tính bảng xác suất chung cho D_1 và D_2 khi đã biết $H = 0$ dựa trên các thông tin trên.
2. Suy ra xác suất bệnh nhân thực sự bị bệnh ($H = 1$) sau khi một xét nghiệm cho kết quả dương tính. (Giả sử xác suất ban đầu $P(H = 1) = 0.0015$ như trước.)

Bài tập số 7 IV

3. Suy ra xác suất bệnh nhân bị bệnh ($H = 1$) sau khi cả hai xét nghiệm đều cho kết quả dương tính.

Lời giải bài tập 7 7.1. Bảng xác suất đồng thời cho $P(D_1, D_2 \mid H = 0)$.
Ta có: Với $H = 0$, hai xét nghiệm không độc lập nhưng có thể mô hình hóa thông qua xác suất đồng thời. Ta cần điền các giá trị còn lại của bảng:

$D_1 \backslash D_2$	$D_2 = 0$	$D_2 = 1$	Tổng
$D_1 = 0$	a	b	0.9
$D_1 = 1$	c	0.02	0.1
Tổng	0.9	0.1	1

Bảng: Bảng xác suất đồng thời

Tính toán:

- $c = P(D_1 = 1 \mid H = 0) - P(D_1 = D_2 = 1 \mid H = 0) = 0.1 - 0.02 = 0.08.$
- $b = P(D_2 = 1 \mid H = 0) - P(D_1 = D_2 = 1 \mid H = 0) = 0.1 - 0.02 = 0.08.$

Bài tập số 7 V

- $a = 0.9 - b = 0.9 - 0.08 = 0.82$.

Kết quả:

$D_1 \backslash D_2$	$D_2 = 0$	$D_2 = 1$	Tổng
$D_1 = 0$	0.82	0.08	0.9
$D_1 = 1$	0.08	0.02	0.1
Tổng	0.9	0.1	1

Bảng: Bảng xác suất đồng thời

7.2. Xác suất bệnh sau một xét nghiệm dương tính ($P(H = 1 \mid D_1 = 1)$) Công thức Bayes:

$$P(H = 1 \mid D_1 = 1) = \frac{P(D_1 = 1 \mid H = 1)P(H = 1)}{P(D_1 = 1)}$$

Tính các thành phần:

Bài tập số 7 VI

- $P(D_1 = 1 \mid H = 1) = 1 - P(D_1 = 0 \mid H = 1) = 1 - 0.01 = 0.99.$
- $P(H = 1) = 0.0015$ (đề bài cho) $\Rightarrow P(H = 0) = 0.9985$.
- $P(D_1 = 1) = P(D_1 = 1 \mid H = 1)P(H = 1) + P(D_1 = 1 \mid H = 0)P(H = 0)$
- $P(D_1 = 1) = 0.99 \times 0.0015 + 0.1 \times 0.9985 \approx 0.101485.$

Kết quả:

$$P(H = 1 \mid D_1 = 1) = \frac{0.99 \times 0.0015}{0.101485} \approx 0.0146 \quad (1.46\%).$$

7.3. Xác suất bệnh sau hai xét nghiệm dương tính ($P(H = 1 \mid D_1 = D_2 = 1)$)
Ta có công thức Bayes:

$$P(H = 1 \mid D_1 = D_2 = 1) = \frac{P(D_1 = D_2 = 1 \mid H = 1)P(H = 1)}{P(D_1 = D_2 = 1)}.$$

Do D_1, D_2 độc lập khi $H = 1$, ta có:

$$P(D_1 = D_2 = 1 \mid H = 1) = P(D_1 = 1 \mid H = 1)^2 = 0.99^2 = 0.9801.$$

Bài tập số 7 VII

- Tử số: $P(D\_1 = D\_2 = 1 \mid H = 1)P(H = 1) = 0.9801 \times 0.0015 \approx 0.00147$.
- Mẫu số:

$$P(D_1 = D_2 = 1) = P(D_1 = D_2 = 1 \mid H = 1)P(H = 1) + P(D_1 = D_2 = 1 \mid H = 0)P(H = 0)$$

$$P(D_1 = D_2 = 1) = 0.9801 \times 0.0015 + 0.02 \times (1 - 0.0015) \approx 0.02147.$$

- Kết quả:

$$P(H = 1 \mid D_1 = D_2 = 1) = \frac{0.9801 \times 0.0015}{0.9801 \times 0.0015 + 0.02 \times (1 - 0.0015)} \approx 0.0685 \quad (6.85\%).$$

Vậy, sau hai xét nghiệm đều dương tính, xác suất bệnh nhân thực sự bị nhiễm là khoảng 6.85%. Dù xác suất này nhỏ hơn 10%, nhưng vẫn lớn hơn nhiều so với xác suất ban đầu 0.15%.

Kết quả của chạy Code Python để tính toán, trực quan hóa kết quả cho bài tập 7

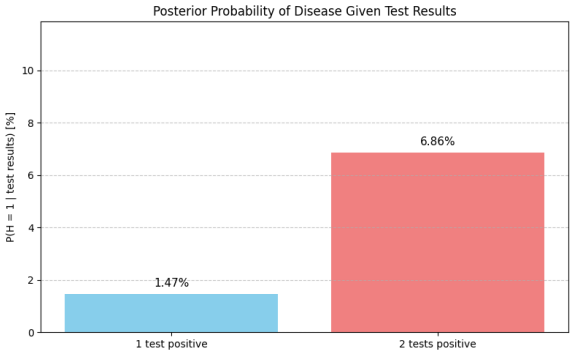
Bài tập số 7 VIII

```
Bảng xác suất với H=0:
P(D1=0, D2=0 | H=0) = 0.82000
P(D1=0, D2=1 | H=0) = 0.08000
P(D1=1, D2=0 | H=0) = 0.08000
P(D1=1, D2=1 | H=0) = 0.02000

Xác suất bệnh khi 1 test dương tính:
P(H=1 | D1=1) ≈ 0.01465

Xác suất bệnh khi 2 test đều dương tính:
P(H=1 | D1=1, D2=1) ≈ 0.06857
```

Hình: Kết quả bài tập 7(Python)

Trực quan hóa hai xác suất hậu nghiệm $P(H = 1 \mid D = 1)$ và $P(H = 1 \mid D_1 = 1, D_2 = 1)$ bằng biểu đồ cột

Bài tập số 7 IX



Hình: Biểu đồ trực quan hóa hai xác suất hậu nghiệm

Bài tập số 7 X

Mô phỏng bài toán bằng thử nghiệm ngẫu nhiên (Monte Carlo)
Cho kết quả tương đồng với lời giải ở trên

$$\text{Monte Carlo estimate of } P(H = 1|D1 = 1, D2 = 1) = 0.0681(\approx 6.81\%)$$

Bài tập số 8 I

Giả sử bạn là một quản lý tài sản tại một ngân hàng đầu tư và bạn có nhiều lựa chọn cổ phiếu s_i để đầu tư. Danh mục đầu tư của bạn cần có tổng trọng số bằng 1 ($\sum_{i=1}^{n} \alpha_i = 1$), với trọng số α_i cho mỗi cổ phiếu. Các cổ phiếu có mức lợi nhuận trung bình

$$\mu = E_{s \sim P}[s]$$

và hiệp phương sai

$$\Sigma = \mathrm{Cov}_{s \sim P}[s].$$

1. Tính toán lợi nhuận kỳ vọng cho một danh mục đầu tư **s** đã cho.
2. Nếu muốn tối đa hóa lợi nhuận của danh mục đầu tư, nên phân bổ đầu tư như thế nào?
3. Tính toán phương sai của danh mục đầu tư.

Bài tập số 8 II

4. Xây dựng bài toán tối ưu để tối đa hóa lợi nhuận trong khi giữ phương sai ở một ngưỡng nhất định. Đây chính là danh mục đầu tư Markowitz đã đoạt giải Nobel (Mangram, 2013). Để giải bài toán này, sẽ cần sử dụng một trình giải tối ưu bậc hai (quadratic programming solver), một công cụ vượt xa phạm vi của cuốn sách này.

Lời giải bài tập 8

Ta có:

- Lợi nhuận trung bình của cổ phiếu: $\mu = E_{s\sim P}[s]$.
- Ma trận hiệp phương sai: $\Sigma = \text{Cov}_{s\sim P}[s]$.
- Danh mục đầu tư $\mathbf{s} = \{s_1, s_2, ..., s_n\}$: Gồm các cổ phiếu s_i với trọng số tương ứng α_i, $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_n\}$, với ràng buộc:

$$\sum_{i=1}^{n} \alpha_i = 1$$

- Lợi nhuận trung bình của cổ phiếu: $\mu = E_{s\sim P}[\mathbf{s}]$.

Bài tập số 8 III

- Ma trân hiệp phương sai: $\Sigma = \text{Cov}_{s\sim P}[\mathbf{s}]$, với vector ngẫu nhiên $\mathbf{s} = [s_1, s_2, \ldots, s_n]^T$, ma trận Σ có dạng:

$$\Sigma = \begin{bmatrix} \text{Var}(s_1) & \text{Cov}(s_1, s_2) & \ldots & \text{Cov}(s_1, s_n) \\ \text{Cov}(s_2, s_1) & \text{Var}(s_2) & \ldots & \text{Cov}(s_2, s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(s_n, s_1) & \text{Cov}(s_n, s_2) & \ldots & \text{Var}(s_n) \end{bmatrix}$$

8.1. Tính lợi nhuận kỳ vọng của danh mục:
- Lợi nhuận kỳ vọng của danh mục là trung bình trọng số của lợi nhuận các cổ phiếu: - Mỗi cổ phiếu s_i đóng góp tỷ suất sinh lợi trung bình là μ_i, đầu tư trọng số α_i vào nó, nên lợi nhuận kỳ vọng của danh mục là

$$E[\mathbf{s}] = \sum_{i=1}^{n} \alpha_i \mu_i = \alpha^\top \mu$$

- Ví dụ: Nếu danh mục gồm 2 cổ phiếu với $\alpha = \{0.6, 0.4\}$ và $\mu = \{0.1, 0.05\}$, thì:

$$E[\alpha] = 0.6 \times 0.1 + 0.4 \times 0.05 = 0.08 \quad (8\%).$$

Bài tập số 8 IV

8.2. Tối đa hóa lợi nhuận:
Để tối đa hóa lợi nhuận kỳ vọng của danh mục đầu tư, ta cần phân bổ toàn bộ vốn vào cổ phiếu có lợi nhuận kỳ vọng lớn nhất. Nếu không xét đến rủi ro (phương sai), lời giải đơn giản là:

$$\alpha_j = \begin{cases} 1 & \text{nếu } j = \arg\max_i \mu_i \\ 0 & \text{với } i \neq j \end{cases}$$

Trong đó, μ_i là lợi nhuận kỳ vọng của cổ phiếu s_i.
Tuy nhiên, cần lưu ý rằng ta chưa xét đến đến rủi ro và lời giải này chỉ đúng trong trường hợp:

- Chỉ quan tâm đến lợi nhuận kỳ vọng (không xét phương sai).
- Không có ràng buộc khác ngoài $\sum \alpha_i = 1$ và $\alpha_i \geq 0$ (không bán khống).

Bài tập số 8 V

Trong thực tế, nhà đầu tư thường cân bằng giữa lợi nhuận và rủi ro, do đó cần xem xét thêm phương sai danh mục (ở phần 8.3) hoặc giải bài toán Markowitz (ở phần 8.4). 8.3. Tính phương sai của danh mục đầu tư: Phương sai danh mục phản ánh độ biến động:

$$\text{Var}(\alpha^T \mathbf{s}) = \alpha^T \Sigma \alpha = \sum_{i,j} \alpha_i \alpha_j \Sigma_{i,j}.$$

Ví dụ: Nếu $\Sigma = \begin{bmatrix} 0.04 & 0.01 \\ 0.01 & 0.02 \end{bmatrix}$ và $\alpha = [0.6, 0.4]^T$:

$$\text{Var}(\alpha^T \mathbf{s}) = 0.6^2 \times 0.04 + 2 \times 0.6 \times 0.4 \times 0.01 + 0.4^2 \times 0.02 = 0.0203.$$

8.4. Bài toán tối ưu danh mục Markovitz:

Bài tập số 8 VI

Tối đa hóa lợi nhuận với ràng buộc rủi ro tối đa $\sigma^2_{\max}$:

$$\begin{aligned} \max_{\alpha} \quad & \alpha^T \mu \\ \text{subject to} \quad & \alpha^T \Sigma \alpha \leq \sigma^2_{\max}, \\ & \sum \alpha_i = 1, \\ & \alpha_i \geq 0 \quad \text{(nếu không cho phép bán khống).} \end{aligned}$$

Để mô phỏng cho bài tập 8: Thực hiện lấy dữ liệu từ nguồn thị trường chính khoán Việt Nam (import vnstock as vn). Đặc tả dữ liệu:

- Chạy thực nghiệm trên 5 mã chứng khoán: VCB, FPT, MWG, VNM, HPG.
- Thời điểm lấy dữ liệu từ ngày 01/01/2023 đến ngày 01/01/2024.
- Điểm dữ liệu: là giá trị trung bình giao dịch trong 1 ngày của tất cả các phiên khớp lệnh trong ngày đó. Có 250 ngày giao dịch $\Rightarrow$ có 250 điểm dữ liệu.

Kết quả chạy thực nghiệm cho bài tập 8:

Bài tập số 8 VII

1. (8.1) Lợi nhuận kỳ vọng hàng ngày và hàng năm của danh mục
   - Lợi suất kỳ vọng hàng ngày của danh mục: 0.1898%
   - Lợi suất kỳ vọng hàng năm của danh mục: 47.8326%
2. (8.2) Tối đa hóa lợi nhuận danh mục đầu tư
   - Tối đa hóa lợi nhuận danh mục:
     - VCB: 0.00%
     - FPT: 100.00%
     - MWG: 0.00%
     - VNM: 0.00%
     - HPG: 0.00%
   - Lợi suất kỳ vọng danh mục tối đa lợi nhuận: 0.19%
   - Rủi ro tương ứng (độ lệch chuẩn): 1.96%
3. (8.3) Tính toán phương sai của danh mục đầu tư
   - Phương sai danh mục Markowitz: 0.000210
4. (8.4) Xây dựng bài toán tối ưu để tối đa hóa lợi nhuận trong khi giữ phương sai ở một ngưỡng nhất định (Tối đa hóa lợi nhuận với ràng buộc phương sai).
   - Danh mục tối ưu với ràng buộc phương sai:

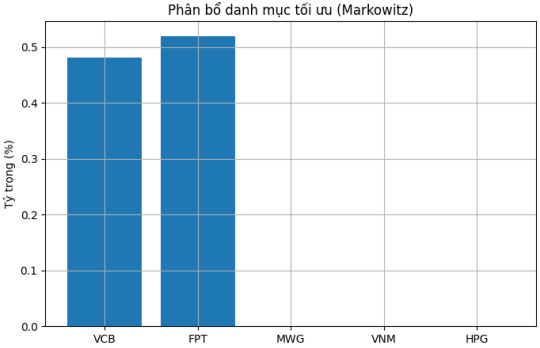
Bài tập số 8 VIII

- VCB: 0.00%
- FPT: 100.00%
- MWG: 0.00%
- VNM: 0.00%
- HPG: 0.00%

- Lợi suất kỳ vọng: 0.19%
- Phương sai danh mục: 0.000383
- Rủi ro tương ứng (độ lệch chuẩn): 1.96%

5. (8.5) Trực quan hóa danh mục đầu tư tối ưu bằng Bar Chart

Bài tập số 8 IX



Hình: Danh mục đầu tư tối ưu

Bài tập nâng cao 1 I

Bài toán nâng cao dựa trên bài tập 5 (phần 2.6.8). Bài toán mô tả như sau:
Cho ba biến cố A, B, C với các xác suất đã biết:

- $P(A) = a$
- $P(B) = b$
- $P(C) = c$

Yêu cầu bài toán: Hãy tìm giới hạn trên và giới hạn dưới có thể có của $P(A \cup B \cup C)$ và $P(A \cap B \cap C)$.

Bài tập nâng cao 1 II

Lời giải bài toán NC1

1. Giới hạn của $P(A \cup B \cup C)$
   - Công thức xuất đầy đủ:

     $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

   - Giới hạn trên:

     $$P(A \cup B \cup C) \leq \min(1, P(A) + P(B) + P(C))$$

     Tối đa là tổng các xác suất, nhưng không vượt quá 1.
   - Giới hạn dưới:

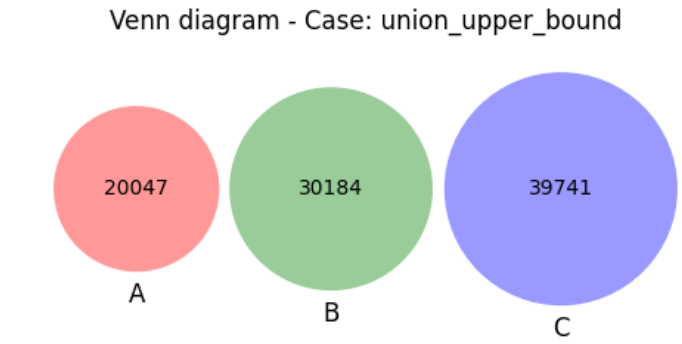
     $$P(A \cup B \cup C) \geq \max(P(A), P(B), P(C))$$

     Hoặc chặt hơn: $\max(P(A) + P(B) + P(C) - 2, \max(P(A), P(B), P(C)))$
     Xảy ra khi các tập giao nhau hoặc bị chồng lấp nhiều.

Bài tập nâng cao 1 III

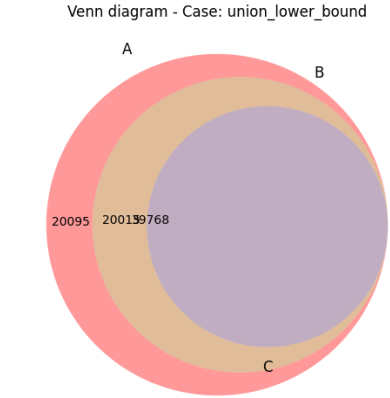
2. Giới hạn của $P(A \cap B \cap C)$
   - Giới hạn trên:
   $$P(A \cap B \cap C) \leq \min(P(A), P(B), P(C))$$
   Lớn nhất khi một tập nằm trong hai tập còn lại.
   - Giới hạn dưới:
   $$P(A \cap B \cap C) \geq \max(0, P(A) + P(B) + P(C) - 2)$$
   Nhỏ nhất khi ba tập gần rời nhau (giao nhỏ nhất).

Bài tập nâng cao 1 IV



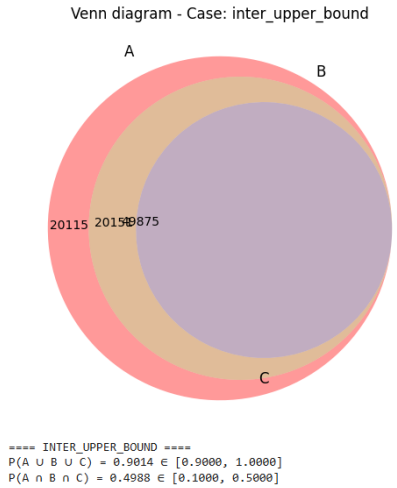
Hình: Mô phỏng trường hợp xảy ra dấu bằng tại giới hạn trên của $P(A \cup B \cup C)$

Bài tập nâng cao 1 V



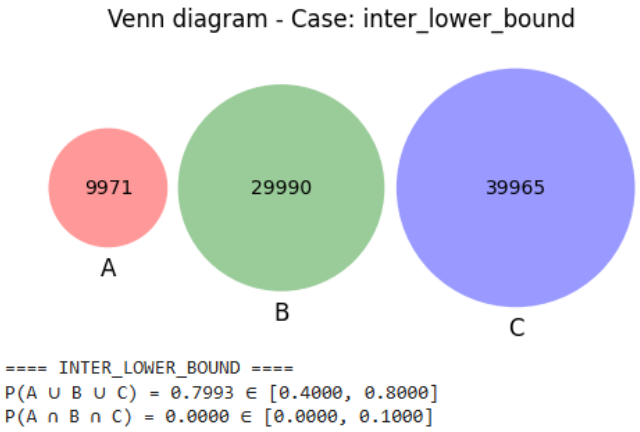
Hình: Mô phỏng trường hợp xảy ra dấu bằng tại giới hạn dưới của $P(A \cup B \cup C)$

Bài tập nâng cao 1 VI



Hình: Mô phỏng trường hợp xảy ra dấu bằng tại giới hạn trên của $P(A \cap B \cap C)$

Bài tập nâng cao 1 VII



Hình: Mô phỏng trường hợp xảy ra dấu bằng tại giới hạn dưới của $P(A \cap B \cap C)$

Bài tập nâng cao 2 I

Bài toán nâng cao dựa trên bài tập 6 (phần 2.6.8). Bài toán mô tả như sau:

Bài toán: Phát hiện gian lận qua mô hình Bayes

Một công ty tài chính đang sử dụng hệ thống phát hiện gian lận giao dịch dựa trên các biến sau:

- F: Biến nhị phân cho biết giao dịch có gian lận hay không (1 = gian lận, 0 = bình thường).
- L: Biến nhị phân cho biết liệu giao dịch được thực hiện từ vị trí lạ không (1 = vị trí lạ, 0 = vị trí quen thuộc).
- T: Biến nhị phân cho biết thời điểm giao dịch có phải vào giờ bất thường không (1 = bất thường, 0 = bình thường).

Bài tập nâng cao 2 II

Giả sử mô hình thỏa mãn mối quan hệ:

- L và T độc lập có điều kiện khi biết F.
- Sơ đồ phụ thuộc có thể được biểu diễn như:

$$F \to L$$

$$F \to T$$

Bài tập nâng cao 2 III

Yêu cầu:

1. Viết biểu thức xác suất đồng thời $P(F, L, T)$ dựa trên cấu trúc phụ thuộc nêu trên.
2. Sử dụng định lý Bayes để viết công thức tính $P(F = 1|L = 1, T = 1)$.
3. Giả sử có các xác suất sau:
   - $P(F = 1) = 0.01$
   - $P(L = 1|F = 1) = 0.9$, $P(L = 1|F = 0) = 0.1$
   - $P(T = 1|F = 1) = 0.8$, $P(T = 1|F = 0) = 0.2$

   Tính giá trị cụ thể của $P(F = 1|L = 1, T = 1)$.

Bài tập nâng cao 2 IV

Lời giải bài toán NC2

1. Viết biểu thức xác suất đồng thời $P(F, L, T)$
   Biểu thức xác suất đồng thời:

$$P(F, L, T) = P(F) \cdot P(L|F) \cdot P(T|F)$$

2. Sử dụng định lý Bayes để tính $P(F = 1|L = 1, T = 1)$
   Theo định lý Bayes:

$$P(F = 1|L = 1, T = 1) = \frac{P(F = 1) \cdot P(L = 1|F = 1) \cdot P(T = 1|F = 1)}{P(L = 1, T = 1)}$$

Mẫu số $P(L = 1, T = 1)$ được tính bằng cách tổng trên tất cả giá trị của $F \in \{0, 1\}$:

$$P(L = 1, T = 1) = \sum_{f \in \{0,1\}} P(F = f) \cdot P(L = 1|F = f) \cdot P(T = 1|F = f)$$

Bài tập nâng cao 2 V

3. Tính giá trị cụ thể của $P(F = 1|L = 1, T = 1)$ với các giá trị xác suất đã cho
Tính mẫu số:

$$P(L = 1, T = 1) = 0.01 \cdot 0.9 \cdot 0.8 + 0.99 \cdot 0.1 \cdot 0.2 = 0.0072 + 0.0198 = 0.027$$

Tính tử số:

$$\text{Tử số} = 0.01 \cdot 0.9 \cdot 0.8 = 0.0072$$

Suy ra:

$$P(F = 1|L = 1, T = 1) = \frac{0.0072}{0.027} \approx 0.2667$$

Ta nhận thấy mặc dù khả năng gian lận gốc chỉ là 1%, nhưng khi thấy **vị trí lạ** + **thời điểm bất thường**, xác suất gian lận tăng lên $\approx 26.67\%$!

Bài tập nâng cao 3 I

Tóm tắt bài toán NC3- INPUT:

1. $H = 1$: Bệnh nhân nhiễm bệnh.
2. $H = 0$: Bệnh nhân khỏe mạnh.
3. D_i: Kết quả của xét nghiệm thứ i ($D_i = 1$ nếu kết quả xét nghiệm dương tính, $D_i = 0$ nếu kết quả xét nghiệm âm tính).
4. Xác suất:
   - Xác suất tiên nghiệm: $P(H = 1) = 0.0015$ ($\Rightarrow P(H = 0) = 0.9985$).
   - Xét nghiệm dương tính giả (False Positive): $P(D_i = 1 \mid H = 0) = 0.1$.
   - $\Rightarrow$ Xét nghiệm âm tính thật (True Positive): $P(D_i = 0 \mid H = 0) = 0.9$.
   - Xét nghiệm âm tính giả (False Negative): $P(D_i = 0 \mid H = 1) = 0.01$.
   - $\Rightarrow$ Xét nghiệm dương tính thật (True positive): $P(D_i = 1 \mid H = 1) = 0.99$.
5. Xét nghiệm độc lập theo H:
   - $P(D_1, D_2, ...D_k \mid H = 1) = P(D_1 \mid H = 1)P(D_2 \mid H = 1)...P(D_k \mid H = 1)$
   - $P(D_1, D_2, ...D_k \mid H = 0) = P(D_1 \mid H = 0)P(D_2 \mid H = 0)...P(D_k \mid H = 0)$

Bài tập nâng cao 3 II

6. Chi phí:
   - Điều trị đúng thì chi phí: 0$/người
   - Điều trị nhầm người không mắc bệnh, bệnh viện phải tốn chi phí: 500$/người
   - Bỏ sót điều trị người mắc bệnh, bệnh viện phải tốn chi phí: 10,000$/người
   - Một lần xét nghiệm bệnh nhân, bệnh viện phải tốn chi phí: 50$
7. Ràng buộc bài toán: Được phép xét nghiệm tối đa 3 lần cho mỗi bệnh nhân, với các lần xét nghiệm độc lập có điều kiện theo H. Sau mỗi lần xét nghiệm, có thể quyết định:
   - Hành động (A): Dừng lại (không xét nghiệm thêm 1 lần nào nữa) và điều trị.
   - Hành động (B): Dừng lại (không xét nghiệm thêm 1 lần nào nữa) và không điều trị.
   - Hành động (C): Tiếp tục xét nghiệm lần 3 (lần cuối).

Câu hỏi bài toán NC3- OUTPUT:

1. Tính xác suất hậu nghiệm $P(H = 1|D_1, D_2, ...D_k)$ sau mỗi bước.
2. Tính chi phí kỳ vọng cho mỗi hành động A, B, C.
3. Mô phỏng thuật toán giúp chọn hành động tối ưu ở mỗi bước xét nghiệm.
4. Sử dụng mô phỏng Monte Carlo để kiểm nghiệm chiến lược trên với 1 triệu bệnh nhân ngẫu nhiên.

Bài tập nâng cao 3 III

Lời giải bài toán NC3

Câu 1: Tính xác suất hậu nghiệm $P(H = 1|D_1, D_2, ..., D_k)$

- Công thức Bayes: Sau k lần xét nghiệm, có s kết quả xét nghiệm dương tính:

$$P(H = 1|D_1, ..., D_k) = \frac{P(D_1, ..., D_k|H = 1) \cdot P(H = 1)}{P(D_1, ..., D_k)} \tag{7}$$

hay

$$P(H = 1 \mid D_1, ..., D_k) = \frac{P(D_1=...=D_s=1|H=1)\cdot P(D_{s+1}=...=D_k=0|H=1)\cdot P(H=1)}{P(D_1,...,D_k|H=1)+P(D_1,...,D_k|H=0)} \tag{8}$$

$$(D_1 = ... = D_s = 1, D_{s+1} = ... = D_k = 0)$$

Vì các test độc lập có điều kiện theo H, ta có:

Bài tập nâng cao 3 IV

Tử số của (8)

$$\begin{aligned} P(D_1, ..., D_k|H = 1) &= P(D_1 = ... = D_s = 1|H = 1).P(D_{s+1} = ... = D_k = 0|H = 1) \\ &= (P(D_1 = 1|H = 1)^s.(P(D_k = 0|H = 1)^{k-s} \\ &= (0.99)^s \cdot (0.01)^{k-s} \end{aligned}$$

và

$$\begin{aligned} P(D_1, ..., D_k|H = 0) &= P(D_1 = ... = D_s = 1|H = 0)P(D_{s+1} = ... = D_k = 0|H = 0) \\ &= (P(D_1 = 1|H = 0))^s.(P(D_k = 0|H = 0))^{k-s} \\ &= (0.10)^s \cdot (0.90)^{k-s} \end{aligned}$$

Do đó:

$$P(H = 1|D_1, ..., D_k) = \frac{(0.99)^s(0.01)^{k-s} \cdot 0.0015}{(0.99)^s(0.01)^{k-s} \cdot 0.0015 + (0.10)^s(0.90)^{k-s} \cdot 0.9985}$$

Bài tập nâng cao 3 V

Câu 2: Tính chi phí kỳ vọng của các hành động
Từ công thức (8), ta đặt:

$$bayes\_posterior(s, k) = P(H = 1|D_1, ..., D_k)$$

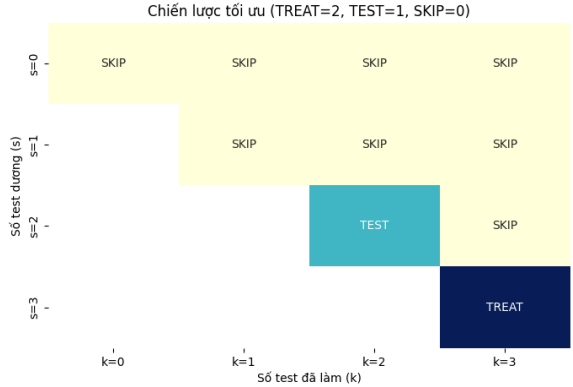
- Giả sử tại bước xét nghiệm thứ k, có số lần xét nghiệm dương tính là s, gọi hàm:

$$p = bayes\_posterior(s, k)$$

  - Nếu hành động (A): Điều trị thì gọi hàm tính chi phí: $cost\_A = (1 - p) * 500 + k * 50$;
  - Nếu hành động (B): Không điều trị thì gọi hàm tính chi phí: $cost\_B = p * 10000 + k * 50$;
  - Nếu hành động (C) tiếp tục test$\rightarrow$ xây dựng cây quyết định đệ quy ở câu 3.

Bài tập nâng cao 3 VI

Câu 3: Mô phỏng thuật toán giúp chọn hành động tối ưu ở mỗi bước xét nghiệm
Mô phỏng, trực quan hóa cây quyết định để tìm chiến lược điều trị tối ưu



Hình: Mô phỏng, trực quan hóa cây quyết định

Bài tập nâng cao 3 VII

Câu 4: Sử dụng mô phỏng Monte Carlo để kiểm nghiệm chiến lược trên với 1 triệu bệnh nhân ngẫu nhiên để tính chi phí trung bình mà bệnh viện phải chi trả.
Ý tưởng

- Sinh $H \sim \text{Bernoulli}(P\_H1)$
- Mỗi bước test sinh ngẫu nhiên dựa trên H
- Áp dụng phương thức $expected\_cost()$ để ra quyết định sau mỗi test
- Tính tổng chi phí thực tế trên 1 triệu bệnh nhân

Kết quả phỏng Monte Carlo để kiểm nghiệm chiến lược trên với 1 triệu bệnh nhân ngẫu nhiên, khi xuất ra màn hình là:

Chi phí trung bình cho mỗi bệnh nhân mà bệnh viện phải chi trả: 15.07$

Kết luận

Nhóm đã thực hiện:

- Trình bày ngắn gọn các kiến thức nền tảng của xác suất thống kê.
- Trình bày về sự bất định trong machine learning
- Các vấn đề về thu thập thêm dữ liệu giúp giảm bất định.
- Trình bày định lý Bayes và các bài toán ứng dụng của định lý Bayes.
- Ứng dụng của kỳ vọng và xác suất vào trong các bài toán thực tế để xây dựng mô hình hóa xác suất.

References

https://d2l.ai/chapter_preliminaries/probability.html

Probability and Statistics for Computer Science, David Forsyth, (2018).

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... et al. (2022). Flamingo: a visual language model for few-shot learning. ArXiv:2204.14198.

Alsallakh, B., Kokhlikyan, N., Miglani, V., Yuan, J., & Reblitz-Richardson, O. (2020). Mind the PAD – CNNs can develop blind spots. ArXiv:2010.02178.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... et al. (2023). PaLM 2 Technical Report. ArXiv:2305.10403.

Nhập Môn Hiện Đại Xác Suất Thống Kê, GS. Nguyễn Tiến Dũng và GS. Đỗ Đức Thái

Robert V. Hogg, Joseph W. McKean, Allen T. Craig. 2013. "Introduction to Mathematical Statistics". 7th edition.

Sheldon Ross. 2020. "A First Course in Probability". 10th edition.

William Mendenhall, Robert J. Beaver, Barbara M. Beaver. 2009. "Introduction to Probability and Statistics". 13th edition.

Q&A

Thank you very much!

We are Group 7

Contact us:

tdhung.sdh242@hcmut.edu.vn
thtam.sdh242@hcmut.edu.vn
nmdai.sdh242@hcmut.edu.vn
vuongminhtoan2@gmail.com
ndnminh.sdh232@hcmut.edu.vn
nxhien.sdh242@hcmut.edu.vn