

TRƯỜNG ĐẠI HỌC BÁCH KHOA TP. HỒ CHÍ MINH  
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH  
CỞ SỞ TOÁN CHO KHOA HỌC MÁY TÍNH (CO5263)

Đề tài:  
Xác suất và Thống kê

GVHD:

TS. Nguyễn An Khương

TS. Trần Tuấn Anh

Học viên:

Ngô Minh Đại - 2470722

Nguyễn Xuân Hiền - 2470749

Trần Đăng Hùng - 2470750

Nguyễn Đình Nhật Minh - 2370736

Trần Hoài Tâm - 2470743

Vương Minh Toàn - 2491057

- 1 Giới thiệu
- 2 Biến ngẫu nhiên
- 3 Đa biến ngẫu nhiên
- 4 Ví dụ
- 5 Kỳ vọng, phương sai, độ lệch chuẩn
- 6 Bài tập

# Xác suất là gì

## Định nghĩa xác suất

Cho một không gian mẫu  $\Omega$ , biến cố  $A \subseteq \Omega$ .

- 1  $0 \leq P(A) \leq 1$  với mọi  $A \subseteq \Omega$ ;
- 2  $P(\emptyset) = 0$ ,  $P(\Omega) = 1$ ;
- 3 Nếu  $A \cap B = \emptyset$  thì  $P(A \cup B) = P(A) + P(B)$ .

# Tung đồng xu

- **Thí nghiệm:** Tung một đồng xu công bằng.
- **Kết quả:** - Ngửa (Heads) - Sấp (Tails)
- **Xác suất lý thuyết:**

$$P(\text{Ngửa}) = 0.5, \quad P(\text{Sấp}) = 0.5.$$

- **Ước lượng thực nghiệm:** Sau  $N$  lần tung, ghi  $H$  lần ngửa, ta có

$$\hat{P}(\text{Ngửa}) = \frac{H}{N}, \quad \hat{P}(\text{Sấp}) = 1 - \frac{H}{N}.$$

# Tung xúc xắc

- **Thí nghiệm:** Tung một con xúc xắc công bằng.
- **Kết quả (không gian mẫu):**  $\Omega = \{1, 2, 3, 4, 5, 6\}$ .
- **Ví dụ biến cố:** - Chẵn:  $A = \{2, 4, 6\}$ . - Lẻ:  $B = \{1, 3, 5\}$ .
- **Xác suất lý thuyết:**

$$P(A) = \frac{3}{6} = 0.5, \quad P(B) = 0.5.$$

# Biến ngẫu nhiên rời rạc

## Định nghĩa

### Định nghĩa

Biến ngẫu nhiên  $X$  là hàm gán giá trị số cho mỗi kết quả trong không gian mẫu  $\Omega$ .

### Ví dụ nhị phân

Lấy số ngẫu nhiên  $\omega \in \{1, \dots, 100\}$ , định nghĩa

$$B(\omega) = \begin{cases} 1, & \omega > 50, \\ 0, & \omega \leq 50. \end{cases}$$

$B$  chỉ nhận hai giá trị: 0 hoặc 1.

# Biến ngẫu nhiên rời rạc

## Định nghĩa

### Phân phối rời rạc (PMF)

$$p_X(x) = P(X = x), \quad \sum_x p_X(x) = 1.$$

### Xác suất trên khoảng

Với  $a \leq b$  và rời rạc:

$$P(a \leq X \leq b) = \sum_{x=a}^b p_X(x).$$

# Biến ngẫu nhiên rời rạc

## Định nghĩa

Ví dụ: Tung 3 đồng xu

$X$  = số lần ra ngửa trong 3 lần tung, nên  $X \in \{0, 1, 2, 3\}$ .

$$P(X = k) = \binom{3}{k} \left(\frac{1}{2}\right)^3, \quad k = 0, 1, 2, 3.$$

Cụ thể:

$$P(X = 0) = \frac{1}{8}, \quad P(X = 1) = \frac{3}{8}, \quad P(X = 2) = \frac{3}{8}, \quad P(X = 3) = \frac{1}{8}.$$

Hàm phân phối tích lũy tại 1:

$$F_X(1) = P(X \leq 1) = P(0) + P(1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}.$$



# Đa biến ngẫu nhiên

## Giới thiệu

Khi xét nhiều biến ngẫu nhiên cùng lúc, chúng ta quan tâm tới mối quan hệ giữa chúng.

- Biết giá trị của biến này giúp cập nhật niềm tin về biến khác.
- **Ví dụ y tế:** Bệnh nhân sốt cao (X) và nổi phát ban (Y)  $\Rightarrow$  xác suất cao bị sốt xuất huyết.

# Đa biến ngẫu nhiên

## Không gian mẫu

**Định nghĩa:** Tập hợp tất cả kết quả có thể của cặp biến  $(A, B)$ .

- *Ví dụ bóng:* Hộp có 2 đỏ (R), 1 xanh (G), rút 2 quả không hoàn lại.
- Không gian mẫu:

$$\Omega = \{(R, R), (R, G), (G, R)\}.$$

# Đa biến ngẫu nhiên

## Hàm xác suất đồng thời

**Định nghĩa:** xác suất cùng lúc  $A = a$  và  $B = b$ .

- Với ví dụ bóng: mỗi cặp có xác suất:

$$P(R, R) = \frac{1}{3}, P(R, G) = \frac{1}{3}, P(G, R) = \frac{1}{3}.$$

- Luôn thỏa:  $0 \leq P(A, B) \leq P(A)$  và  $\sum_{(a,b) \in \Omega} P(A = a, B = b) = 1$ .

# Đa biến ngẫu nhiên

## Phân phối riêng (marginal)

**Khái niệm:** Phân phối riêng (marginal) cho biết xác suất của một biến, bất chấp giá trị của biến kia.

- Công thức:**

$$P(A = a) = \sum_{v \in \text{Range}(B)} P(A = a, B = v), \quad P(B = b) = \sum_{u \in \text{Range}(A)} P(A = u, B = b).$$

- Ý nghĩa:** Khi không quan tâm đến biến thứ hai, ta lấy tổng xác suất của tất cả các giá trị có thể của biến đó. Đây là cách tính xác suất biên từ phân phối xác suất chung.

**Ví dụ bóng:**

- Xác suất quả đầu tiên là đỏ ( $A=R$ ):

$$P(A = R) = P(R, R) + P(R, G) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

- Xác suất quả thứ hai là đỏ ( $B=R$ ):

$$P(B = R) = P(R, R) + P(G, R) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

- Giải thích: Dù không biết quả đầu tiên là G hay R, ta vẫn tính được xác suất.

# Đa biến ngẫu nhiên

## Có điều kiện & Bayes

**Có điều kiện:**

$$P(B = b \mid A = a) = \frac{P(A = a, B = b)}{P(A = a)}.$$

*Ví dụ bóng:*

$$P(B = R \mid A = R) = \frac{P(R, R)}{P(A = R)} = \frac{1/3}{2/3} = \frac{1}{2}.$$

**Định lý Bayes:**

$$P(A = a \mid B = b) = \frac{P(B = b \mid A = a) P(A = a)}{P(B = b)}.$$

*Ứng dụng:* Trong thống kê Bayes,  $P(H)$  là tiên nghiệm,  $P(H \mid E)$  là hậu nghiệm.

# Ví dụ minh họa Định lý Bayes

**Bài toán:** Giả sử:

- 1% dân số mắc bệnh A:  $P(H) = 0.01$ ,  $P(\neg H) = 0.99$
- Nếu có bệnh: xét nghiệm dương tính với xác suất  $P(E | H) = 0.99$
- Nếu không bệnh: xét nghiệm dương tính sai với xác suất  $P(E | \neg H) = 0.05$

**Công thức tính xác suất tổng quát:**

$$P(E) = P(E | H) \cdot P(H) + P(E | \neg H) \cdot P(\neg H)$$

$$P(E) = 0.99 \cdot 0.01 + 0.05 \cdot 0.99 = 0.0594$$

**Áp dụng định lý Bayes:**

$$P(H | E) = \frac{P(E | H) \cdot P(H)}{P(E)} = \frac{0.99 \cdot 0.01}{0.0594} \approx 0.167$$

**Kết luận:** Dù xét nghiệm dương tính, xác suất thật sự mắc bệnh chỉ khoảng 16.7%.

# Đa biến ngẫu nhiên

## Độc lập & Độc lập có điều kiện

### Độc lập

A và B độc lập khi  $P(A = a, B = b) = P(A = a)P(B = b)$  với mọi  $(a, b)$ .

Ví dụ: Tung 2 đồng xu:

$$P(H, T) = P(H)P(T) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

### Độc lập có điều kiện

A và B độc lập khi biết C nếu

$$P(A, B \mid C) = P(A \mid C)P(B \mid C).$$

Ví dụ: Biết trời mưa (Z), việc mang dù (X) và mặc áo mưa (Y) độc lập:

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z).$$

## 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  $P(D_1 | H)$**

<b>Xác suất có điều kiện</b>	$H = 1$	$H = 0$
$P(D_1 = 1   H)$	1	0.01
$P(D_1 = 0   H)$	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)  $P(D_1 = 1 \mid H = 0) = 0.01$  (tỷ lệ dương tính giả = 1%)

Giả sử tỷ lệ nhiễm HIV trong dân số là:  $P(H = 1) = 0.0015$

Vậy tỷ lệ không bị nhiễm HIV trong dân số là:

$$P(H = 0) = P(H) - P(H = 1) = 1 - 0.0015 = 0.9985$$

## Tính xác suất biên

Để áp dụng định lý Bayes, trước hết ta cần sử dụng quy tắc biên hóa để tính:

$P(D_1 = 1) = P(D_1 = 1, H = 1) + P(D_1 = 1, H = 0)$  (áp dụng Luật xác suất toàn phần- Law of Total Probability, vì  $H = 0$  và  $H = 1$  là hai biến cố đầy đủ và loại trừ lẫn nhau)

Áp dụng công thức nhân xác suất có điều kiện:  $P(D_1 = 1) = P(D_1 = 1, H = 1) + P(D_1 = 1, H = 0) = P(D_1 = 1, H = 1)P(H = 1) + P(D_1 = 1, H = 0)P(H = 0)$  Thay giá trị vào công thức, ta có:  $1 * 0.0015 + 0.01 * 0.9985 = 0.0015 + 0.00985 = 0.011485$

## 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:

$$\frac{1 * 0.0015}{0.011485} \approx 0.1306$$

Trước khi test, trực giác chỉ ra có 0.15% khả năng nhiễm bệnh. Sau khi xét nghiệm dương tính, xác suất nhiễm bệnh được cập nhật nhưng vì test cũng có 1% dương tính giả, nên trong số rất nhiều người khỏe, vẫn có khá nhiều “dương tính giả”. Kết quả: chỉ  $\approx 13\%$  trong các trường hợp dương tính là thật sự có bệnh.

## 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. Ví dụ, 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

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). Ví dụ, 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. Ví dụ: Nếu một đồng xu có 50% ra sấp (giá trị 0) và 50% ra ngửa (giá trị 1), thì:  $E[X] = 0 * 0.5 + 1 * 0.5 = 0.5 \Rightarrow$  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  $p(x)$ . Điều này tương tự như tính trung bình có trọng số, với trọng số là xác suất liên tục trên mỗi giá trị  $x$ .



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

Tương tự, Khi giá trị đầu vào của phương trình  $f(x)$  là một biến ngẫu nhiên cho trước theo phân phối  $P$  với các giá trị  $x$  khác nhau, kỳ vọng của  $f(x)$  sẽ được tính theo công thức:

Công thức rời rạc:

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

Công thức liên tục:

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

Giải thích: Thay vì tính trung bình giá trị của  $X$ , ta tính trung bình của hàm số  $f(X)$ . Đây là cách đánh giá các biến đổi phi tuyến của  $X$  – ví dụ như log, bình phương, hàm lợi ích (utility), v.v. Cực kỳ quan trọng trong các bài toán như:

- Kỳ vọng lợi nhuận trong tài chính
- Kỳ vọng mất mát trong học máy
- Tính entropy, information gain trong học thống kê

## Phương sai

Phương sai là một thước đo định lượng khoảng dao động quanh giá trị kỳ vọng của một biến ngẫu nhiên.

Phương sai của biến ngẫu nhiên được tính theo công thức:

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

Do không thể lấy kỳ vọng của hiệu:

$$E[X - E[X]]$$

Vì:

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

Nên không thể dùng hiệu này để đo mức dao động; phải bình phương lên để loại bỏ dấu âm.

## Phương sai

Công thức phương sai của một hàm số của biến ngẫu nhiên được định nghĩa tương tự:

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

Giải thích: Cách tính tương tự, nhưng áp dụng cho hàm số  $f(X)$ . Hữu ích khi:  $f(X)$  là biến đổi phi tuyến của  $X$  và phân tích rủi ro hoặc biến động sau khi qua một mô hình

## Độ lệch chuẩn

Độ lệch chuẩn luôn có thể suy ra bằng cách lấy căn bậc hai của phương sai:

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

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:  $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

# Bài tập số 1 I

Hãy cho một ví dụ về việc khi quan sát nhiều dữ liệu hơn có thể làm giảm lượng không chắc chắn về kết quả đầu ra xuống mức thấp một cách tùy ý. (Nguồn: Bài tập 1 Phần 2.6.8)

- Gọi  $p$  là xác suất tung đồng xu được mặt hình  $p \in (0, 1)$ .
- Tung 1 lần  $\rightarrow p=1$ ; 10 lần  $\rightarrow p=0.7$ ; 100 lần  $\rightarrow p=0.58$ ;...; 1.000.000 lần  $\rightarrow p=0,50035$ .
- **Giảm độ không chắc chắn:** Theo Luật số lớn, *Law of large numbers*, khi số lần thử tăng lên giá trị trung bình của kết quả thu được sẽ gần với giá trị kỳ vọng và tỷ lệ mẫu của mặt hình sẽ hội tụ về xác suất thực  $p$ .
- **Mức thấp tùy ý:** Khoảng tin cậy cho  $p$  là phạm vi giá trị của  $p$  thực thuộc về. Khi  $n \rightarrow \infty$  thì khoảng tin cậy tiến tới 0 hay ta có thể giảm sự không chắc chắn của xác suất  $p$  xuống mức thấp tùy ý gần với giá trị kỳ vọng.

## Bài tập số 2 I

Hãy cho một ví dụ về việc giảm lượng bất định đến một điểm nào đó khi tăng quan sát nhiều dữ liệu hơn. Giải thích vì sao chúng ta muốn xác định được điểm kỳ vọng này. (*Nguồn: Bài tập 2 Phần 2.6.8*)

- Dùng thước nhựa đo có chia vạch đến mm để đo chiều dài một cái bàn. Kết quả của sự không chắc chắn là chiều dài thực, chính xác của cái bàn.
- **Giảm độ không chắc chắn:** Theo Luật số lớn, *Law of large numbers*, khi số lần đo tăng lên thì giá trị trung bình của phép đo sẽ gần với chiều dài thực, nếu chỉ xảy ra lỗi ngẫu nhiên.
- **Lỗi hệ thống và dụng cụ đo:** không thể khắc phục được bằng cách tăng số phép đo.

## Bài tập số 3 I

Chứng minh bằng thực nghiệm sự hội tụ đến giá trị trung bình cho phép tung đồng xu. Tính toán phương sai của ước lượng xác suất nhìn thấy mặt hình sau khi tung  $n$  lần.

- 1. Phương sai tỷ lệ với số lượng quan sát như thế nào?
- 2. Sử dụng bất đẳng thức Chebyshev để giới hạn độ lệch khỏi kỳ vọng.
- 3. Mọi liên quan đến định lý giới hạn tập trung, *central limit theorem*?
- (Nguồn: Bài tập 3 Phần 2.6.8)

## Bài tập số 3 II

Ví dụ về tung đồng xu hai mặt hình và chữ.

- $X_i = 1$  là hiện mặt hình,  $X_i = 0$  là hiện mặt chữ.  $X_i$  là biến ngẫu nhiên độc lập.
- Kỳ vọng:  $E(X_i) = p$
- Phương sai (Bernnoulli):  $Var(X_i) = p.(1 - p)$
- Ước lượng tỷ lệ mẫu có xác suất mặt hình:  $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$
- Ước lượng phương sai mẫu:  $Var(\hat{p}_n) = \frac{p.(1-p)}{n}$



## Bài tập số 3 III

1. Mối quan hệ giữa phương sai và số lượng quan sát.

- Ước lượng phương sai mẫu:  $Var(\hat{p}_n) = \frac{p \cdot (1-p)}{n}$
- Phương sai **tỷ lệ nghịch** với độ lớn của mẫu, nghĩa là khi tăng lượng quan sát lên thì phương sai sẽ nhỏ hay độ chính xác sẽ tăng lên.

## Bài tập số 3 IV

2. Vận dụng bất đẳng thức Chebyshev để chặn độ lệch so với giá trị kỳ vọng.

- Bất đẳng thức Chebyshev:  $P(|Y - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$
- Xét biến ngẫu nhiên  $\hat{p}_n$ , kỳ vọng  $E[\hat{p}_n] = p$ , phương sai  $Var(\hat{p}_n) = \frac{p \cdot (1-p)}{n}$
- $\Rightarrow P(|\hat{p}_n - p| \geq \epsilon) \leq \frac{p \cdot (1-p)}{n \cdot \epsilon^2}$
- Theo *Luật số lớn*, khi  $n \rightarrow \infty$  thì  $\hat{p}_n \rightarrow p$ .

## Bài tập số 3 V

3. Mối liên hệ giữa phương sai và Định lý giới hạn tập trung.

- Định lý giới hạn tập trung, *Central Limit theorem*: Tổng (hoặc trung bình) của số lượng lớn biến ngẫu nhiên, mỗi biến có kỳ vọng, phương sai hữu hạn, sẽ có phân phối chuẩn.
- Xét biến ngẫu nhiên  $\hat{p}_n$ :
- Kỳ vọng:  $E[\hat{p}_n] = p$
- Phương sai:  $Var(\hat{p}_n) = \frac{p \cdot (1-p)}{n}$
- Độ lệch chuẩn:  $SE(\hat{p}_n) = \sqrt{\frac{p \cdot (1-p)}{n}}$

## 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  $x_1, x_2, \dots, x_m$  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  $x_1, x_2, \dots, x_m$  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  $x_1, x_2, \dots, x_m$  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

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:

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

# Bài tập số 4 I

- Phương sai:

Phương sai tổng của hai biến ngẫu nhiên:

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

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

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

Áp dụng:

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

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



# 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 I

(phần 2.6.8)

Given two events with probability  $P(\mathcal{A})$  and  $P(\mathcal{B})$ , 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

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

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 nâng cao 1 - NC1 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 - NC1 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 - NC1 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.*

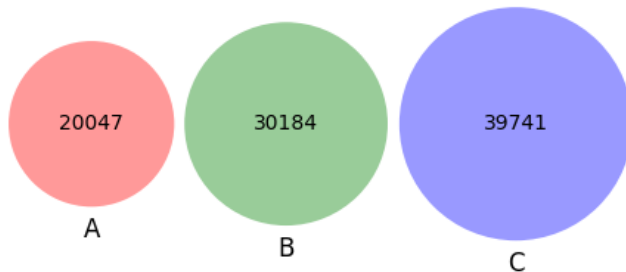
- Giới hạn dưới:

$$P(A \cap B \cap C) \geq \max(0, P(A) + P(B) + P(C) - 2)$$

*Nhỏ nhất khi ba tập gần rời nhau (giao nhỏ nhất).*

# Bài tập nâng cao 1 - NC1 IV

Venn diagram - Case: union\_upper\_bound



==== UNION\_UPPER\_BOUND ====

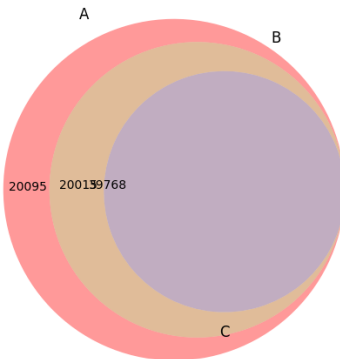
$$P(A \cup B \cup C) = 0.8997 \in [0.4000, 0.9000]$$

$$P(A \cap B \cap C) = 0.0000 \in [0.0000, 0.2000]$$

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 - NC1 V

Venn diagram - Case: union\_lower\_bound

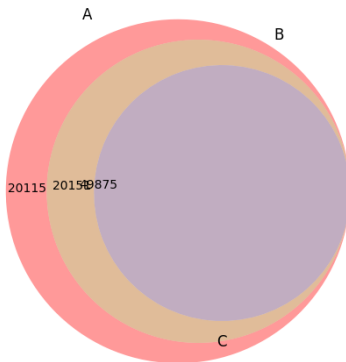


==== UNION\_LOWER\_BOUND ====  
 $P(A \cup B \cup C) = 0.7988 \in [0.8000, 1.0000]$   
 $P(A \cap B \cap C) = 0.3977 \in [0.0000, 0.4000]$

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 - NC1 VI

Venn diagram - Case: inter\_upper\_bound

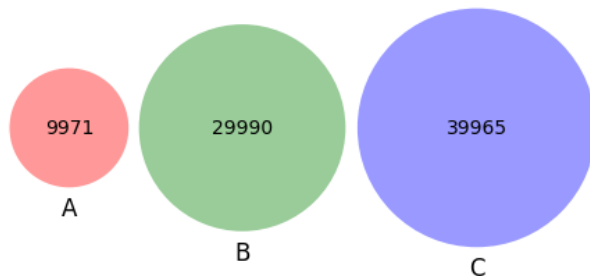


```
==== INTER_UPPER_BOUND ====
P(A ∪ B ∪ C) = 0.9014 ∈ [0.9000, 1.0000]
P(A ∩ B ∩ C) = 0.4988 ∈ [0.1000, 0.5000]
```

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 - NC1 VII

Venn diagram - Case: inter\_lower\_bound



```
==== INTER_LOWER_BOUND ====
P(A ∪ B ∪ C) = 0.7993 ∈ [0.4000, 0.8000]
P(A ∩ B ∩ C) = 0.0000 ∈ [0.0000, 0.1000]
```

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 - NC2 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 - NC2 II

Giả sử mô hình thỏa mãn mỗi quan hệ:

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

$$F \rightarrow L$$

$$F \rightarrow T$$

## Bài tập nâng cao 2 - NC2 III

### Lời giải bài toán NC2

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

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

- 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 - NC2 IV

- 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 - NC3 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 dương tính,  $D_i = 0$  nếu 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  $H$ :
  - $P(D_1, D_2, \dots, D_k \mid H = 1) = P(D_1 \mid H = 1)P(D_2 \mid H = 1)\dots P(D_k \mid H = 1)$
  - $P(D_1, D_2, \dots, D_k \mid H = 0) = P(D_1 \mid H = 0)P(D_2 \mid H = 0)\dots P(D_k \mid H = 0)$

## Bài tập nâng cao 3 - NC3 II

### 6 Chi phí:

- Điều trị đúng chi phí: 0\$
- Điều trị nhầm người không mắc bệnh, bệnh viện phải tốn chi phí: 500\$
- Bỏ sót điều trị người mắc bệnh, bệnh viện phải tốn chi phí: 10,000\$
- 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:

- (A): Dừng lại và điều trị.
- (B): Dừng lại và không điều trị.
- (C): Tiếp tục xét nghiệm lần 3 (lần cuối).

### **Câu hỏi bài toán NC3- OUTPUT:**

- 1 Tính xác suất hậu nghiệm  $P(H = 1|D_1, D_2, \dots, D_k)$  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.

## Bài tập nâng cao 3 - NC3 III

### Lời giải bài toán NC3

**Phần 1:** Tính xác suất hậu nghiệm  $P(H = 1|D_1, D_2, \dots, D_k)$

- Công thức Bayes: Sau  $k$  lần xét nghiệm, với  $s$  lần dương tính:

$$\begin{aligned} P(H = 1|D_1, \dots, D_k) &= \frac{P(D_1, \dots, D_k|H = 1) \cdot P(H = 1)}{P(D_1, \dots, D_k)} \\ &= \frac{P(D_1 = \dots = D_s = 1|H = 1) \cdot P(D_{s+1} = \dots = D_k = 0|H = 1) \cdot P(H = 1)}{P(D_1, \dots, D_k|H = 1) + P(D_1, \dots, D_k|H = 0)} \\ &\quad (D_1 = \dots = D_s = 1, D_{s+1} = \dots = D_k = 0) \end{aligned}$$

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

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

## Bài tập nâng cao 3 - NC3 IV

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

$$\begin{aligned} \Rightarrow P(H = 1 | D_1, \dots, D_k) &= \frac{(0.99)^s (0.01)^{k-s} \cdot 0.0015}{(0.99)^s (0.01)^{k-s} \cdot 0.0015 + (0.10)^s (0.90)^{k-s} \cdot 0.9985} \\ \text{bayes\_posterior}(s, k) &= P(H = 1 | D_1, \dots, D_k) \end{aligned} \quad (1)$$

## Bài tập nâng cao 3 - NC3 V

**Phần 2:** Tính chi phí kỳ vọng của các hành động Tính chi phí kỳ vọng:

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

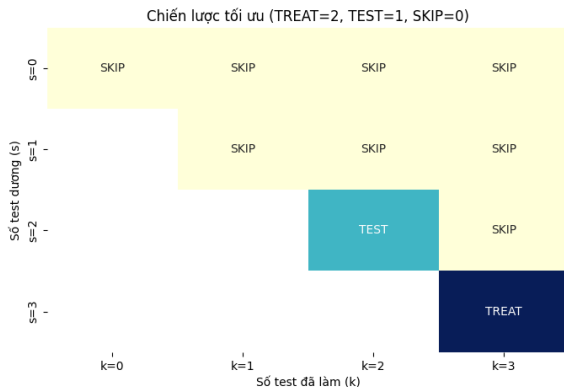
- Giả sử tại bước  $k$ , số dương là  $s$ , gọi hàm:  $p = bayes\_posterior(s, k)$ 
  - Nếu hành động (A): Điều trị thì gọi hàm tính chi phí:  $cost\_A = (1 - p) * 500 + k * 50$ ;
  - Nếu hành động (B): Không điều trị thì gọi hàm tính chi phí:  $cost\_B = p * 10000 + k * 50$ ;
  - Nếu hành động (C) tiếp tục test → xây dựng cây quyết định đệ quy ở Phần 3.



# Bài tập nâng cao 3 - NC3 VI

## Phần 3: Xây dựng thuật toán chọn hành động tối ưu

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

# Kết luận

Nhóm đã thực hiện:

- Trình bày ngắn gọn các kiến thức nền tảng của xác suất thống kê.
- Sự bất định trong machine learning và việc 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](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.

# Q&A

# Thank you very much!

## We are Group 4

Contact us: [nmdai.sdh242@hcmut.edu.vn](mailto:nmdai.sdh242@hcmut.edu.vn)  
[nxhien.sdh242@hcmut.edu.vn](mailto:nxhien.sdh242@hcmut.edu.vn)  
[tdhung.sdh242@hcmut.edu.vn](mailto:tdhung.sdh242@hcmut.edu.vn)  
[ndnminh.sdh232@hcmut.edu.vn](mailto:ndnminh.sdh232@hcmut.edu.vn)  
[ttham.sdh242@hcmut.edu.vn](mailto:ttham.sdh242@hcmut.edu.vn)  
[vuongminhtoan2@gmail.com](mailto:vuongminhtoan2@gmail.com)