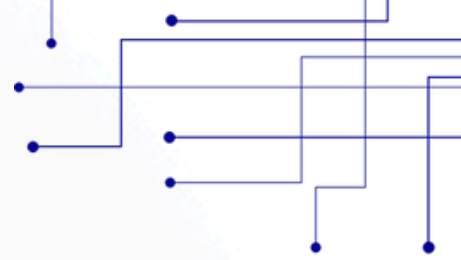


HCMUT EE MACHINE LEARNING & IOT LAB

# Buổi 2

## Cơ sở toán

Presentation By: Trương Thịnh



# Table of Content

**I** Đại số tuyến tính

---

**II** Đạo hàm

---

**III** Xác suất

---



# I. Đại số tuyến tính

# I. Giới thiệu

Giả sử ta có ma trận

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Ta nói một tập hợp các vectơ  $\mathbf{v}_1, \dots, \mathbf{v}_k$  phụ thuộc tuyến tính nếu tồn tại các hệ số  $a_1, \dots, a_k$  đồng thời không bằng 0 sao cho

$$\sum_{i=1}^k a_i \mathbf{v}_i = \mathbf{0}.$$

Hạng của một ma trận  $A$  là số lượng hàng độc lập tuyến tính lớn nhất trong mọi tập con các hàng của ma trận đó

Ma trận chuyển vị

$$A^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

Ma trận đơn vị

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

# II. Tính chất


Cộng

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

Tích ma trận với một số

$$\lambda \cdot \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & \ddots & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{bmatrix}$$

Nhân hai ma trận


$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1j} & \cdots & b_{1s} \\ b_{21} & \cdots & b_{2j} & \cdots & b_{2s} \\ \vdots & & \vdots & & \vdots \\ b_{n1} & \cdots & b_{nj} & \cdots & b_{ns} \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1j} & \cdots & c_{1s} \\ \vdots & & \vdots & & \vdots \\ c_{i1} & \cdots & c_{ij} & \cdots & c_{is} \\ \vdots & & \vdots & & \vdots \\ c_{m1} & \cdots & c_{mj} & \cdots & c_{ms} \end{bmatrix}$$

# III. Tính nghịch đảo

Một ma trận  $A^{-1}$  được gọi là ma trận nghịch đảo nếu

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Chú ý: Một ma trận có các hàng phụ thuộc tuyến tính thì không khả nghịch.

Ví dụ nếu  $A$  là ma trận 2x2

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

thì ma trận nghịch đảo của  $A$  sẽ là

$$\frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

# IV. Định thức

Giả sử ta có ma trận  $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}$

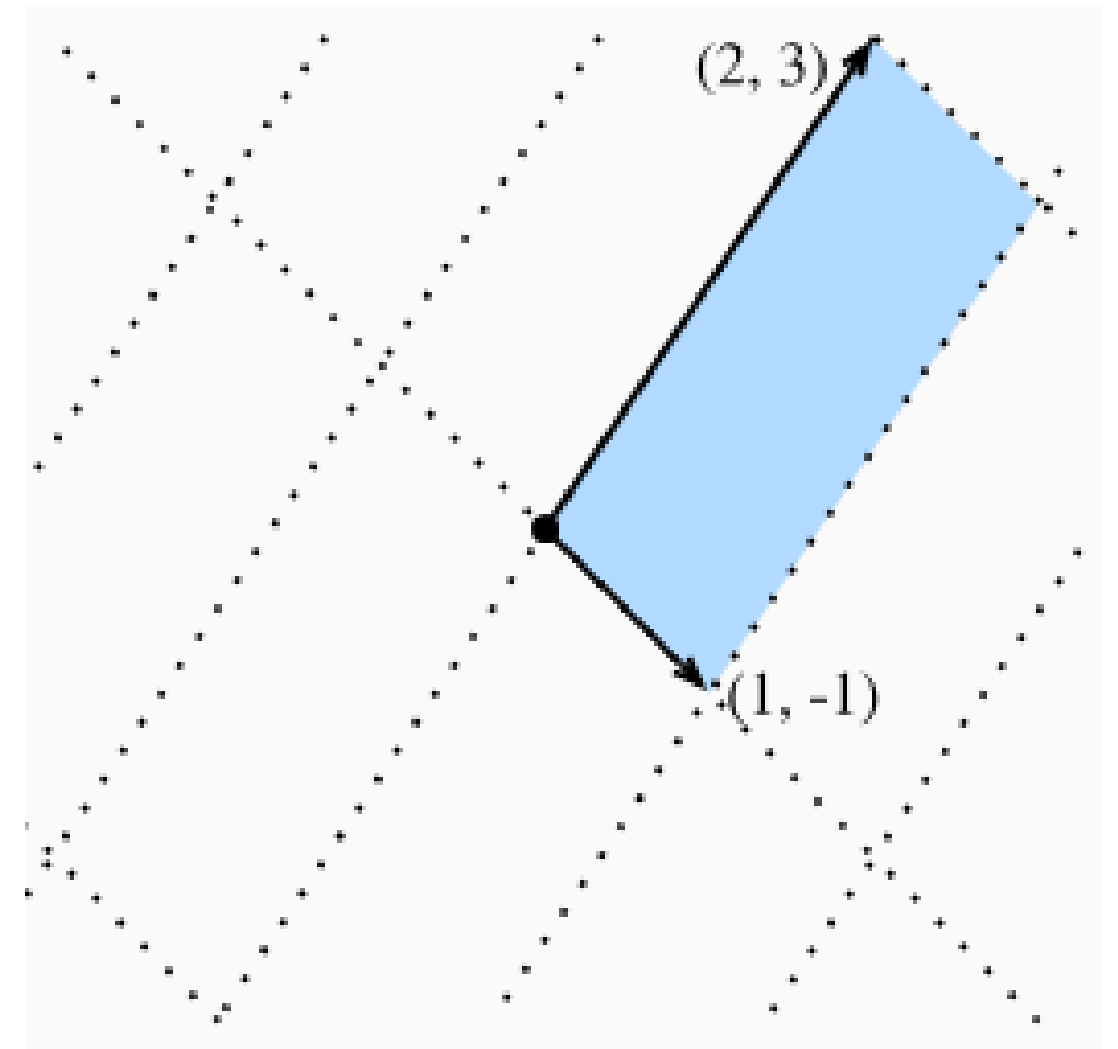
Dễ dàng tính được diện tích hình bình hành bằng 5.

Tổng quát, với  $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

Ta cũng có thể tính ra diện tích của hình bình hành là  **$ad - bc$** . Diện tích này được coi là *định thức*.

Định thức có thể **bằng 0** khi ma trận nén không gian xuống ít chiều hơn. Vì khi đó hình bình hành ban đầu sẽ bị nén xuống thành đường thẳng nên diện tích bằng 0.

Do đó chúng ta suy ra được hệ quả sau: ma trận  $\mathbf{A}$  khả nghịch khi và chỉ khi nó có định thức khác không.



# V. Trị riêng, vecto riêng

Giả sử ta có ma trận

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$$

Nếu ta áp dụng  $\mathbf{A}$  lên bất kỳ vector  $\mathbf{v} = [x, y]^T$  nào, ta nhận được vector  $\mathbf{Av} = [2x, -y]^T$ . Điều này có thể được diễn giải trực quan như sau: kéo giãn vector  $\mathbf{v}$  dài gấp đôi theo phương  $x$ , rồi lấy đối xứng theo phương  $y$ .

Tuy nhiên, sẽ có một vài vector không thay đổi phương, chỉ bị kéo giãn. Ta gọi những vector ấy là *vector riêng* và các hệ số mà chúng giãn ra là *trị riêng*.

Tổng quát, nếu ta tìm được một số  $\lambda$  và một vector  $\mathbf{v}$  mà

$$\mathbf{Av} = \lambda \mathbf{v}.$$

Ta nói rằng  $\mathbf{v}$  là một vector riêng và  $\lambda$  là một trị riêng của  $\mathbf{A}$ .



# V. Trị riêng, vecto riêng

Để tìm trị riêng, ta biến đổi phương trình

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}.$$

Dễ thấy  $(\mathbf{A} - \lambda \mathbf{I})$  phải nén một số chiều xuống không,

=>  $(\mathbf{A} - \lambda \mathbf{I})$  không thể nghịch đảo được nên có định thức bằng không. Do đó, ta có thể tìm các trị riêng bằng cách tìm giá trị  $\lambda$  sao cho  $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$ .

Một khi tìm được các trị riêng, ta có thể giải phương trình  $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}$  để tìm (các) vector riêng tương ứng.

## II. Đạo hàm

# I. Đạo hàm đơn biến

$$\begin{aligned}\frac{df}{dx}(x) &= \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon) - f(x)}{\epsilon} \implies \frac{df}{dx}(x) \approx \frac{f(x + \epsilon) - f(x)}{\epsilon} \\ &\implies \epsilon \frac{df}{dx}(x) \approx f(x + \epsilon) - f(x) \\ &\implies f(x + \epsilon) \approx f(x) + \epsilon \frac{df}{dx}(x).\end{aligned}$$

Cần phải nói rõ hơn về phương trình cuối cùng. Nó cho chúng ta biết rằng nếu ta chọn một hàm số bất kỳ và thay đổi đầu vào một lượng nhỏ, sự thay đổi của đầu ra sẽ bằng với lượng nhỏ đó nhân với đạo hàm.

Bằng cách này, chúng ta có thể hiểu đạo hàm là hệ số tỷ lệ cho biết mức độ biến thiên của đầu ra khi đầu vào thay đổi.

# II. Quy tắc đạo hàm

## Các Đạo hàm phổ biến

Đạo hàm hằng số:  $\frac{d}{dx} c = 0$ .

Đạo hàm hàm tuyến tính:  $\frac{d}{dx} (ax) = a$ .

Quy tắc lũy thừa:  $\frac{d}{dx} x^n = nx^{n-1}$ .

Đạo hàm hàm mũ cơ số tự nhiên:  $\frac{d}{dx} e^x = e^x$ .

Đạo hàm hàm logarit cơ số tự nhiên:  $\frac{d}{dx} \log(x) = \frac{1}{x}$ .

## Các Quy tắc tính Đạo hàm

Quy tắc tổng.  $\frac{d}{dx} (g(x) + h(x)) = \frac{dg}{dx}(x) + \frac{dh}{dx}(x)$ .

Quy tắc tích.  $\frac{d}{dx} (g(x) \cdot h(x)) = g(x) \frac{dh}{dx}(x) + \frac{dg}{dx}(x) h(x)$ .

Quy tắc dây chuyền.  $\frac{d}{dx} g(h(x)) = \frac{dg}{dh}(h(x)) \cdot \frac{dh}{dx}(x)$ .

# III. Gradient Descent

$$f(x + \epsilon) \approx f(x) + \epsilon \frac{df}{dx}(x).$$

$$L(w) \approx L(w_t) + \frac{d}{d(w_t)} \cdot (w - w_t)$$

Mục tiêu: Tìm hướng cập nhật  $w$  để  $J(w)$  giảm nhiều nhất. Ta cần

$$\frac{d}{d(w_t)} \cdot (w - w_t) < 0$$

Giải pháp: Chọn  $w$  sao cho  $(w - w_t)$  ngược hướng với đạo hàm:

$$w = w_t - \eta \cdot \frac{d}{d(w_t)}$$

1. Xét hàm một biến  $f(x) = x^2$ . Hàm này có một điểm cực tiểu duy nhất tại  $x^* = 0$ . Chúng ta đang sử dụng gradient descent (GD) để tìm điểm cực tiểu này, và tại thời điểm  $t$ , ta đến điểm  $x_t = 2$ . Kích thước bước (step size) bao nhiêu sẽ đưa ta đến  $x^*$  tại thời điểm  $t + 1$ ?

A. 0.1.

B. 0.5.

C. 0.2.

D. 0.9.

# IV. Đạo hàm nhiều biến

$$L(w_1 + \epsilon_1, w_2, \dots, w_N) \approx L(w_1, w_2, \dots, w_N) + \epsilon_1 \frac{d}{dw_1} L(w_1, w_2, \dots, w_N).$$

Chúng ta sẽ gọi đạo hàm của một biến trong khi không thay đổi những biến còn lại là đạo hàm riêng (partial derivative), và ký hiệu đạo hàm này là

$$\frac{\partial}{\partial w_1}$$

$$L(w_1 + \epsilon_1, w_2 + \epsilon_2, \dots, w_N + \epsilon_N) \approx L(w_1, w_2, \dots, w_N) + \sum_i \epsilon_i \frac{\partial}{\partial w_i} L(w_1, w_2, \dots, w_N).$$

Đặt  $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_N]^\top$  và  $\nabla_{\mathbf{x}} L = \left[ \frac{\partial L}{\partial x_1}, \dots, \frac{\partial L}{\partial x_N} \right]^\top$

Ta có  $L(\mathbf{w} + \boldsymbol{\epsilon}) \approx L(\mathbf{w}) + \boldsymbol{\epsilon} \cdot \nabla_{\mathbf{w}} L(\mathbf{w})$  Ta gọi vector  $\nabla_{\mathbf{w}} L$  là gradient của  $L$ .

# IV. Đạo hàm nhiều biến

Để thuận tiện, ta giả định hướng của chúng ta có độ dài bằng một và sử dụng  $\theta$  để biểu diễn góc giữa  $\mathbf{v}$  và  $\nabla_{\mathbf{w}}L(\mathbf{w})$

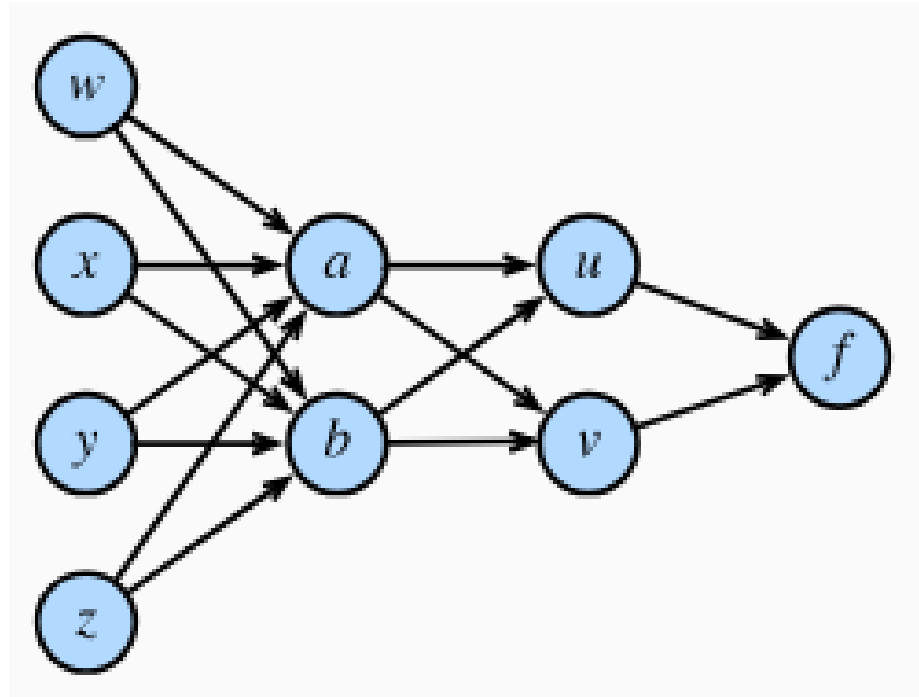
$$L(\mathbf{w} + \mathbf{v}) \approx L(\mathbf{w}) + \mathbf{v} \cdot \nabla_{\mathbf{w}}L(\mathbf{w}) = L(\mathbf{w}) + \|\nabla_{\mathbf{w}}L(\mathbf{w})\| \cos(\theta).$$

Nếu muốn  $L$  giảm càng nhanh, ta sẽ muốn giá trị của biểu thức trên càng âm càng tốt.

Cách duy nhất để chọn hướng đi trong phương trình này là thông qua  $\cos(\theta)$ , giá trị âm nhất của hàm này là  $\cos(\theta) = -1$ , là khi góc giữa vector gradient và hướng cần chọn là  $\pi$  radian hay 180 độ.

=> chọn  $\mathbf{v}$  theo hướng hoàn toàn ngược chiều với  $\nabla_{\mathbf{w}}L(\mathbf{w})$

# V. Backpropagation



$$\begin{aligned}\frac{\partial f}{\partial w} &= \frac{\partial f}{\partial u} \frac{\partial u}{\partial w} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial w}, \\ \frac{\partial u}{\partial w} &= \frac{\partial u}{\partial a} \frac{\partial a}{\partial w} + \frac{\partial u}{\partial b} \frac{\partial b}{\partial w}, \\ \frac{\partial v}{\partial w} &= \frac{\partial v}{\partial a} \frac{\partial a}{\partial w} + \frac{\partial v}{\partial b} \frac{\partial b}{\partial w}.\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial w} &= \frac{\partial f}{\partial a} \frac{\partial a}{\partial w} + \frac{\partial f}{\partial b} \frac{\partial b}{\partial w}, \\ \frac{\partial f}{\partial a} &= \frac{\partial f}{\partial u} \frac{\partial u}{\partial a} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial a}, \\ \frac{\partial f}{\partial b} &= \frac{\partial f}{\partial u} \frac{\partial u}{\partial b} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial b}.\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial a} \frac{\partial a}{\partial x} + \frac{\partial f}{\partial b} \frac{\partial b}{\partial x}, \\ \frac{\partial f}{\partial y} &= \frac{\partial f}{\partial a} \frac{\partial a}{\partial y} + \frac{\partial f}{\partial b} \frac{\partial b}{\partial y}, \\ \frac{\partial f}{\partial z} &= \frac{\partial f}{\partial a} \frac{\partial a}{\partial z} + \frac{\partial f}{\partial b} \frac{\partial b}{\partial z}.\end{aligned}$$



### **III. Xác suất**

# I. Tiên đề

Xác suất có thể được xem là một hàm số ánh xạ một tập hợp các sự kiện tới một số thực. Xác suất của sự kiện  $A$  trong không gian mẫu  $S$ , được kí hiệu là  $P(A)$ , phải thoả mãn những tính chất sau:

- Với mọi sự kiện  $A$ , xác suất của nó là không âm, tức là:  $P(A) \geq 0$
- Xác suất của toàn không gian mẫu luôn bằng 1, tức:  $P(S)=1$
- Đối với mọi dãy sự kiện có thể đếm được  $A_1, A_2, \dots$  xung khắc lẫn nhau, xác suất có ít nhất một sự kiện xảy ra sẽ là tổng của những giá trị xác suất riêng lẻ, hay

$$P(\bigcup_{i=1}^{\infty} \mathcal{A}_i) = \sum_{i=1}^{\infty} P(\mathcal{A}_i)$$

## II. Xác suất

Xác suất để  $B = b$ , với điều kiện  $A = a$  đã xảy ra là

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

Nhân lên ta có  $P(A, B) = P(B \mid A)P(A)$

Định lý Bayes

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

Hai biến ngẫu nhiên  $A$  và  $B$  độc lập nghĩa là việc một sự kiện của  $A$  xảy ra không tiết lộ bất kỳ thông tin nào về việc xảy ra một sự kiện của  $B$ . Trong trường hợp này

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = P(A) \text{ tương đương với } P(A, B) = P(A)P(B).$$

Tương tự, cho một biến ngẫu nhiên  $C$  khác, hai biến ngẫu nhiên  $A$  và  $B$  là độc lập có điều kiện khi và chỉ khi  $P(A, B \mid C) = P(A \mid C)P(B \mid C)$

# III. Naive Bayes

## Định lý Bayes

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

Xét bài toán classification với  $C$  classes  $1, 2, \dots, C$ . Ta có thể xác định class của điểm dữ liệu  $\mathbf{x}$  bằng cách chọn ra class có xác suất cao nhất

$$c = \arg \max_{c \in \{1, \dots, C\}} p(c | \mathbf{x}) = \arg \max_c \frac{p(\mathbf{x} | c)p(c)}{p(\mathbf{x})} = \arg \max_c p(\mathbf{x} | c)p(c)$$

$$p(\mathbf{x} | c) = p(x_1, x_2, \dots, x_d | c) = \prod_{i=1}^d p(x_i | c)$$

$$c = \arg \max_{c \in \{1, \dots, C\}} p(c) \prod_{i=1}^d p(x_i | c)$$

# III. Naive Bayes

Hãy tìm label cho điểm dữ liệu (Height = Small, Hair = Dark, Eyes = Brown)

Height	Hair	Eyes	Attractive?
Small	Blonde	Brown	No
Tall	Dark	Brown	No
Tall	Blonde	Blue	Yes
Tall	Dark	Blue	No
Small	Dark	Blue	No
Tall	Red	Blue	Yes
Tall	Blonde	Brown	No
Small	Blonde	Blue	Yes

Khi d lớn, tính toán có thể gặp sai số. Để giải quyết việc này, ta sẽ dùng log như sau:

$$c = \arg \max_{c \in \{1, \dots, C\}} = \log(p(c)) + \sum_{i=1}^d \log(p(x_i|c))$$

# IV. Kỳ vọng và Phương sai

Kỳ vọng (hay trung bình) của một biến ngẫu nhiên  $X$ , được ký hiệu là

$$\mu_X = E[X] = \sum_i x_i p_i$$

Một vài tính chất của kỳ vọng:

- Với bất kỳ biến ngẫu nhiên  $X$  và hai số  $a$  và  $b$  nào,  $\mu_{aX+b} = a\mu_X + b$ .
- Với hai biến ngẫu nhiên  $X$  và  $Y$ ,  $\mu_{X+Y} = \mu_X + \mu_Y$ .

Trong nhiều trường hợp, chúng ta muốn đo độ lệch của biến ngẫu nhiên  $X$  so với kỳ vọng của nó. Đại lượng này có thể được đo bằng phương sai

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

Một vài tính chất của phương sai:

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

# V. Độ lệch chuẩn, hiệp phương sai

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

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.

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

# THANK YOU

## CONTACT US

-  403.1 H6, BKHCM Campus 2
-  [mliandiotlab@gmail.com](mailto:mliandiotlab@gmail.com)
-  [mliotlab.github.io](https://mliotlab.github.io)
-  [facebook.com/hcmut.ml.iot.lab](https://facebook.com/hcmut.ml.iot.lab)
-  [youtube.com/@mliotlab](https://youtube.com/@mliotlab)