# CHƯƠNG 1: MỞ ĐẦU

**Lưu Giang Nam**

Bộ môn Ứng dụng Tin học
Khoa Toán - Tin học
Trường Đại học KHTN, ĐHQG TPHCM

**09/2025**

**Nhận dạng mẫu (Pattern Recognition):** Là quá trình *tự động phát hiện các quy luật trong dữ liệu* và *sử dụng các quy luật đó để thực hiện hành động* – chẳng hạn như **phân loại dữ liệu vào các nhóm khác nhau**.

**Ví dụ:** Nhận dạng chữ viết tay.

- **Đầu vào (Input):** một vector **x** gồm các giá trị điểm ảnh (pixel values).
- **Đầu ra (Output):** một chữ số từ 0 đến 9.

# Nhận dạng mẫu và Học máy



**Học máy (Machine Learning):** Một tập lớn các vector đầu vào $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, gọi là *tập huấn luyện (training set)*, được sử dụng để *điều chỉnh các tham số của một mô hình thích nghi (adaptive model)*. Loại (category) của một vector đầu vào được biểu diễn bằng vector mục tiêu (target vector) $\mathbf{t}$.
- **Kết quả của thuật toán học máy:** $\mathbf{y}(\mathbf{x})$, trong đó đầu ra $\mathbf{y}$ được mã hóa dưới dạng các vector mục tiêu.

# Mục lục



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHTN
ĐHQG-HCM

# Phần 1

# Ví dụ về Polynomial curve fitting

# Sự Khớp Đường Cong Đa Thức

- Cho tập huấn luyện (training set) với $N$ quan sát.:

$$\mathbf{x} = (x_1, x_2, \ldots, x_N)^T, \quad \mathbf{t} = (t_1, t_2, \ldots, t_N)^T$$

- Nhiệm vụ: khai thác tập huấn luyện để dự đoán $\hat{t}$ khi có input mới $\hat{x}$. Ta cần tìm một hàm $y$ để ánh xạ $x \mapsto t$. Tuy nhiên, dữ liệu thường chứa *nhiễu* (noise) nên việc tìm hàm chính xác là không khả thi.
- Một cách tiếp cận trực quan: **Polynomial curve fitting** (Khớp đường cong đa thức). $\Rightarrow$ Tìm phương trình đa thức sao cho khớp với các điểm trong tập huấn luyện.

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{i=0}^{M} w_i x^i \quad (1)$$

trong đó: $\mathbf{w} = (w_0, w_1, w_2, \ldots, w_M)^T$ là vector hệ số của hàm số, và $M$ là bậc của đa thức.

# Sự Khớp Đường Cong Đa Thức

- Để tìm được đường mong muốn thì với mỗi điểm $x_i$ trong tập huấn luyện (training set), sai lệch giữa $y = f(x_i)$ và $t_i$ là nhỏ nhất.
- Ta sẽ xây dựng hàm lỗi bằng nguyên lý tổng bình phương sai số (**SSE**):

$$E(\mathbf{w}) = \sum_{n=1}^{N} \left(y(x_n, \mathbf{w}) - t_n\right)^2 \quad (2)$$

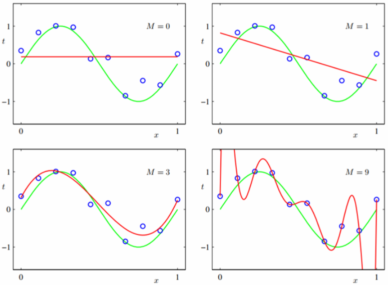
- Phần còn lại là cực tiểu hóa hàm $E(\mathbf{w})$, tức là tìm $\mathbf{w}$ để hàm lỗi $E(\mathbf{w})$ đạt giá trị nhỏ nhất. Ta có thể dùng hàm lỗi $\frac{1}{2}E(\mathbf{w})$ để thuận tiện cho việc đạo hàm để tìm $\mathbf{w}$.
- Việc cực tiểu hóa hàm lỗi $E(\mathbf{w})$ sẽ cho ta một nghiệm duy nhất $\mathbf{w}^*$ và kết quả đa thức mong muốn là

$$f(x) = y(x, \mathbf{w}^*) \quad (3)$$

# Sự Khớp Đường Cong Đa Thức

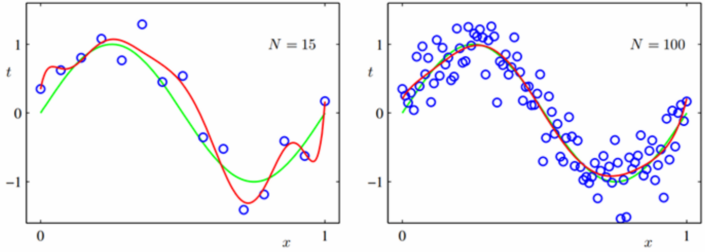
**Vấn đề còn lại là chọn bậc của đa thức như thế nào?**

- Ví dụ: Hàm cần tìm là $sin(2\pi x)$




- Như vậy, ta chỉ cần tìm M vừa đủ
- **Mục tiêu của huấn luyện** là *tổng quát hóa*, chứ không phải là *overfit* tập dữ liệu huấn luyện.

# Làm thế nào để giảm over-fit?




- Tuy nhiên, khi tăng số lượng mẫu huấn luyện, mô hình $M = 9$ không còn đủ khả năng khớp toàn bộ dữ liệu, giúp giảm hiện tượng *overfitting*.
- Số lượng điểm dữ liệu không nên ít hơn (5 - 10)lần số lượng tham số trong mô hình. Nên chọn độ phức tạp của mô hình dựa trên độ phức tạp của vấn đề đang giải quyết.

Phần 2

# Lý Thuyết Xác Suất

# Phần 2

# Lý Thuyết Xác Suất

## Mục 1: Xác suất và xác suất có điều kiện

# Xác suất của một sự kiện

- Lý thuyết xác suất (Probabilistic Theory) là nền tảng cho nhận dạng mẫu, sẽ cho ta một framework để diễn đạt sự tương đối theo cách chính xác và định lượng.
- Xác suất của một sự kiện được định nghĩa là tỷ lệ mà số lần sự kiện đó xảy ra so với tổng số lần thử nghiệm (Số lần thử nghiệm có thể đạt đến vô cùng).
- Gọi $A$ là một sự kiện, công thức $\mathbb{P}(A)$ là xác suất sự kiện $A$ xảy ra, với $0 \leq \mathbb{P}(A) \leq 1$ hay $\mathbb{P}(A) \in [0, 1]$.

# Xác suất của một sự kiện

- $\mathbb{P}(A) = 0$ i.e sự kiện $A$ chắc chắn không xảy ra.
- $\mathbb{P}(A) = 1$ i.e sự kiện $A$ chắc chắn xảy ra.
- Ta viết $\mathbb{P}(\bar{A})$ để chỉ xác suất sự kiện $A$ không xảy ra. Từ đó, ta có $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.

# Biến Ngẫu Nhiên

- Cho hai biến ngẫu nhiên $X, Y$. Trong đó, $X$ có thể nhận giá trị $x_i, i = 1, 2, \dots, M$ và $Y$ có thể nhận giá trị $y_j, j = 1, 2, \dots, L$. Gọi $N$ là tổng số lần thử trong đó ta lấy mẫu cả hai biến $X, Y$ và $n_{i,j}$ là số lần thử mà trong đó $X = x_i$ và $Y = y_j$. Gọi $c_i$ là số lần thử mà $X = x_i$ ( không quan tâm $Y$ có giá trị nào) và tương tự $r_j$ là số lần thử mà $Y = y_j$.
- Xác suất $X = x_i$ và $Y = y_j$ được gọi là **xác suất đồng thời** (joint probability) được định nghĩa là

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{4}$$

- Quy tắc cộng của xác suất (sum rule of probability) đôi khi cũng có thể gọi là xác suất biên (marginal probability).

$$\mathbb{P}(X = x_i) = \sum_{j=1}^{L} \mathbb{P}(X = x_i, Y = y_j) \tag{5}$$

# Xác suất có điều kiện

- Xác suất biến ngẫu nhiên $Y$ nhận giá trị $y_j$ khi biết biến ngẫu nhiên $X$ nhận giá trị $x_i$ được gọi là xác suất có điều kiện của $Y = y_j$ cho trước $X = x_i$, được định nghĩa

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \text{ và } \mathbb{P}(X = x_i) = \frac{c_i}{N} \tag{6}$$

Khi đó ta có:

$$\begin{aligned}\mathbb{P}(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} \\ &= \mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i)\end{aligned} \tag{7}$$

- Đây là **quy tắc nhân của xác suất** (product rule of probability) . Như vậy, ta có công thức

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i) \tag{8}$$

# Công Thức Bayes

- Chú ý là $\mathbb{P}(X, Y) = \mathbb{P}(Y, X)$
- Biến đổi một chút, ta được công thức Bayes:

$$\mathbb{P}(Y \mid X) = \frac{\mathbb{P}(X \mid Y)\,\mathbb{P}(Y)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X \mid Y)\,\mathbb{P}(Y)}{\sum_{y} \mathbb{P}(X \mid y)\,\mathbb{P}(y)} \tag{9}$$

Trong đó,

- $\mathbb{P}(Y|X)$ là hậu nghiệm (posterior)
- $\mathbb{P}(Y)$ là tiên nghiệm (prior)
- $\mathbb{P}(X|Y)$ là hàm hợp lí (likelihood)
- $\sum_Y \mathbb{P}(Y|X)\mathbb{P}(X)$ là hằng số chuẩn hóa (normalization)

Phần 2

# Lý Thuyết Xác Suất

## Mục 2: Hàm Mật độ Xác Suất

# Hàm mật độ xác suất

- Xét một biến $x$ mang giá trị thực, xác suất để $x$ nằm trong vùng giá trị $(x, x+\delta)$ được tính bằng công thức

$$p(x).\delta x, \;\; \delta x \to 0 \tag{10}$$

với $p(x)$ được gọi là hàm mật độ xác suất (probability density) của $x$.

- Xác suất $X$ nằm trong khoảng $(a, b)$ được tính bởi

$$\mathbb{P}(X \in (a, b)) = \int_a^b p(x)dx \tag{11}$$

- Vì xác suất luôn không âm và giá trị của phải nằm trong một khoảng thực nào đó, nên hàm mật độ xác suất $p(x)$ phải thỏa mãn các điều kiện sau.

$$p(x) \geq 0, \quad \int_{-\infty}^{+\infty} p(x)dx = 1 \tag{12}$$

# Hàm mật độ xác suất



- Xét hàm mật độ xác suất $p_x(x)$ biến đổi thành $p_y(y)$ thì những điểm quan sát nằm trong khoảng $(x + \delta x)$ cũng sẽ biến đổi thành $(y + \delta y)$. Ta có:

$$p_x(x)\,\delta x \;=\; p_y(y)\,\delta y \quad \Leftrightarrow \quad p_y(y) = p_x(g(y))g'(y) \tag{13}$$

- Xác suất để $X$ nằm trong khoảng $(-\infty, z)$ là

$$P(z) = \mathbb{P}(X \leq z) = \int_{-\infty}^{z} p(x)dx \tag{14}$$

- Khi đó ta có $p(x) = P'(x)$
- Với vector $\mathbf{x} = (x_1, \ldots, x_D)$, $(x_1, \ldots, x_n)$ là biến ngẫu nhiên liên tục thì hàm mật độ xác suất chung có công thức là:

$$p(\mathbf{x}) = p(x_1, \ldots, x_D) \tag{15}$$

# Phần 2

# Lý Thuyết Xác Suất

## Mục 3: Kỳ Vọng Và Hiệp Phương Sai

# Kỳ Vọng

- Giá trị trung bình của một hàm số $f(x)$ dưới dạng một hàm phân phối xác suất $p(x)$ được gọi là kỳ vọng của $f(x)$ và được ký hiệu là $\mathbb{E}[f]$.
- Nếu là một phân phối rời rạc thì $\mathbb{E}[f]$ được định nghĩa bởi công thức

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{16}$$

item Nếu là biến liên tục thì kỳ vọng được biểu diễn dưới dạng một tích phân theo hàm mật độ tương ứng

$$\mathbb{E}[f] = \int p(x)f(x)dx \tag{17}$$

- Ta cũng có thể xét kỳ vọng có điều kiện theo một phân phối có điều kiện như sau

$$\mathbb{E}[f|y] = \sum_x p(x|y)f(x) \tag{18}$$

# Phương sai



- Phương sai của một hàm $f(x)$ được định nghĩa

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \tag{19}$$

- Phương sai của $f(x)$ cung cấp một thước đo về mức độ biến thiên của $f(x)$ quanh giá trị trung bình $\mathbb{E}[f(x)]$ của nó.
- Khai triển hằng đẳng thức, công thức trên có thể viết lại thành

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{20}$$

- Cho hai biến ngẫu nhiên $x, y$, hiệp phương sai được định nghĩa như sau

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\ &= \mathbb{E}_{xy}[x, y] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{21}$$

Điều này biểu thị mức độ mà $x$ và $y$ biến thiên cùng nhau. Nếu $x, y$ độc lập, thì hiệp phương sai của chúng bằng 0.

# Hiệp Phương Sai



- Trong trường hợp $\mathbf{x}, \mathbf{y}$ là hai vector biến ngẫu nhiên, hiệp phương sai là một ma trận

$$\begin{aligned} \operatorname{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T])] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \end{aligned} \tag{22}$$

- Nếu ta xét hiệp phương sai giữa các thành phần của một vector $\mathbf{x}$ với nhau thì ta sử dụng ký hiệu đơn giản hơn $\operatorname{cov}[\mathbf{x}] = \operatorname{cov}[\mathbf{x}, \mathbf{x}]$

# Phần 2

# Lý Thuyết Xác Suất

## Mục 4: Phân Phối Gauss

# Phân Phối Gauss



- Phân phối Gauss hay còn gọi là Phân phối Chuẩn là một trong những phân phối xác suất quan trọng cho biến liên tục
- Trong trường hợp một biến thực duy nhất $x$, phân phối Chuẩn được định nghĩa như sau

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu^2)}{2\sigma^2}} \tag{23}$$

- Phân phối này được xác định bởi hai tham số : $\mu$ gọi là giá trị trung bình (mean), và $\sigma^2$ được gọi là phương sai (variance) còn $\sigma$ được gọi là độ lệch chuẩn (standard deviation).
- Nghịch đảo của phương sai, viết là $\beta = \frac{1}{\sigma^2}$, được gọi là độ chính xác (precision).
- Phân phối Gauss thỏa mãn tính chất

$$\mathcal{N}(x|\mu,\sigma^2) > 0, \quad \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)dx = 1 \tag{24}$$

# Phân Phối Gauss



- Kỳ vọng của hàm theo biến **x** dưới phân phối Gauss được tính

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \mathrm{d}x = \mu \tag{25}$$

- Tương tự với kỳ vọng bậc 2

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \mathrm{d}x = \mu^2 + \sigma^2 \tag{26}$$

- Từ đó ta có thể suy ra phương sai của $x$:

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \tag{27}$$

- Phân phối Gauss xác định với vectơ **x** có $\mathcal{D}$ chiều trong biến ngẫu nhiên liên tục là:

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})} \tag{28}$$

# Phân Phối Gauss

Giả sử ta có một tập dữ liệu quan sát $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$ biểu diễn $N$ lần quan sát của biến vô hướng $x$. Giả sử các quan sát được rút ra độc lập từ phân phối chuẩn có trung bình $\mu$ và phương sai $\sigma^2$ chưa biết. Nhiệm vụ của ta là xác định các tham số $\mu$ và $\sigma^2$ từ tập dữ liệu này

- Các điểm dữ liệu được rút ra độc lập từ cùng một phân phối được gọi là independent and identically distributed - i.i.d.
- Vì tập dữ liệu $\mathbf{x}$ là i.i.d nên ta có thể viết lại xác suất của tập dữ liệu là

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \quad (29)$$

- Bằng cách sử dụng ước lượng hợp lí cực đại (**MLE**)
- Ta tìm được

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

# Phần 2

# Lý Thuyết Xác Suất

## Mục 5: Trở Lại Bài Toán Khớp Đường Cong

# Khớp đường cong



- Ta xem xét lại ví dụ này dưới góc nhìn xác suất. Từ đó bắt đầu có hướng tiếp cận dưới góc nhìn của Bayes
- Với tập huấn luyện $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$ và tập nhãn $\mathbf{t} = (t_1, t_2, \ldots, t_N)^T$ với $N$ là số giá trị đầu vào. Nhiệm vụ là dự đoán biến mục tiêu $t$ khi cho một giá trị mới $x$ dựa trên tập huấn luyện.
- Ta có thể biểu diễn sự bất định về giá trị của biến mục tiêu bằng một phân phối xác suất. Để làm được điều này, giả sử rằng, với mỗi giá trị $x$ đã cho, giá trị tương ứng của $t$ tuân theo một phân phối Gauss có trung bình bằng giá trị $y(x, \mathbf{w})$ của đa thức được cho ở ví dụ trên. Do đó, ta có

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{30}$$

trong đó $\beta$ là tham số độ chính xác.

# Khớp đường cong

- Ta sử dụng dữ liệu huấn luyện $\{\mathbf{x}, \mathbf{t}\}$ để xác định giá trị của các tham số chưa biết $\mathbf{w}$ và $\beta$ bằng nguyên lý ước lượng cực đại (MLE).

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \tag{31}$$

- Lấy Log-likelihood, ta được

$$\ell(\mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \tag{32}$$

- Việc còn lại là cực đại hóa hàm $\ell$ để tìm $\mathbf{w}_{MLE}$ và $\beta_{MLE}$

# Khớp đường cong



- Một cách tiếp cận khác là ta lấy (Negative log-likelihood) - NLL, lúc này hàm $\ell(\mathbf{w}, \beta)$ - hàm lỗi tổng bình phương (sum-of-squares error function) và việc ta cần làm là cực tiểu hóa $\ell(\mathbf{w}, \beta)$ để tìm $\mathbf{w}_{MLE}$ và $\beta_{MLE}$.
- Ta có thể suy ra $\beta_{MLE}$ từ công thức trên

$$\frac{1}{\beta_{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 \tag{33}$$

- Vì hiện giờ ta có một mô hình xác suất, nên các dự đoán này được biểu diễn dưới dạng phần phối dự đoán (predictive distribution) i.e phân phối xác suất của $t$, thay vì chỉ là các ước lượng điểm đơn lẻ, và phân phối này được thu được bằng cách thay các tham số hợp lí cực đại vào. Từ đó, ta có

$$p(t|x, \mathbf{w}_{MLE}, \beta_{MLE}) = \mathcal{N}(t|y(x, \mathbf{w}_{MLE}), \beta_{MLE}^{-1}) \tag{34}$$

# Hướng tiếp cận Bayes



- Bằng cách sử dụng một phân phối tiên nghiệm cho các hệ số đa thức **w**. Để đơn giản, xét một phân phối Chuẩn có dạng

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} e^{\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)} \quad (35)$$

- Bằng cách sử dụng một phân phối tiên nghiệm cho các hệ số đa thức **w**. Để đơn giản, xét một phân phối Chuẩn có dạng

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} e^{\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)} \quad (36)$$

vói $\alpha$ là độ chính xác của phân phối và $M+1$ là tổng số phần tử trong vector **w** có bậc đa thức thứ $M$.

# Hướng tiếp cận Bayes



- $\alpha$ là tham số điều khiển phân phối của mô hình tham số, được gọi là siêu tham số (hyperparameters)
- Sử dụng định lý Bayes, phân phối hậu nghiệm của **w** tỷ lệ thuận với tích phân phối tiên nghiệm và hàm hợp lí

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \qquad (37)$$

- Bây giờ ta có thể xác định **w** bằng cách tìm giá trị có khả năng xảy ra cao nhất của **w** khi biết dữ liệu, nói cách khác là bằng cách cực đại hóa phân phối hậu nghiệm. Kỹ thuật này được gọi là Ước lượng hậu nghiệm cực đại ( Maximum A Posterior) - MAP.
- Lấy NLL của $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ và $p(\mathbf{w}|\alpha)$, đồng thời kết hợp với công thức trên, ta được

$$-\ln p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto \frac{\beta}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \qquad (38)$$

Phần 3

# Lựa Chọn Mô Hình

# Lựa chọn mô hình



- Trong ví dụ *Polynomial Curve Fitting*, ta đã thấy bậc đa thức quyết định độ phức tạp của mô hình.
- Khi thêm hệ số chuẩn hoá $\lambda$, mô hình có thể kiểm soát được mức độ phức tạp và tránh overfitting.
- Với các mô hình phức tạp hơn như *mixture models* hay *neural networks*, nhiều tham số cùng ảnh hưởng đến độ phức tạp.

# Lựa chọn mô hình



- Mục tiêu là đạt hiệu suất dự đoán tốt nhất trên dữ liệu mới.
- Khi dữ liệu đủ lớn, có thể chia thành **tập huấn luyện**, **tập kiểm định (validation set)** và **tập kiểm tra (test set)**.
- Nếu dữ liệu hạn chế, dùng **đánh giá chéo (cross-validation)** để tận dụng tối đa dữ liệu, ví dụ kỹ thuật *leave-one-out*.

# Lựa chọn mô hình

- Cross-validation đòi hỏi huấn luyện lặp lại nhiều lần, tốn chi phí tính toán.
- Một hướng khác là dùng các **tiêu chí thông tin (Information Criteria)** để thêm hệ số phạt chống overfitting.
- **Akaike Information Criterion (AIC)** chọn mô hình sao cho:

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MLE}}) - M \text{ là lớn nhất.} \tag{39}$$

- Biến thể khác là **Bayesian Information Criterion (BIC)**.

Phần 4

# Lý thuyết ra quyết định (Decision Theory)

# Giới thiệu

- Bài toán ra quyết định: cho $\mathbf{x}$, dự đoán $t$ theo mô hình xác suất $p(\mathbf{x}, t)$.
- Xét trường hợp **phân loại nhị phân**: $t \in \{0, 1\} \Leftrightarrow \{C_1, C_2\}$.
- Đại lượng quan trọng: xác suất hậu nghiệm $p(C_k|\mathbf{x})$.

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}, C_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_{j=1}^{2} p(\mathbf{x}|C_j)p(C_j)} \qquad (40)$$

$\Rightarrow$ Suy luận trung tâm là tìm $p(\mathbf{x}, C_i)$.

- Trực giác: chọn lớp $k$ sao cho $p(C_k|\mathbf{x})$ lớn nhất.

# Ra quyết định – Phân loại nhị phân



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHTN
ĐHQG-HCM

- **Miền quyết định:** $R_i = \{\mathbf{x} : \text{pred}(\mathbf{x}) = C_i\}$.
- **Xác suất sai phân loại:**

$$p(\text{mis}) = \int_{R_1} p(\mathbf{x}, C_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x}, C_1) d\mathbf{x} \tag{41}$$

- Để tối thiểu hóa sai số: gán $\mathbf{x}$ vào $R_1$ nếu

$$p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2) \Leftrightarrow p(C_1|\mathbf{x}) > p(C_2|\mathbf{x}) \tag{42}$$

- Với $K$ lớp:

$$\text{pred}(\mathbf{x}) = \arg\max_k p(C_k|\mathbf{x}) \tag{43}$$

# Ra quyết định có tổn thất (Loss-sensitive decision)

- **Ma trận tổn thất:** $L_{kj}$ – tổn thất khi dự đoán $C_j$ trong khi thực tế là $C_k$.
- Mục tiêu: **tối thiểu hóa tổn thất kỳ vọng**

$$\mathbb{E}[L] = \sum_j \int_{R_j} \left( \sum_k L_{kj} p(\mathbf{x}, C_k) \right) d\mathbf{x} \tag{44}$$

- Giải pháp: với mỗi $\mathbf{x}$, chọn lớp $C_j$ sao cho

$$\sum_k L_{kj} p(C_k|\mathbf{x}) \text{ là nhỏ nhất.} \tag{45}$$

# Ví dụ: Chẩn đoán y khoa

- Hai lớp: $C_1$ = bệnh, $C_2$ = khỏe.
- Ma trận tổn thất: $L = \begin{pmatrix} 0 & 100 \\ 1 & 0 \end{pmatrix}$. Tổn thất kỳ vọng:

$$\mathbb{E}[L] = \int_{R_2} 100\, p(\mathbf{x}, C_1) d\mathbf{x} + \int_{R_1} p(\mathbf{x}, C_2) d\mathbf{x} \tag{46}$$

- Điều này chẩn đoán sai người bệnh là khỏe nguy hiểm gấp 100 lần so với chẩn đoán sai người khỏe là bệnh. Quyết định tối ưu sẽ nghiêng về "bắt bệnh cho chắc" (ưu tiên giảm rủi ro cao hơn).
- Nếu chỉ xét xác suất sai phân loại:

$$p(\text{mis}) = \int_{R_1} p(\mathbf{x}, C_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x}, C_1) d\mathbf{x} \tag{47}$$

$\Rightarrow$ tương ứng với hàm mất mát 0/1: $L = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

# Tuỳ chọn từ chối (Reject Option)



- Với hàm mất mát 0/1:

$$\text{pred}(\mathbf{x}) = \arg\max_k p(C_k|\mathbf{x}) \tag{48}$$

- Ta có:

$$\frac{1}{K} \leq \max_k p(C_k|\mathbf{x}) \leq 1 \tag{49}$$

- Khi $\max_k p(C_k|\mathbf{x}) \approx \frac{1}{K}$, độ tin cậy của dự đoán giảm.
- **Chiến lược từ chối:** chỉ ra quyết định nếu $\max_k p(C_k|\mathbf{x}) > \sigma$.

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHTN
ĐHQG-HCM

# Thiết lập hồi quy (Regression setting)

- Mục tiêu: dự đoán biến liên tục $t \in \mathbb{R}$.
- Hàm mất mát phổ biến:

$$L(t, y(\mathbf{x})) = (y(\mathbf{x}) - t)^2 \tag{50}$$

- Tổn thất kỳ vọng:

$$\mathbb{E}[L] = \int_{\mathcal{X}} \int_{\mathbb{R}} (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t)\, d\mathbf{x} dt \tag{51}$$

- Giải pháp tối ưu:

$$y(\mathbf{x}) = \int_{\mathbb{R}} t\, p(t|\mathbf{x})\, dt \tag{52}$$

- Hàm $y(\mathbf{x})$ biểu diễn **giá trị trung bình có điều kiện** của $t$ khi biết $\mathbf{x}$. Đây là nghiệm trực quan và có ý nghĩa: dự đoán giá trị trung bình của biến mục tiêu dựa trên phân bố có điều kiện.

# Các cách tiếp cận trong lý thuyết quyết định

1. **Mô hình sinh (Generative model)**: Ước lượng $p(\mathbf{x}|C_k)$ và kết hợp với $p(C_k)$ để suy ra $p(C_k|\mathbf{x})$.
2. **Mô hình phân biệt (Discriminative model)**: Trực tiếp học $p(C_k|\mathbf{x})$ – đủ cho bài toán phân loại.
3. **Hàm phân biệt (Discriminant function)**: Học hàm $f(\mathbf{x})$ ánh xạ trực tiếp từ đầu vào sang nhãn lớp.

**Ưu điểm – Nhược điểm:**

- **Generative:** Có thể ước lượng $p(\mathbf{x}) \rightarrow$ phát hiện ngoại lai dễ dàng. Ước lượng $p(\mathbf{x}, C_k)$ phức tạp, đòi hỏi dữ liệu lớn.
- **Discriminative:** Ít yêu cầu dữ liệu hơn; phù hợp cho phân loại trực tiếp. Không mô hình hóa toàn bộ phân bố dữ liệu.
- **Discriminant function:** Bài toán học duy nhất, đơn giản hóa quá trình ra quyết định. Không có thông tin về $p(C_k|\mathbf{x})$, khó kết hợp mô hình hoặc từ chối mẫu.

Phần 5

# Lý thuyết thông tin

# Entropy rời rạc



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN KHTN ĐHQG-HCM

- Xét biến ngẫu nhiên rời rạc $X$.
- Ta cần một độ đo $h(x)$ thể hiện mức độ "ngạc nhiên" khi quan sát $X = x$.
- Yêu cầu tự nhiên:
  - Nếu $p(x)$ nhỏ $\Rightarrow$ $h(x)$ lớn (và ngược lại).
  - $h(x)$ giảm đơn điệu theo $p(x)$.
  - Nếu $X, Y$ độc lập $\Rightarrow$ $h(x, y) = h(x) + h(y)$.

Điều này dẫn đến:

$$h(x) = -\log p(x) \tag{53}$$

Entropy của biến $X$:

$$H[X] = \mathbb{E}[h(X)] = -\sum_x p(x) \log p(x) \tag{54}$$

*(Quy ước: $p \log p = 0$ nếu $p = 0$)*

# Entropy – Các tính chất quan trọng

- $H[X] \geq 0$ vì $p \in [0,1] \Rightarrow p \log p \leq 0$.
- $H[X] = 0$ nếu tồn tại $x$ sao cho $p(x) = 1$.
- Phân bố có entropy lớn nhất là **phân bố đều**.

Tối ưu hoá:

$$\max_{p(x_i)} \left( H[X] + \lambda \left( \sum_i p(x_i) - 1 \right) \right) \tag{55}$$

Giải ra: $p(x_i) = \frac{1}{M} \Rightarrow H[X] = \log M$

$$0 \leq H[X] \leq \log M$$

**Ý nghĩa:** $H[X]$ là giới hạn dưới về số bit cần thiết để mã hoá nhị phân $X$. Phân bố đều $\rightarrow$ không thể nén thêm. Phân bố không đều $\rightarrow$ có thể tối ưu mã hóa.

# Entropy – Biến ngẫu nhiên liên tục



- Với biến liên tục:

$$H[X] = -\int p(x) \ln p(x)\, dx \tag{56}$$

- $p(x)$ có thể $> 1$, nên $H[X]$ có thể âm.
- Ví dụ: $X \sim U(0, \frac{1}{2}) \Rightarrow H[X] = -\ln 2$
- Với ràng buộc $\mu, \sigma$: phân bố có entropy lớn nhất là Gaussian:

$$p(x) = \mathcal{N}(\mu, \sigma^2), \quad H[X] = \frac{1}{2}(1 + \ln(2\pi\sigma^2)) \tag{57}$$

Entropy có điều kiện:

$$H[Y|X] = -\iint p(x,y) \ln p(y|x)\, dxdy \tag{58}$$

Dẫn đến: $H[X,Y] = H[Y|X] + H[X]$

# Kullback–Leibler Divergence

- Đo sự khác biệt giữa hai phân bố $p(x)$ (thật) và $q(x)$ (xấp xỉ):

$$KL(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} \, dx \tag{59}$$

- $KL(p||q) \neq KL(q||p)$
- $KL(p||p) = 0$
- $KL(p||q) \geq 0$

**Chứng minh:** dùng bất đẳng thức Jensen, vì $-\ln(x)$ là hàm lồi:

$$\mathbb{E}[-\ln q(x)/p(x)] \geq -\ln \mathbb{E}[q(x)/p(x)] = 0 \tag{60}$$

**Diễn giải:** KL đo lượng thông tin "thừa" cần thêm khi mã hóa $p(x)$ bằng $q(x)$.

*Cảm Ơn Thầy Và Các Bạn Đã Lắng Nghe!*