

Nội dung môn học

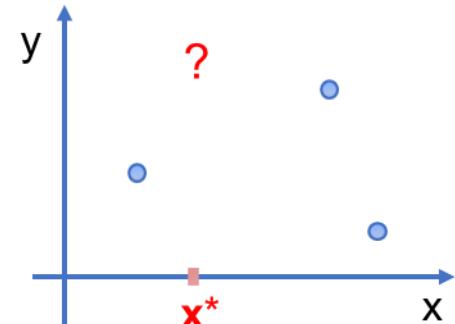
- Lecture 1: Giới thiệu về Học máy và khai phá dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4+5: Phân cụm
- Lecture 6: Phân loại và Đánh giá hiệu năng
- Lecture 7: dựa trên láng giềng gần nhất (KNN)
- Lecture 8: Cây quyết định và Rừng ngẫu nhiên
- **Lecture 9: Học dựa trên xác suất**
- Lecture 10: Mạng nơron (Neural networks)
- Lecture 11: Máy vector hỗ trợ (SVM)
- Lecture 12: Khai phá tập mục thường xuyên và các luật kết hợp
- Lecture 13: Thảo luận ứng dụng trong thực tế

Tại sao cần mô hình hóa xác suất?

- Việc suy diễn từ dữ liệu thường không chắc chắn
- Lý thuyết xác suất: mô hình hóa tính không chắc chắn thay vì bỏ qua tình chất này.
- Việc suy diễn và dự đoán có thể thực hiện được nhờ vào công cụ xác suất
- Ứng dụng trong: Học máy, khai phá dữ liệu, trí giác máy tình, NLP, công nghệ tin sinh,...
- Mục đích bài giảng:
 - Cái nhìn tổng quan về mô hình hóa xác suất
 - Các khái niệm quan trọng
 - Ứng dụng trong bài toán phân lớp

Dữ liệu

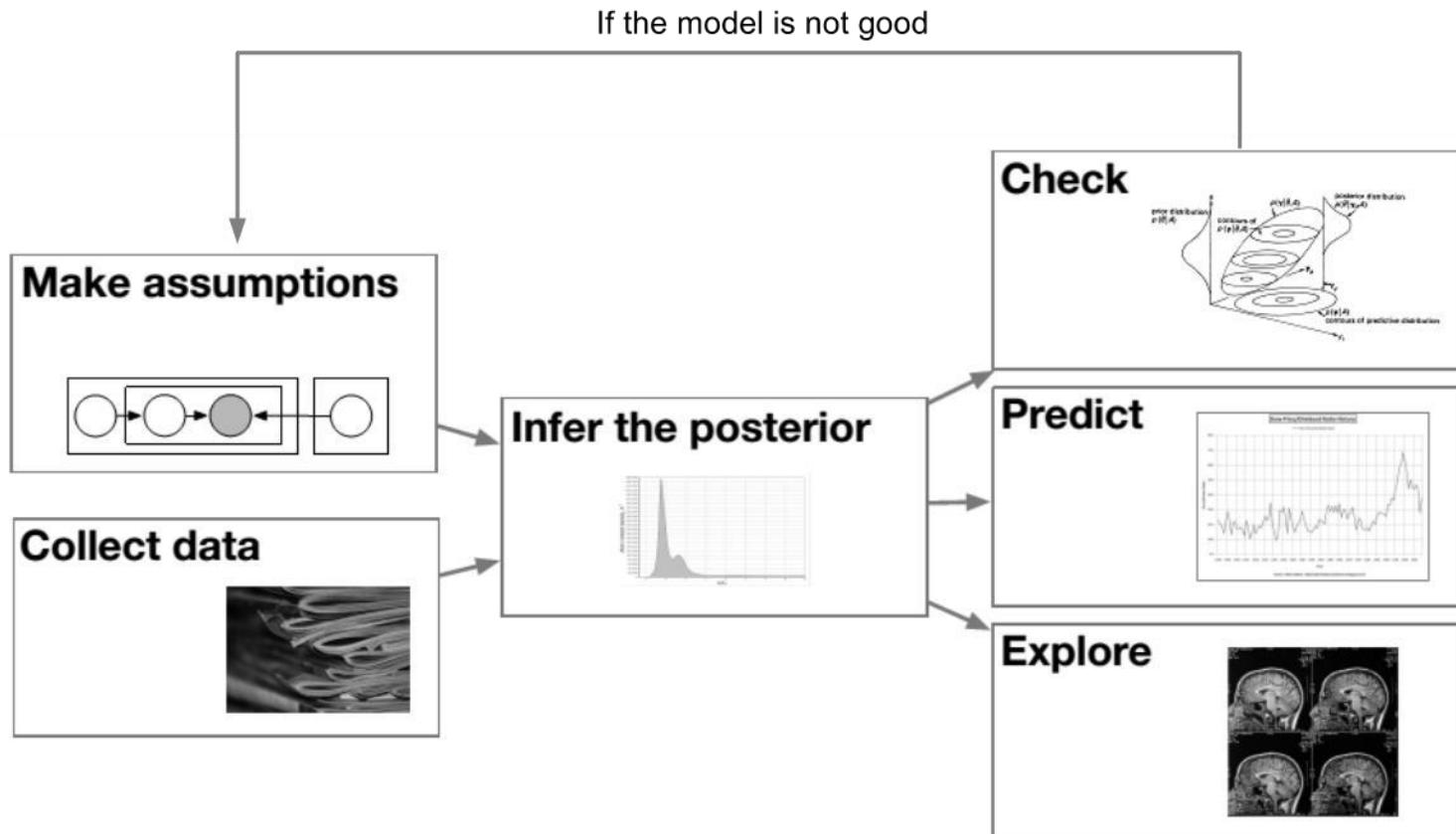
- Gọi $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_M, y_M)\}$ là tập dữ liệu cỡ M
 - Mỗi quan sát x_i là một biến n chiều
vd: $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})$ với mỗi chiều là một thuộc tính.
 - y là đầu ra đơn biến
- Dự đoán: cho vào tập dữ liệu D , có thể nhận xét gì về y^* cho một giá trị x^* chưa biết.
- Để dự đoán, chúng ta cần có giả thuyết
- Mô hình (model) H mã hóa những giả thuyết này và thường phụ thuộc vào một vài tham số θ , ví dụ:
$$y = f(x|\theta)$$
- Quá trình học chính là tìm được H từ tập D .



Sự không chắc chắn

- Sự không chắc chắn xuất hiện trong bất kỳ bước nào
 - Sự không chắc chắn do đo đạc (D)
 - Sự không chắc chắn của tham số (θ)
 - Sự không chắc chắn về tính chính xác của mô hình (H)
- Sự không chắc chắn do đo đạc
 - Sự không chắc chắn có thể xảy ra ở cả đầu vào và đầu ra?
- **Làm thế nào để biểu diễn sự không chắc chắn?**
-> Lý thuyết xác suất

Quá trình mô hình hóa



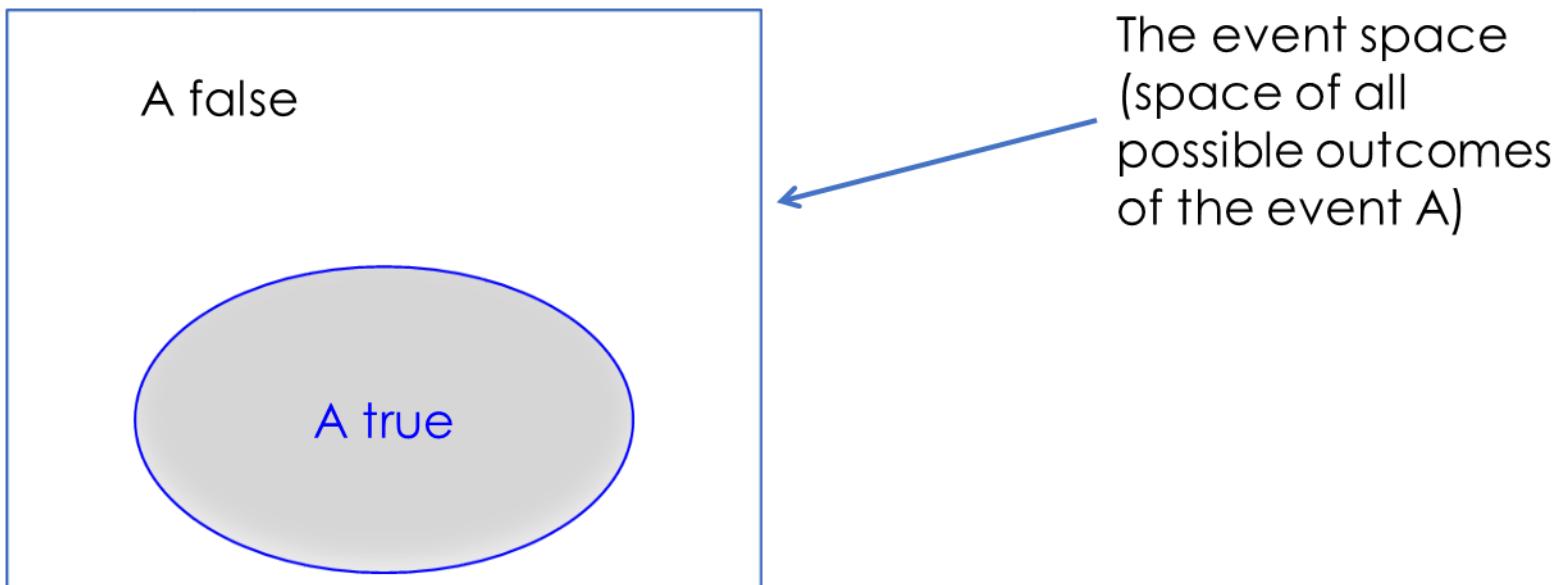
Lý thuyết xác suất cơ bản

Các khái niệm cơ bản

- Giả sử thực hiện thử nghiệm với các kết quả ngẫu nhiên,
Ví dụ: tung một con xúc xắc.
- Không gian S của kết quả: tập hợp tất cả các kết quả có thể có của một phép thử
 - Ví dụ: $S = \{1, 2, 3, 4, 5, 6\}$ cho việc tung con xúc xắc
- Sự kiện E: một tập con của không gian kết quả S.
 - Vd: $E = \{1\}$ sự kiện con xúc xắc xuất hiện 1.
 - Vd: $E = \{1, 3, 5\}$ trường hợp con xúc xắc xuất hiện lẻ.
- Không gian W của sự kiện: không gian của tất cả các sự kiện có thể xảy ra
 - Ví dụ: W chứa tất cả các lần tung có thể
- Biến ngẫu nhiên: đại diện cho một sự kiện ngẫu nhiên và có xác suất xuất hiện liên quan của sự kiện đó.

Biểu diễn xác suất

- Xác suất biểu diễn cho khả năng một sự kiện A có thể xảy ra.
 - Ký hiệu bởi $P(A)$
 - $P(A)$ là tỉ lệ của phần không gian con mà A là đúng.



Biến ngẫu nhiên nhị phân

- Một biến ngẫu nhiên nhị phân (boolean) chỉ có thể nhận giá trị Đúng hoặc Sai.
- Một số tiên đề:
 - $0 \leq P(A) \leq 1$
 - $P(\text{true}) = 1$
 - $P(\text{false}) = 0$
 - $P(A \text{ hoặc } B) = P(A) + P(B) - P(A, B)$
- Một số hệ quả:
 - $P(\text{không phải } A) = P(\sim A) = 1 - P(A)$
 - $P(A) = P(A, B) + P(A, \sim B)$

Các biến ngẫu nhiên đa thức

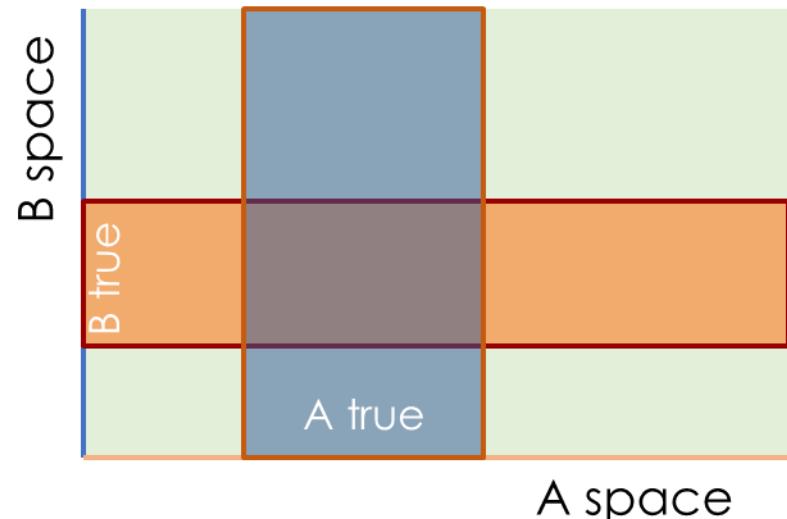
- Một biến ngẫu nhiên đa thức có thể nhận một từ K giá trị có thể có của $\{v_1, v_2, \dots, v_k\}$.
- $P(A = v_i, A = v_j) = 0$ nếu $i \neq j$

$$P\left(\bigcup_{n=1}^m (A = v_n)\right) = \sum_{n=1}^m P(A = v_n)$$

$$P\left(\bigcup_{n=1}^k (A = v_n)\right) = \sum_{n=1}^k P(A = v_n) = 1$$

Xác suất đồng thời

- Xác suất đồng thời:
 - Khả năng xảy ra của A và B cùng lúc.
 - $P(A, B)$ là tỷ lệ của không gian trong đó cả A và B đều đúng.

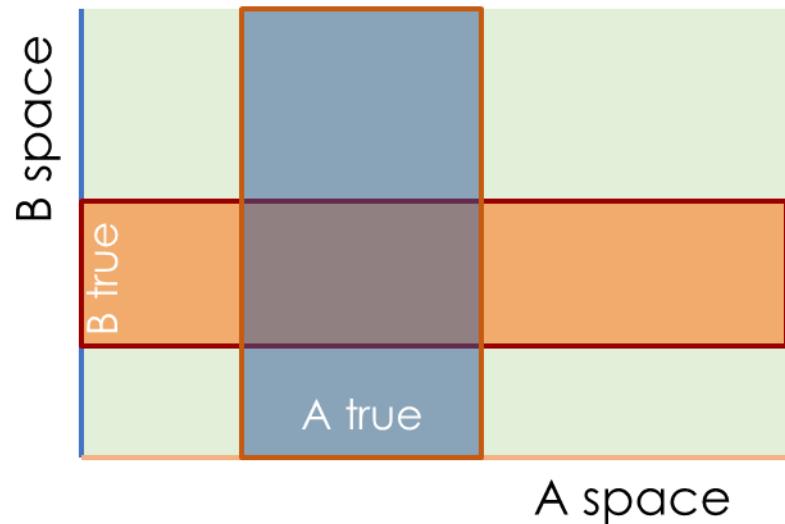


- Ví dụ:
 - A: Tôi sẽ chơi bóng đá vào ngày mai.
 - B: John sẽ không chơi bóng đá.
 - $P(A, B)$: xác suất mà ngày mai tôi sẽ chơi bóng còn John thì không.

Xác suất đồng thời (2)

- Ký hiệu S_A là không gian của A
- Ký hiệu S_B là không gian của B
- Ký hiệu S_{AB} là không gian của biến đồng thời (A, B)

$$S_{AB} = S_A \times S_B$$



- Khi đó:

$$P(A, B) = |T_{AB}| / |S_{AB}|$$

- T_{AB} là không gian mà cả A và B đều đúng
- $|X|$ là kích thước của không gian X

Xác suất có điều kiện

- Xác suất có điều kiện:
 - $P(A|B)$: khả năng A xảy ra khi B đã xảy ra.
 - $P(A|B)$: là tỉ lệ của không gian trong đó A xảy ra, biết rằng B đúng.
- Ví dụ:
 - A: Tôi sẽ chơi bóng đá vào ngày mai.
 - B: ngày mai trời sẽ không mưa.
 - $P(A | B)$: xác suất để tôi đá bóng đá, với điều kiện ngày mai trời không mưa.
- Sự khác nhau giữa xác suất đồng thời và xác suất có điều kiện?

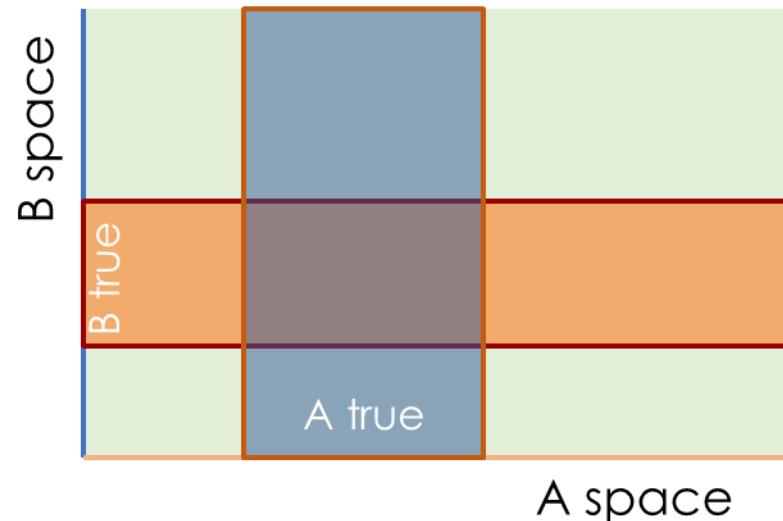
Xác suất có điều kiện (2)

- Xác suất có điều kiện:

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

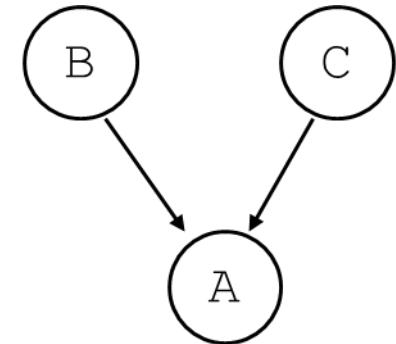
- Một số hệ quả:

- $P(A, B) = P(A|B).P(B)$
- $P(A|B) + P(\sim A|B) = 1$
- $\sum_{i=1}^k P(A = v_i|B) = 1$



Xác suất có điều kiện

- $P(A|B, C)$ là xác suất của A cho rằng B và C đã xảy ra.



- Ví dụ:
 - A: Sáng mai, tôi sẽ đi lang thang gần sông.
 - B: Thời tiết sáng mai rất đẹp.
 - C: Tôi sẽ thức dậy sớm vào sáng mai.
 - $P(A | B, C)$: xác suất đi lang thang qua gần con sông, với điều kiện trời rất đẹp và sáng mai tôi sẽ thức dậy sớm.

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

Độc lập thống kê

- Hai sự kiện A và B được gọi là **Độc lập thống kê** nếu xác suất A xảy ra không thay đổi bởi sự kiện B.

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

- Ví dụ:
 - A: Tôi sẽ chơi bóng vào ngày mai.
 - B: Biển Thái Bình Dương có nhiều cá.
 - $P(A|B) = P(A)$: việc biển Thái Bình Dương chứa nhiều cá không ảnh hưởng đến quyết định chơi bóng vào ngày mai của tôi.

Độc lập thống kê

- Giả sử $P(A|B) = P(A)$, ta có:
 - $P(\sim A|B) = P(\sim A)$
 - $P(B|A) = P(B)$
 - $P(A, B) = P(A).P(B)$
 - $P(\sim A, B) = P(\sim A).P(B)$
 - $P(A, \sim B) = P(A).P(\sim B)$
 - $P(\sim A, \sim B) = P(\sim A).P(\sim B)$

Độc lập có điều kiện

- Hai biến cố A và C được gọi là **Độc lập có điều kiện** cho trước B nếu $P(A|B, C) = P(A|B)$
- Ví dụ:
 - A: Tôi sẽ chơi bóng vào ngày mai.
 - B: trận đấu bóng đá sẽ diễn ra trong nhà vào ngày mai.
 - C: ngày mai trời sẽ không mưa.
 - $P(A|B, C) = P(A|B)$

Một số quy luật

- Luật chuỗi:
 - $P(A, B) = P(A|B).P(B) = P(B|A).P(A) = P(B, A)$
 - $P(A|B) = \frac{P(A,B)}{P(B)} = P(B|A).\frac{P(A)}{P(B)}$
 - $P(A, B|C) = \frac{P(A,B,C)}{P(C)} = P(A|B,C).\frac{P(B,C)}{P(C)} = P(A|B,C).P(B|C)$
- Luật độc lập:
 - $P(A|B) = P(A)$ nếu A và B độc lập thống kê
 - $P(A, B|C) = P(A|C).P(B|C)$ nếu A và B độc lập có điều kiện C
 - $P(A_1, A_2, \dots, A_n | C) = P(A_1|C) \dots P(A_n|C)$ nếu A_1, A_2, \dots, A_n là độc lập với điều kiện C.

Quy tắc nhân và tổng

- Coi x và y là các biến ngẫu nhiên rời rạc. Miền của chúng lần lượt là X và Y
- **Quy tắc nhân:**

$$P(x, y) = P(x|y)P(y)$$

- **Quy tắc tổng:** Xác suất của biến x bằng tổng các xác suất đồng thời của x với **tất cả các giá trị có thể** của y.

$$P(X) = \sum_{y \in Y} P(x, y) \quad P(x) = \sum_{y \in Y} P(x, y)$$

- Tổng sẽ chuyển thành tích phân nếu biến y liên tục

$$P(x) = \int P(x, y) dy$$

Định lý Bayes

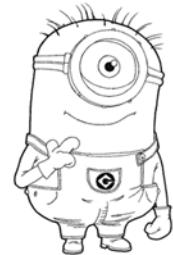
$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- $P(\theta)$ là xác suất tiên nghiệm (Prior) của biến θ
 - Sự không chắc chắn của chúng ta về θ trước khi quan sát dữ liệu.
- $P(D)$ xác suất tiên nghiệm mà chúng ta có thể quan sát tập dữ liệu D.
- $P(D|\theta)$ xác suất (likelihood) chúng ta có thể quan sát được tập dữ liệu D khi biết trước biến θ
- $P(\theta|D)$ xác suất hậu nghiệm của θ khi đã quan sát được tập dữ liệu D
 - Cách tiếp cận Bayesian dựa trên thông số này.

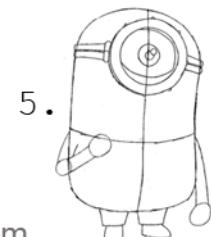
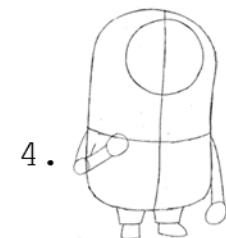
Mô hình xác suất

Mô hình xác suất

- Giả thuyết của chúng ta về quá trình dữ liệu được sinh ra như thế nào.
- VD: **Một câu được tạo ra như thế nào?**
 - Chúng ta giả sử bộ não hoạt động theo quy trình sau:
 - Đầu tiên, chọn chủ đề cho câu nói
 - Sinh từng từ một để tạo thành câu
- **TIM được tạo ra như thế nào?**

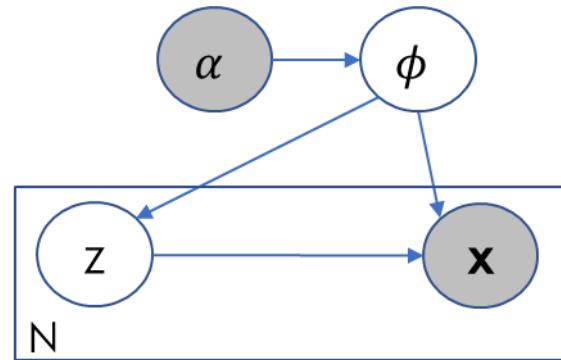


drawinghowtodraw.com



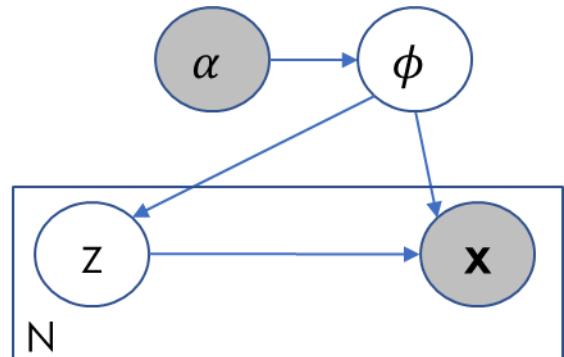
Mô hình xác suất

- Một mô hình đôi khi bao gồm
 - **Biến quan sát được:** mô tả những thứ quan sát hoặc thu thập được (ví dụ: x)
 - **Biến ẩn:** mô tả những thứ không quan sát được (ví dụ: z, ϕ)
 - **Biến cục bộ:** liên kết với một quan sát (ví dụ: z, x)
 - **Biến toàn cục:** chung cho các dữ liệu và thường dùng để đại diện cho mô hình (ví dụ: ϕ)
 - **Mối quan hệ giữa các biến**
- Mỗi biến tuân theo một phân phối xác suất nào đó.



Các loại mô hình

- **Mô hình đồ thị xác suất (PGM)**
 - Mỗi đỉnh đại diện cho một biến ngẫu nhiên, màu xám biểu diễn biến quan sát được, màu trắng biểu diễn biến ẩn
 - Mỗi cạnh đại diện cho mối quan hệ phụ thuộc có điều kiện giữa hai biến
 - Mô hình đồ thị có hướng: mỗi cạnh tuân theo một chiều
 - Mô hình đồ thị vô hướng: không có chiều trên các cạnh.
- **Mô hình biến ẩn:** một PGM có ít nhất 1 biến ẩn
- **Mô hình Bayes:** một PGM có xác suất tiên nghiệm trên các tham số mô hình.



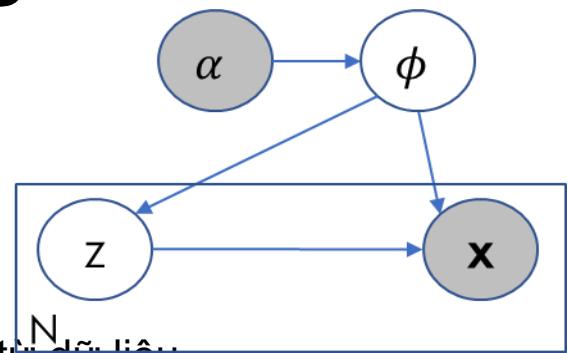
Mô hình xác suất

Mô hình xác suất là gì?

Một mô hình xác suất là cách biểu diễn **các biến và mối quan hệ giữa chúng thông qua xác suất**. Nó giúp chúng ta:

- Hiểu được cách dữ liệu được tạo ra,
- Mô phỏng lại quá trình sinh ra dữ liệu,
- Học các tham số ẩn từ dữ liệu quan sát được.

Các thành phần chính trong mô hình xác suất



1. Biến quan sát được (x)

- Là những giá trị mà ta có thể **đo lường, ghi nhận trực tiếp** từ dữ liệu.
- Ví dụ: chiều cao, cân nặng, điểm thi, hình ảnh...

2. Biến ẩn (z, ϕ)

- Là những đại lượng **không thể đo trực tiếp** từ dữ liệu.
- Chúng được giả định tồn tại và **có ảnh hưởng đến dữ liệu quan sát được**.
- Ví dụ:

- Chủ đề (topic) của một tài liệu,
- Cụm (cluster) của dữ liệu,
- Kỹ năng học sinh trong mô hình kiến thức (Knowledge Tracing)...

3. Biến cục bộ (z, x)

- Là những biến **gắn liền với từng điểm dữ liệu**.
- Ví dụ:
 - Nếu bạn có 10 bức ảnh, mỗi ảnh có 1 nhãn (z_i) → đó là biến cục bộ.

Các thành phần chính trong mô hình xác suất

4. Biến toàn cục (ϕ)

- Là những biến dùng chung cho toàn bộ tập dữ liệu.
- Ví dụ:

- Trọng số chủ đề trong toàn bộ tập tài liệu,
- Phân phối chung các đặc trưng của dữ liệu...

5. Mối quan hệ giữa các biến

- Mỗi biến được giả định có quan hệ xác suất với biến khác.
- Mối quan hệ này thường được mô tả bằng **đồ thị** (sẽ giải kỹ hơn ở slide sau).

💡 Tất cả các biến đều tuân theo một phân phối xác suất nào đó

- Ví dụ: $x \sim \mathcal{N}(\mu, \sigma^2)$, nghĩa là x tuân theo phân phối chuẩn.
- Biến ẩn có thể tuân theo phân phối khác (Bernoulli, Multinomial,...)

Các loại mô hình

1. Mô hình đồ thị xác suất (PGM = Probabilistic Graphical Model)

◇ **Mỗi đỉnh trong đồ thị:**

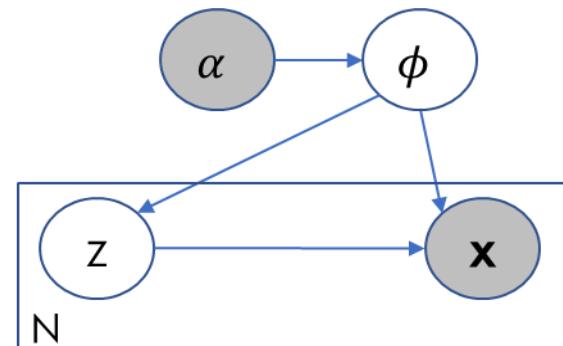
- Đại diện cho một **biến ngẫu nhiên**.
- Nếu đỉnh màu xám: biến **quan sát được**.
- Nếu đỉnh màu trắng: biến **ẩn**.

◇ **Mỗi cạnh trong đồ thị:**

- Đại diện cho **mối phụ thuộc điều kiện** giữa các biến.
- Nếu có hướng (mũi tên): mô hình có hướng (thường là mô hình Bayes).
- Nếu không có hướng: mô hình vô hướng (như MRF – Markov Random Field).

Ví dụ trong hình:

- x là biến quan sát được (màu xám),
- z, ϕ, α là biến ẩn (màu trắng),
- ϕ phụ thuộc α , z phụ thuộc ϕ , x phụ thuộc z .



Các loại mô hình

2. Mô hình biến ẩn

- Một mô hình được gọi là "**mô hình biến ẩn**" nếu trong đó có ít nhất **1 biến không quan sát được** (ẩn).
- Ví dụ: Mô hình LDA (Latent Dirichlet Allocation), HMM (Hidden Markov Model)...

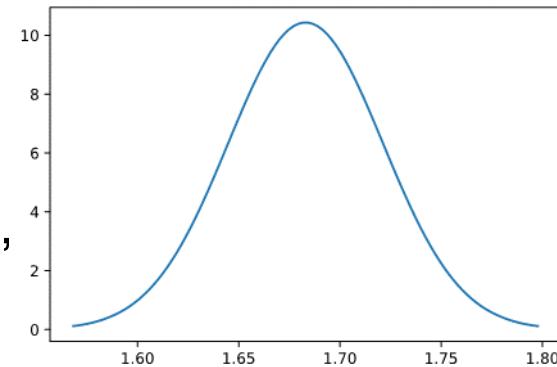
3. Mô hình Bayes

- Là một PGM trong đó có **phân phối xác suất tiên nghiệm (prior)** trên các tham số mô hình.
- Tức là trước khi có dữ liệu, ta giả định tham số đã phân phối theo một phân phối nào đó.
- Giúp mô hình học tốt hơn khi dữ liệu ít (giống như "đoán có cơ sở").

Phân phối chuẩn đơn biến

- Bài toán: vd Chúng ta muốn mô hình hóa chiều cao của một người
 - Tập dữ liệu từ 10 người ở Hà Nội:
 $D = \{1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62\}$
- Gọi x là biến ngẫu nhiên đại diện cho chiều cao của một người
- Ta giả thuyết: x tuân theo phân phối chuẩn (Gaussian) với hàm mật độ xác suất (PDF):

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



- Trong đó $\{\mu, \sigma^2\}$ là giá trị trung bình và phương sai,
 σ : độ lệch chuẩn (standard deviation)

Ý nghĩa mô hình hóa

- Khi ta nói: "chiều cao tuân theo phân phối chuẩn",
 - Nghĩa là hầu hết mọi người sẽ có chiều cao gần trung bình (μ),
 - Những người quá thấp hay quá cao sẽ xuất hiện ít hơn.

Phân phối chuẩn đơn biến

- Ghi chú:
 - $\mathcal{N}(x | \mu, \sigma^2)$ đại diện cho lớp các phân phối chuẩn
 - Lớp này được tham số hóa bởi $\theta = (\mu, \sigma^2)$
- Quá trình học: chúng ta cần biết các giá trị cụ thể của $\{\mu, \sigma^2\}$

Khái niệm

Giải thích

Biến quan sát

Có thể đo được từ dữ liệu (x)

Biến ẩn

Không thấy trực tiếp, phải suy diễn (z, ϕ)

PGM

Biểu diễn mối quan hệ xác suất qua đồ thị

Phân phối chuẩn

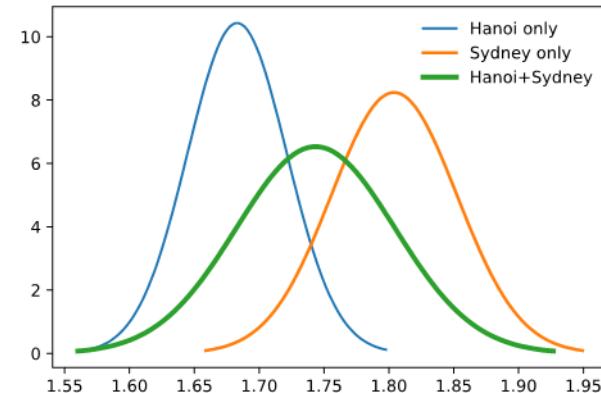
Phân phối liên tục hình chuông, rất phổ biến

Mục tiêu học

Tìm ra tham số (μ, σ^2) phù hợp với dữ liệu

Phân phối chuẩn đơn biến

- Mục tiêu là mô hình hóa chiều cao của một người
- Tập dữ liệu từ 10 người ở Hà Nội + 10 người ở Sydney
 - $D = \{1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62, 1.75, 1.80, 1.85, 1.65, 1.91, 1.78, 1.88, 1.79, 1.82, 1.81\}$
- Gọi x là biến ngẫu nhiên đại diện cho chiều cao
- Nếu chúng ta sử dụng phân phối Chuẩn:
 - Đường màu xanh lam mô hình chiều cao ở Hà Nội
 - Đường màu cam mô hình chiều cao ở Sydney
 - Đường màu xanh lục mô hình toàn bộ D
- Gaussian không mô hình hóa tốt cho dữ liệu này
→ Mô hình hỗn hợp?



Mô hình Gaussian hỗn hợp đơn biến (Univariate Gaussian Mixture Model – GMM)

1. Bối cảnh và giả định

Giả định: Dữ liệu bạn đang quan sát không đến từ một phân phối chuẩn duy nhất, mà đến từ sự kết hợp (hỗn hợp) của nhiều phân phối Gaussian khác nhau.

Ví dụ trực quan:

- Chiều cao của một lớp học có cả nam và nữ.
 - Nam: cao hơn, phân phối chuẩn riêng.
 - Nữ: thấp hơn, phân phối chuẩn khác.
- Dữ liệu chung là hỗn hợp của 2 phân phối chuẩn này.

2. Quá trình sinh dữ liệu (Generative process)

Mỗi điểm dữ liệu x được sinh ra theo 2 bước:

1. Chọn chỉ số phân phối:

- Gọi z là biến chọn phân phối (\hat{z}).
- z được lấy từ phân phối rời rạc **Multinomial** với tham số $\phi \backslash \text{phi}$
- (multinomial ở đây là dạng đơn giản: phân phối nhị phân 0 hoặc 1).
- Nghĩa là:
 - Với xác suất ϕ chọn Gaussian 1.
 - Với xác suất $1 - \phi$, chọn Gaussian 2.

2. Sinh dữ liệu thực tế:

- Nếu $z = 1$, thì sinh x từ $\mathcal{N}(x | \mu_1, \sigma_1^2)$
- Nếu $z = 2$, thì sinh x từ $\mathcal{N}(x | \mu_2, \sigma_2^2)$

3. Hàm mật độ của mô hình hỗn hợp

Khi không biết z là gì (biến ẩn), ta lấy kỳ vọng (trung bình) của hai phân phối có trọng số:

$$f(x) = \phi \cdot \mathcal{N}(x|\mu_1, \sigma_1^2) + (1 - \phi) \cdot \mathcal{N}(x|\mu_2, \sigma_2^2)$$

👉 Hiểu đơn giản:

- Phân phối của x là **sự cộng gộp** có trọng số của hai phân phối Gaussian.
- ϕ : là **trọng số** hay xác suất chọn Gaussian 1.

Mô hình này dùng để làm gì?

- **Phân cụm (clustering)**: tìm các nhóm ẩn trong dữ liệu.
- **Phát hiện bất thường (anomaly detection)**: điểm dữ liệu rời quá xa cả 2 phân phối \rightarrow có thể là bất thường.
- **Tăng khả năng mô hình hóa**: khi dữ liệu không "đẹp", ta dùng hỗn hợp nhiều Gaussian để "xấp xỉ" phân phối thật.

Thành phần**Ý nghĩa** μ_1, σ_1^2

Trung bình và phương sai Gaussian 1

 μ_2, σ_2^2

Trung bình và phương sai Gaussian 2

 ϕ

Xác suất chọn Gaussian 1

 $1 - \phi$

Xác suất chọn Gaussian 2

 z

Biến ẩn chọn phân phối (0 hoặc 1)

 x

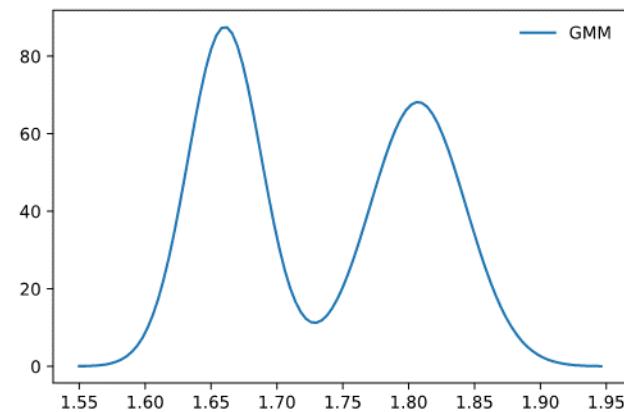
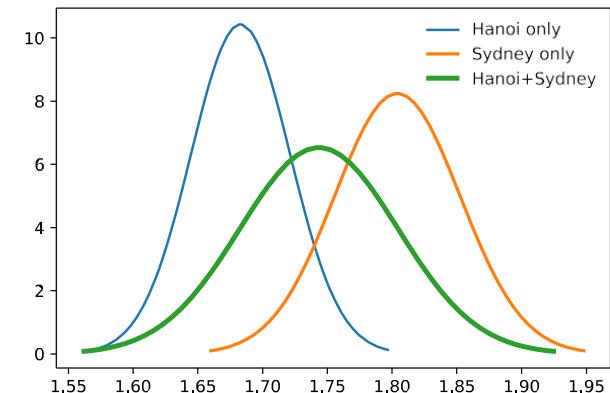
Dữ liệu quan sát được

Tóm tắt :Mô hình Gaussian hỗn hợp đơn biến

- Giả định: dữ liệu được tạo từ hai phân bố Gaussian khác nhau và mỗi quan sát được sinh bởi một trong số đó.

Quá trình sinh:

- Chọn chỉ số $z \sim Multinomial(z|\phi)$
- Sinh mẫu $x \sim Normal(x | \mu_z, \sigma_z^2)$
- Đây là mô hình hỗn hợp Gauss
 - (μ_1, σ_1^2) đại diện cho phân phối Gaussian thứ nhất
 - (μ_2, σ_2^2) đại diện cho Gaussian thứ hai
 - $\phi \in [0,1]$ là tham số của phân phối Đa thức, và
 - $P(z = 1 | \phi) = \phi = 1 - P(z = 2 | \phi)$
- Hàm mật độ:
$$\phi \mathcal{N}(x | \mu_1, \sigma_1^2) + (1 - \phi) \mathcal{N}(x | \mu_2, \sigma_2^2)$$



GMM đa biến

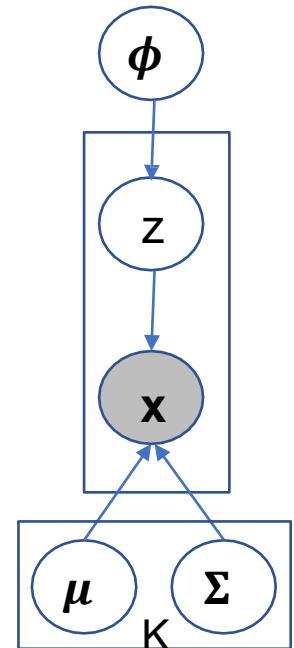
- Xét trường hợp mỗi x thuộc không gian n chiều.
- GMM: giả định rằng dữ liệu là các mẫu từ K phân bố Gaussian khác nhau.
- Mỗi x được tạo ra từ một trong K Gaussian theo **quá trình sinh** như sau:
 - Lấy chỉ số: $z \sim Multinomial(z|\phi)$
 - Sinh: $x \sim Normal(x | \mu_z, \Sigma_z)$
- Hàm mật độ:

$$p(x|\mu, \Sigma, \phi) = \sum_{k=1}^K \phi_k \mathcal{N}(x | \mu_k, \Sigma_k)$$

- $\phi = (\phi_1, \dots, \phi_K)$ chứa trọng số của từng phân bố con

- Mỗi Gaussian có hàm mật độ:

$$\mathcal{N}(x | \mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$



Một số mô hình phổ biến

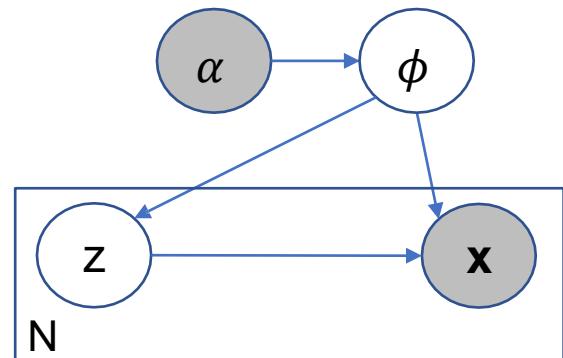
- Mô hình hỗn hợp Gaussian (GMM)
 - Mô hình dữ liệu có giá trị thực
- Mô hình Latent Dirichlet Allocation
 - Mô hình các chủ đề ẩn trong dữ liệu văn bản
- Mô hình Markov ẩn (HMM)
 - Mô hình chuỗi thời gian, dữ liệu theo thời gian hoặc có bản chất tuần tự
- Trường ngẫu nhiên có điều kiện (CRF)
 - Cho dự đoán cấu trúc
- Mô hình sinh sâu (Deep generative models)
 - Mô hình các cấu trúc ẩn, sinh dữ liệu nhân tạo

Mô hình xác suất: hai bài toán

- **Suy diễn:** cho một trường hợp nhất định x_n
 - Khôi phục biến cục bộ (ví dụ: z_n)
 - Sự phân phối của các biến cục bộ (VD: $P(z_n, x_n | \phi)$)
 - Ví dụ: đối với GMM, ta muốn biết z_n đại diện cho phân bố con nào đã tạo ra x_n
- **Học (ước lượng):**

Cho trước một tập dữ liệu, hãy ước lượng phân phối đồng thời của các biến

 - Ví dụ: ước lượng $P(\phi, z_1, \dots, z_n, x_1, \dots, x_n | \alpha)$
 - Ví dụ: ước lượng $P(x_1, \dots, x_n | \alpha)$
 - Ví dụ: ước lượng α
 - Suy diễn của các biến cục bộ thường là cần thiết



1. Suy diễn (Inference):

Cho biết một số biến quan sát (như x_n), ta muốn suy ra giá trị của biến ẩn (như z_n).

- **Ví dụ:** Nếu x_n là dữ liệu (âm thanh, văn bản, ảnh...), thì z_n là “nhãn” hoặc “nhóm” nào đã tạo ra dữ liệu đó (ẩn).
- **Khôi phục biến ẩn:** Dùng để tìm ra z_n từ x_n – ví dụ xác định một điểm dữ liệu thuộc cụm nào trong bài toán phân cụm (clustering).
- **Tính phân phối:** Tìm $P(z_n, x_n | \phi)$, tức là xác suất xảy ra đồng thời của z_n và x_n , với ϕ là tham số của mô hình.

👉 Trong GMM (Gaussian Mixture Model), z_n là chỉ số cụm (cluster), ta cần suy ra z_n ứng với từng điểm x_n .

2. Học (ước lượng tham số):

Cho tập dữ liệu, hãy học (ước lượng) các phân phối xác suất.

- Mục tiêu là **ước lượng phân phối đồng thời của các biến** từ dữ liệu. Bao gồm:
 - $P(\phi, z_1, \dots, z_n, x_1, \dots, x_n | \alpha)$
 - $P(x_1, \dots, x_n | \alpha)$
 - $P(z_1, \dots, z_n | \alpha)$
- **Ước lượng tham số mô hình:**
 - ϕ : tham số quan sát.
 - α : tham số siêu cấp (hyperparameter).
- Trong quá trình học, ta thường cần **suy diễn** các biến ẩn z song song.

Suy diễn và học

Một số cách suy diễn

- Gọi D là dữ liệu và h là giả thuyết
 - Giả thuyết : tham số chưa biết, biến ẩn, ...
- **Cực đại hóa khả năng** (Maximum Likelihood Estimation - MLE)

$$h^* = \arg \max_{h \in H} P(D|h)$$

- Tìm h^* (trong không gian giả thuyết H) tối đa hóa khả năng xảy ra của dữ liệu.
- Nói cách khác: MLE đưa ra suy luận về mô hình có nhiều khả năng đã tạo ra dữ liệu.
- Suy diễn Bayes (Bayesian inference) xem xét việc biến đổi tri thức tiên nghiệm $P(h)$ của chúng ta, thông qua dữ liệu D, thành tri thức hậu nghiệm $P(h|D)$
- Từ luật Bayes: $P(h|D) = P(D|h)P(h)/P(D)$

$$P(h|D) \propto P(D|h) * P(h)$$

(Posterior \propto Likelihood * Prior)

Một số cách suy diễn (2)

- Trong một số trường hợp, chúng ta có thể biết phân phối tiên nghiệm của h .
- Cực đại hóa hậu nghiệm** (Maximum a Posterior Estimation - MAP)

$$\begin{aligned} h^* &= \arg \max_{h \in \mathcal{H}} P(h|\mathbf{D}) = \arg \max_{h \in \mathcal{H}} P(\mathbf{D}|h) P(h)/P(\mathbf{D}) \\ &= \arg \max_{h \in \mathcal{H}} P(\mathbf{D}|h) P(h) \end{aligned}$$

- Tìm h^* tối đa hóa xác suất hậu nghiệm của h .
- MAP tìm một điểm, không phải phân phối \rightarrow **Ước lượng điểm**
- MLE là một trường hợp đặc biệt của MAP, khi sử dụng phân phối đều cho h .
- Suy diễn Bayes đầy đủ cố gắng ước lượng phân phối hậu nghiệm đầy đủ $P(h|D)$, không chỉ một điểm h^* .
- Ghi chú: MLE, MAP hoặc Bayes đầy đủ có thể được áp dụng cho cả quá trình học và suy diễn.

MLE: ví dụ Gaussian

- Chúng ta muốn mô hình hóa chiều cao của một người bằng tập dữ liệu
 $D = \{1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62\}$
- Gọi x là biến ngẫu nhiên đại diện cho chiều cao của một người.
- Mô hình: giả sử rằng x tuân theo phân phối Gaussian với giá trị trung bình μ và phương sai σ^2

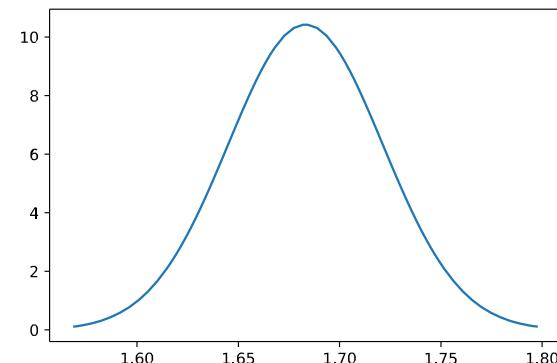
Câu hỏi:

Tìm giá trị **ước lượng hợp lý nhất** (Maximum Likelihood Estimation – MLE) của **trung bình** μ và **độ lệch chuẩn** σ cho phân phối chuẩn $\mathcal{N}(x|\mu, \sigma^2)$

=>**Quá trình học:** tìm (μ, σ) từ dữ liệu đã cho $D = \{x_1, \dots, x_{10}\}$.

- Gọi $f(x|\mu, \sigma)$ là hàm mật độ của họ Gaussian, được tham số hóa bởi (μ, σ) .
 - $f(x_n|\mu, \sigma)$ là khả năng xảy ra của trường hợp x_n
 - $f(D|\mu, \sigma)$ là hàm khả năng xảy ra của D .
- Sử dụng MLE, chúng ta đi tìm

$$(\mu_*, \sigma_*) = \arg \max_{\mu, \sigma} f(D|\mu, \sigma)$$



Với phân phối chuẩn, các công thức MLE cụ thể là:

- Trung bình MLE:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

- Phương sai MLE:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

- Độ lệch chuẩn: $\sigma = \sqrt{\sigma^2}$

Bước 1: Tính trung bình

Tập dữ liệu:

$$D = \{1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62\}$$

Cộng tổng:

$$\text{Tổng} = 1.6 + 1.7 + 1.65 + 1.63 + 1.75 + 1.71 + 1.68 + 1.72 + 1.77 + 1.62 = 16.83$$

Số mẫu: $n = 10$

$$\mu = \frac{16.83}{10} = 1.683$$

Bước 2: Tính phương sai

Công thức:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Tính từng bình phương hiệu số:

x_i	$x_i - \mu$	$(x_i - \mu)^2$	Tổng cộng:
1.6	-0.083	0.006889	$\sum(x_i - \mu)^2 = 0.02921$
1.7	0.017	0.000289	Phương sai:
1.65	-0.033	0.001089	$\sigma^2 = \frac{0.02921}{10} = 0.002921$
1.63	-0.053	0.002809	Độ lệch chuẩn:
1.75	0.067	0.004489	$\sigma = \sqrt{0.002921} \approx 0.054$
1.71	0.027	0.000729	
1.68	-0.003	0.000009	
1.72	0.037	0.001369	
1.77	0.087	0.007569	
1.62	-0.063	0.003969	

Kết quả cuối cùng (MLE)

- $\mu_* = 1.683$
- $\sigma_* \approx 0.054$

MLE: ví dụ Gaussian (2)

- Giả thuyết i.i.d.: giả định rằng dữ liệu được sinh ra một cách độc lập với nhau
- Khi đó $P(D|\mu, \sigma) = P(x_1, \dots, x_{10}|\mu, \sigma) = \prod_{i=1}^{10} P(x_i|\mu, \sigma)$
- Sử dụng giả thuyết này, MLE sẽ tìm

$$\begin{aligned}(\mu_*, \sigma_*) &= \arg \max_{\mu, \sigma} \prod_{i=1}^{10} f(x_i|\mu, \sigma) = \arg \max_{\mu, \sigma} \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\&= \arg \max_{\mu, \sigma} \log \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \quad \text{← Log trick, } \log \stackrel{\text{def}}{=} \ln \\&= \arg \max_{\mu, \sigma} \left[\sum_{i=1}^{10} \left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 - \log \sqrt{2\pi\sigma^2} \right) \right]\end{aligned}$$

Log trick,
 $\log \stackrel{\text{def}}{=} \ln$

- Sử dụng đạo hàm (cho biến μ, σ), ta sẽ tìm được

$$\mu_* = \frac{1}{10} \sum_{i=1}^{10} x_i = 1.683, \quad \sigma_*^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - \mu_*)^2 \approx 0.0015$$

Gaussian Naïve Bayes bao gồm:

Học (learning): ước lượng tham số từ dữ liệu có nhãn
Và suy diễn (inference): dự đoán nhãn cho dữ liệu

mời

1. Giai đoạn Học (Learning) – có giám sát:

- Được thực hiện khi huấn luyện mô hình.
- Cho tập dữ liệu có nhãn: (x_i, y_i)
- Ta ước lượng các tham số:

- $P(y)$: xác suất nhãn (prior)
- $P(x_j | y)$: phân phối của từng thuộc tính điều kiện trên lớp – trong Gaussian Naïve Bayes thì $P(x_j | y)$ là phân phối chuẩn (Gaussian).

$$P(x_j | y) = \mathcal{N}(x_j | \mu_{y,j}, \sigma_{y,j}^2)$$

- Đây là bài toán học có giám sát (supervised learning) vì biết trước nhãn.

Ước lượng các tham số $\mu_{y,i}$, $\sigma_{y,i}$ từ dữ liệu có nhãn.

2. Giai đoạn Suy diễn (Inference) – dự đoán:

- Khi gặp dữ liệu mới x^* , ta dùng Bayes để suy ra nhãn y^* :

$$P(y | x^*) \propto P(y) \prod_j P(x_j | y)$$

- Đây là bài toán suy diễn: tìm nhãn y^* tốt nhất cho dữ liệu mới.

MAP: Gaussian naïve Bayes

- Xét bài toán phân lớp
 - Dữ liệu huấn luyện $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_M, y_M)\}$ với M quan sát, C lớp.
 - Mỗi \mathbf{x}_i là một vectơ trong không gian n chiều \mathbb{R}^n , $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$
- *Mô hình:* giả sử có C phân bõ khác nhau tạo ra dữ liệu trong D và dữ liệu có nhãn c được tạo ra từ phân phõi Gaussian được tham số hóa bởi (μ_c, Σ_c)
 - μ_c là vectơ trung bình, Σ_c là ma trận hiệp phương sai kích thước $n \times n$.
- *Quá trình học:* ta xét $P(\mu, \Sigma, c | \mathbf{D})$, với $(\mu, \Sigma) = (\mu_1, \Sigma_1, \dots, \mu_C, \Sigma_C)$

$$(\mu_*, \Sigma_*) \stackrel{\text{def}}{=} \arg \max_{\mu, \Sigma, c} P(\mu, \Sigma, c | \mathbf{D}) = \arg \max_{\mu, \Sigma, c} P(\mathbf{D} | \mu, \Sigma, c)$$

Định lý Bayes,
bỏ $P(\mathbf{D})$,
giả thuyết phân bõ
tiên nghiệm đều
cho μ, Σ

- Ước lượng $P(c)$ là tỷ lệ của lớp c trong D:
 $P(c) = |\mathbf{D}_c| / |\mathbf{D}|$ trong đó \mathbf{D}_c chứa tất cả các dữ liệu có nhãn c trong D.
- Vì các lớp C là độc lập, chúng ta có thể học cho mỗi lớp

$$(\mu_{c*}, \Sigma_{c*}) \stackrel{\text{def}}{=} \arg \max_{\mu_c, \Sigma_c} P(\mathbf{D}_c | \mu_c, \Sigma_c) P(c) = \arg \max_{\mu_c, \Sigma_c} P(\mathbf{D}_c | \mu_c, \Sigma_c)$$

MAP: Gaussian Naïve Bayes (2)

- Giả sử các mẫu là độc lập, ta có:

$$\begin{aligned}(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) &= \arg \max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} G \quad P(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \arg \max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \mathbb{E} \log P(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \\&= \arg \max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \mathbb{E} \sum_{x \in D_c} \log \left[\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_c)}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) \right) \right] \\&= \arg \max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{x \in D_c} -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) - \log \sqrt{\det(2\pi\boldsymbol{\Sigma}_c)}\end{aligned}$$

- Sử dụng gradient theo $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$, chúng ta đạt được:

$$\boldsymbol{\mu}_{c*} = \frac{1}{|D_c|} \sum_{x \in D_c} \mathbf{x}, \quad \boldsymbol{\Sigma}_{c*} = \frac{1}{|D_c|} \sum_{x \in D_c} (\mathbf{x} - \boldsymbol{\mu}_{c*})(\mathbf{x} - \boldsymbol{\mu}_{c*})^T$$

- Do đó sau quá trình huấn luyện chúng ta đạt được $(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, P(c))$ cho mỗi lớp c

MAP: Gaussian Naïve Bayes (3)

- Mô hình được huấn luyện: $(\mu_{c*}, \Sigma_{c*}, P(c))$ cho mỗi lớp c
- Dự đoán** cho dữ liệu mới z bằng cách tìm nhãn lớp mà có xác suất hậu nghiệm cao nhất:

Bayes' rule

$$\begin{aligned} c_z &= \arg \max_{c \in \{1, \dots, C\}} P(c|z, \mu_{c*}, \Sigma_{c*}) = \arg \max_{c \in \{1, \dots, C\}} P(z|\mu_{c*}, \Sigma_{c*}, c)P(c) \\ &= \arg \max_{c \in \{1, \dots, C\}} \log P(z|\mu_{c*}, \Sigma_{c*}, c) + \log P(c) \\ &= \arg \max_{c \in \{1, \dots, C\}} -\frac{1}{2}(z - \mu_{c*})^T \Sigma_{c*}^{-1} (z - \mu_{c*}) - \log \sqrt{\det(2\pi\Sigma_{c*})} + \log P(c) \end{aligned}$$

- Nếu sử dụng MLE, chúng ta không cần sử dụng xác xuất tiên nghiệm P(c)

MAP: Multinomial Naïve Bayes (1)

- Xét bài toán phân loại văn bản (dữ liệu có tính rời rạc)
 - Tập huấn luyện $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_M, y_M)\}$ với M tài liệu, C lớp.
 - TF: mỗi tài liệu \mathbf{x}_i được biểu diễn bằng một vectơ có V chiều, ví dụ: $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$ mỗi x_{ij} là tần suất của từ j trong tài liệu \mathbf{x}_i
- *Mô hình:* giả sử có C phân bố khác nhau tạo ra dữ liệu trong D và dữ liệu có nhãn c được tạo ra từ một phân phối đa thức được tham số hóa bởi θ_c và có hàm khối lượng xác suất
$$f(x_1, \dots, x_V | \theta_{c1}, \dots, \theta_{cV}) = \frac{\Gamma(\sum_{j=1}^V x_j + 1)}{\prod_{j=1}^V \Gamma(x_j + 1)} \prod_{k=1}^C \theta_{ck}^{x_k}$$
 - $\theta_{cj} = P(x = j | \theta_{cj})$ là xác suất mà từ $j \in \{1, \dots, V\}$ xuất hiện, thỏa mãn $\sum_{k=1}^C \theta_{ck} = 1$ và Γ là hàm gamma.
- *Quá trình học:* chúng ta có thể làm tương tự với Gaussian Naïve Bayes để ước lượng $\theta_c = (\theta_{c1}, \dots, \theta_{cV})$ và $P(c)$.

MAP: Multinomial Naïve Bayes (2)

- Mô hình đã huấn luyện: $(\theta_{c*}, P(c))$ cho mỗi lớp c
- Dự đoán cho dữ liệu mới $\mathbf{z} = (z_1, \dots, z_V)^T$:

$$\begin{aligned} c_z &= \arg \max_{c \in \{1, \dots, C\}} P(c | \mathbf{z}, \theta_{c*}) = \arg \max_{c \in \{1, \dots, C\}} P(\mathbf{z} | \theta_{c*}, c) P(c) \\ &= \arg \max_{c \in \{1, \dots, C\}} \log P(\mathbf{z} | \theta_{c*}) + \log P(c) \end{aligned} \quad (\text{MNB.1})$$

$$\begin{aligned} &= \arg \max_{c \in \{1, \dots, C\}} \log \frac{\Gamma(\sum_{j=1}^V z_j + 1)}{\prod_{j=1}^V \Gamma(z_j + 1)} G \prod_{k=1}^V \theta_{ck*}^{z_k} + \log P(c) \\ &= \arg \max_{c \in \{1, \dots, C\}} \log G \prod_{k=1}^V \theta_{ck*}^{z_k} + \log P(c) \\ &= \arg \max_{c \in \{1, \dots, C\}} \log G \prod_{k=1}^V P(z_k | \theta_{ck*}) + \log P(c) \end{aligned} \quad (\text{MNB.2})$$

- Nhận cho xác suất hậu nghiệm cao nhất
- Lưu ý: về cơ bản chúng ta giả thuyết rằng *các thuộc tính là độc lập với nhau* (từ hai phương trình MNB.1 và MNB.2)

Nhìn lại GMM

- Xem xét việc học GMM, với phân phối K Gaussian, từ dữ liệu huấn luyện $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$.
- Hàm mật độ là $p(x|\mu, \Sigma, \phi) = \sigma_{k=1}^K \phi_k \mathcal{N}(x | \mu_k, \Sigma_k)$

- $\phi = (\phi_1, \dots, \phi_K)$ đại diện cho trọng số của các Gaussian
- Mỗi Gaussian đa biến có mật độ:

$$\mathcal{N}(x | \mu_k, \Sigma_k) = \frac{1}{\sqrt{\det(2\pi\Sigma_k)}} \exp \left[-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right]$$

- MLE cõ gắng cực đại hóa hàm log-likelihood:

$$M \quad K$$

$$L(\mu, \Sigma, \phi) = \sum_{i=1}^M \log \sum_{k=1}^K \phi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k)$$

- Chúng ta không thể tìm thấy một nghiệm có công thức tường minh!

- Cần các thuật toán xấp xỉ.

Vài tình huống khó khăn

- Không tìm được ngay công thức nghiệm:
 - Các ví dụ trước đây đều là các ví dụ đơn giản bởi vì có thể tìm được lời giải ngay bằng gradient
 - Nhiều mô hình G khác không có dạng công thức nghiệm cụ thể như vậy
- Không có công thức tường minh để tính toán
- Bài toán suy diễn không khả thi:
 - Inference in many probabilistic models is NP-hard.
[Sontag & Roy, 2011; Tosh & Dasgupta, 2019]

Tài liệu tham khảo

- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." *Journal of the American Statistical Association* 112, no. 518 (2017): 859-877.
- Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight Uncertainty in Neural Network." In *International Conference on Machine Learning (ICML)*, pp. 1613-1622. 2015.
- Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In *International Conference on Machine Learning*, pp. 1050-1059. 2016.
- Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." *Nature* 521, no. 7553 (2015): 452-459.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." In *International Conference on Learning Representations (ICLR)*, 2014.
- Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349, no. 6245 (2015): 255-260.
- Tosh, Christopher, and Sanjoy Dasgupta. "The Relative Complexity of Maximum Likelihood Estimation, MAP Estimation, and Sampling." In *Proceedings of the 32nd Conference on Learning Theory*, in PMLR 99:2993-3035, 2019.
- Sontag, David, and Daniel Roy, "Complexity of inference in latent dirichlet allocation" in: *Proceedings of Advances in Neural Information Processing System*, 2011.