ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Học máy cơ bản

Naïve Bayes

Thân Quang Khoát

Nội dung môn học

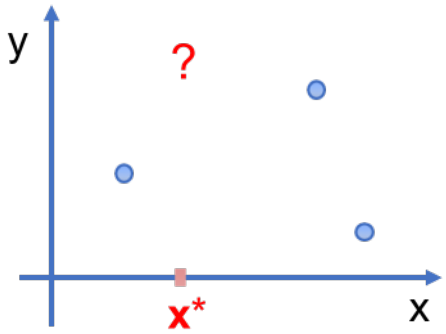
- Buổi 1: Giới thiệu về Học máy
- Buổi 2: Quy trình xây dựng hệ thống học máy
- Buổi 3: Hồi quy tuyến tính
- Buổi 4: Học dựa trên láng giềng gần nhất (KNN)
- Buổi 5: Cây quyết định và Rừng ngẫu nhiên
- **Buổi 6: Naïve Bayes**
- Buổi 7: Máy vector hỗ trợ (SVM)
- Buổi 8: Đánh giá hiệu quả của mô hình học máy
- Buổi 9: Phân cụm
- Buổi 10-11: Kiểm tra giữa kỳ và trình bày ý tưởng làm dự án cuối kỳ
- Buổi 12-20: Học sâu

Tại sao cần mô hình hóa xác suất?

- Việc suy diễn từ dữ liệu thương 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, tri 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:

$$y = f(x|\theta)$$

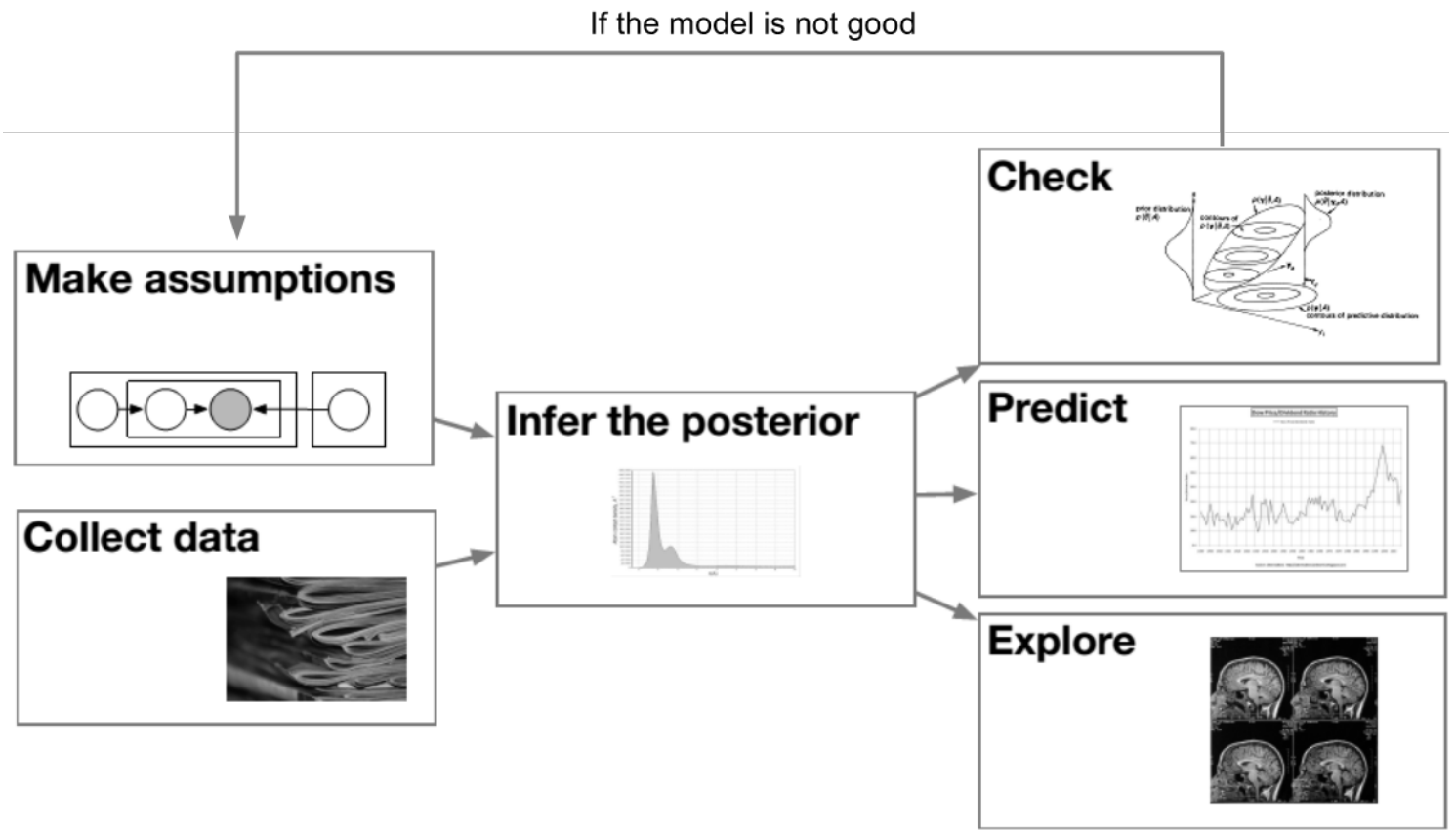
- Quá trình học chính là tìm được **H** từ tập D.

Sự không chắc chắn

- Sự không chắc chắn xuất hiện trong bất kỳ bước nào
  - Sự không chắc chắn trong dữ liệu (**D**)
  - Sự không chắc chắn của tham số (**θ**)
  - Sự không chắc chắn về mô hình (**H**)
- Sự không chắc chắn trong dữ liệu
  - 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 (Probability theory)

Quá trình mô hình hóa



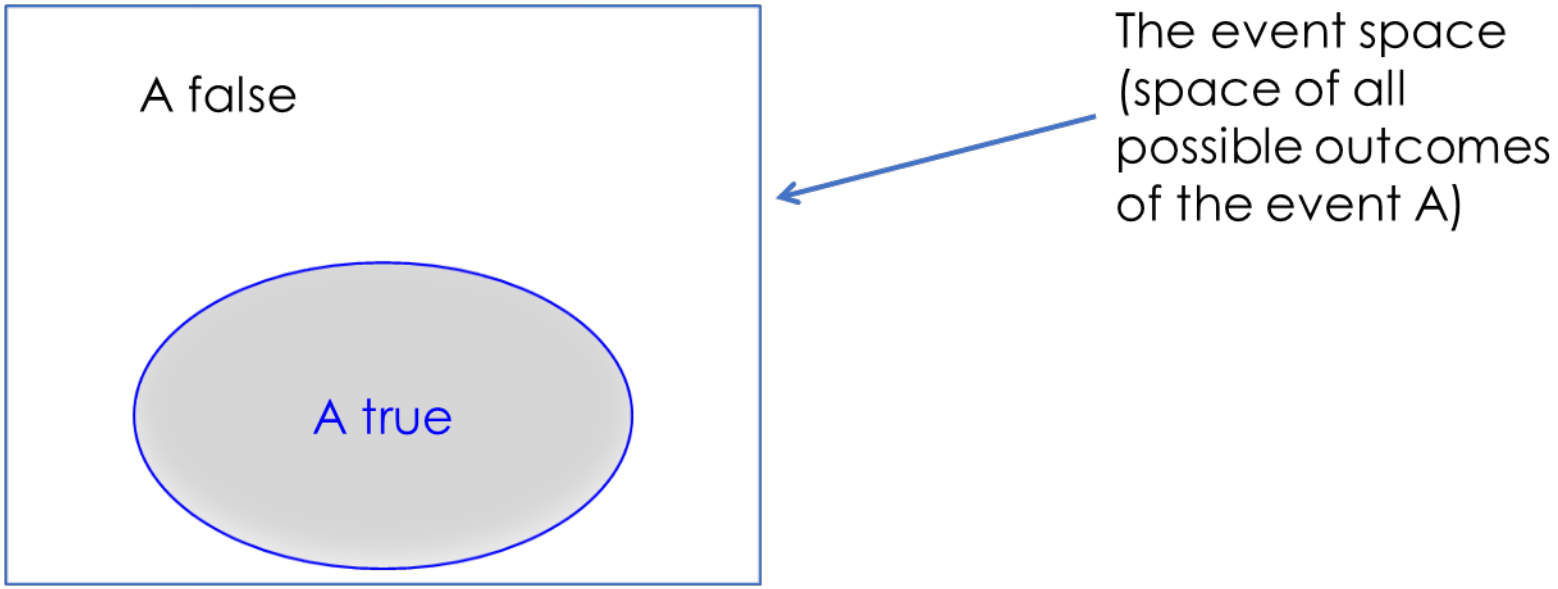
Lý thuyết xác suất cơ bản

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(true) = 1
  - P(false) = 0
  - $P(A$ hoặc $B) = P(A) + P(B) - P(A, B)$
- Một số hệ quả:
  - P(không phải A) = P(~ A) = 1 - P(A)
  - P(A) = P(A, B) + P(A, ~ B)

Các biến ngẫu nhiên đa thức

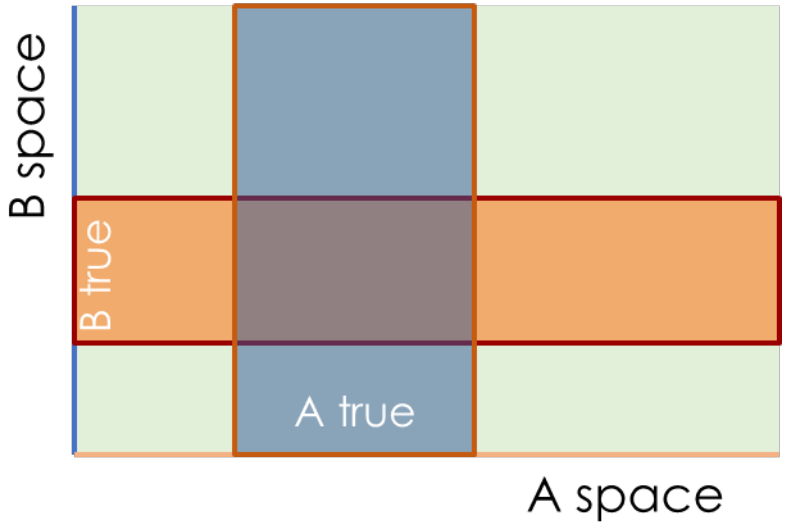
- Một biến ngẫu nhiên đa thức có thể nhận một từ *k* giá trị có thể có của $\{v_1, v_2, \dots, v_k\}$.
- $P(A = v_i, A = v_j) = 0 \;\; nếu \;\; i \neq j$

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

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

Xác suất đồng thời

- ***Xác suất đồng thời (joint probability):***
  - 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.

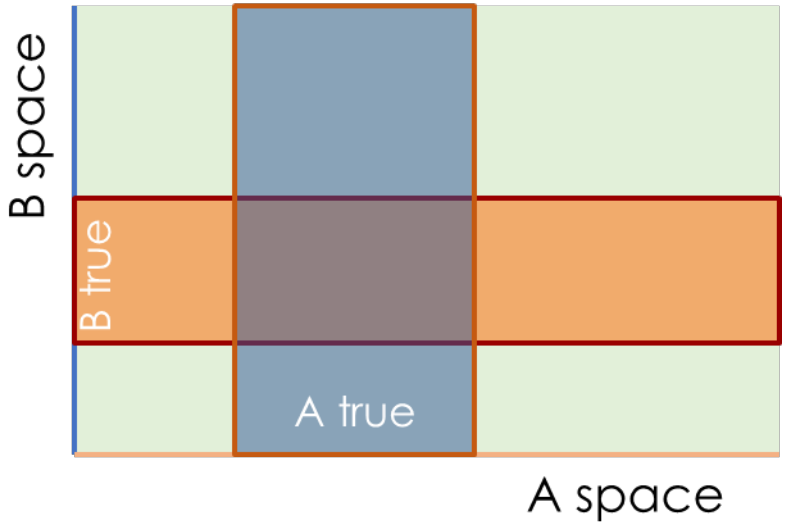


- Ví dụ:
  - A: Tôi sẽ chơi bóng đá vào ngày mai.
  - B: John sẽ không chơi bóng đá.
  - P (A, B): xác suất mà ngày mai tôi sẽ chơi bóng còn John thì không.

Xác suất đồng thời (2)

- Ký hiệu S_A là không gian của A
- Ký hiệu S_B là không gian của B
- Ký hiện 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

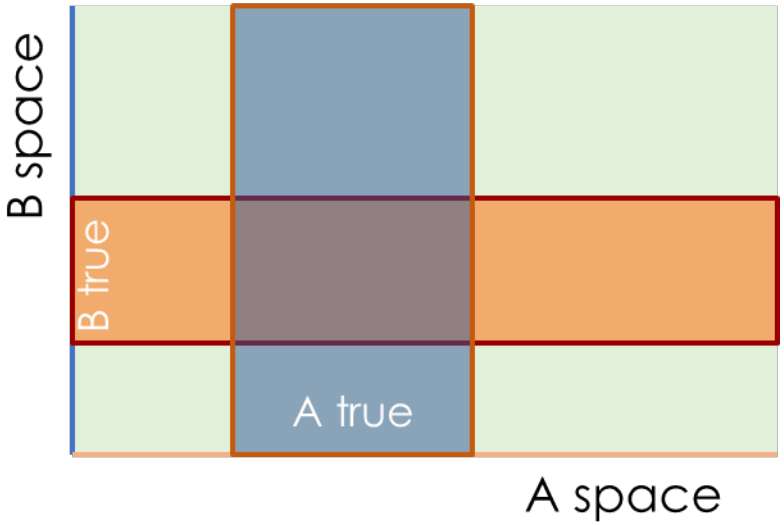
- ***Xác suất có điều kiện (Conditional probability):***
  - $P(A|B)$: khả năng A xảy ra khi B đã xảy ra.
  - $P(A|B)$: là tỉ lệ của không gian trong đó A xảy ra, biết rằng B đúng.
- Ví dụ:
  - A: Tôi sẽ chơi bóng đá vào ngày mai.
  - B: ngày mai trời sẽ không mưa.
  - P (A | B): xác suất để tôi đá bóng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

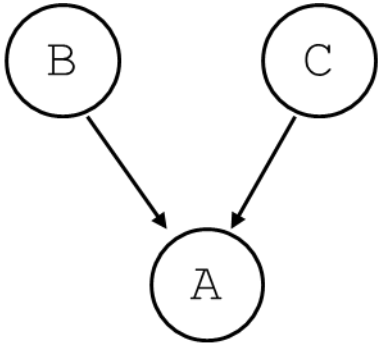
$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- 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.



P(A|B,C)

- 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.

Độc lập thống kê

- Hai sự kiện A và B được gọi là **Độc lập thống kê** (statistically independent) 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** (conditionally independent) 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:**

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

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

Định lý Bayes

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

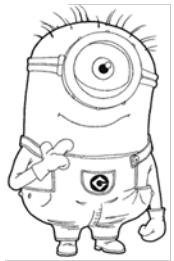
- $P(\theta)$: **là xác suất tiên nghiệm** (prior probability) của θ
  - Sự không chắc chắn của chúng ta về θ trước khi quan sát dữ liệu.
- $P(D)$: xác suất tiên nghiệm mà chúng ta có thể quan sát dữ liệu D.
- $P(D|\theta)$: **xác suất (likelihood)** chúng ta có thể quan sát được dữ liệu D khi biết trước biến θ
- $P(\theta|D)$: **xác suất hậu nghiệm** (posterior probability) của θ khi đã quan sát được dữ liệu D
  - Cách tiếp cận Bayes dựa trên đại lượng này.

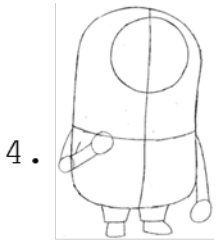
Mô hình xác suất

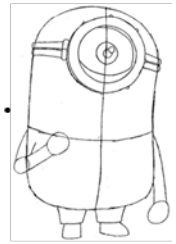
Mô hình xác suất

- 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 hoàn chỉnh*
- TIM được tạo ra như thế nào?



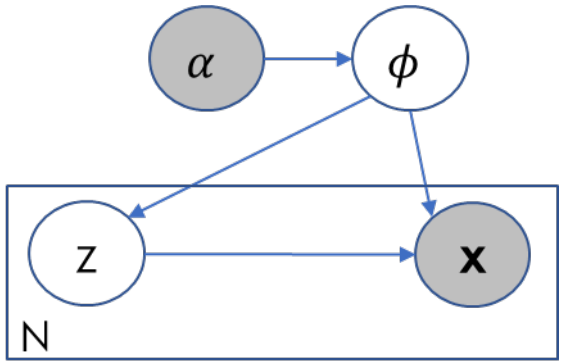




drawinghowtodraw.com

Mô hình xác suất

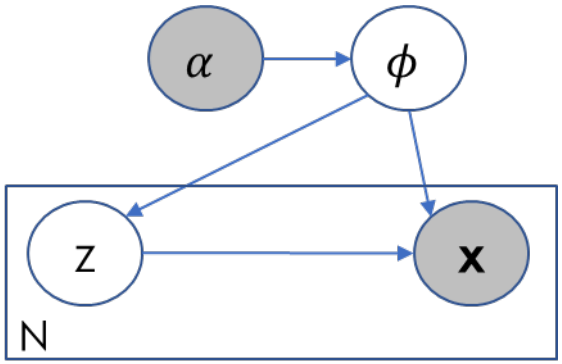
- Một mô hình đôi khi bao gồm
  - **Biến quan sát được:** mô tả những thứ quan sát hoặc thu thập được (ví dụ: $\boldsymbol{x}$)
  - **Biến ẩn**: mô tả những thứ không quan sát được (ví dụ: z, ϕ)
  - **Biến cục bộ:** liên kết với một quan sát (ví dụ: z, $\boldsymbol{x}$)
  - **Biến toàn cục:** chung cho 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**
  (Probabilistic graphical model - PGM)
  - Mỗi đỉnh đại diện cho một biến ngẫu nhiên, màu xám biểu diễn biến quan sát được, màu trắng biểu diễn biến ẩn



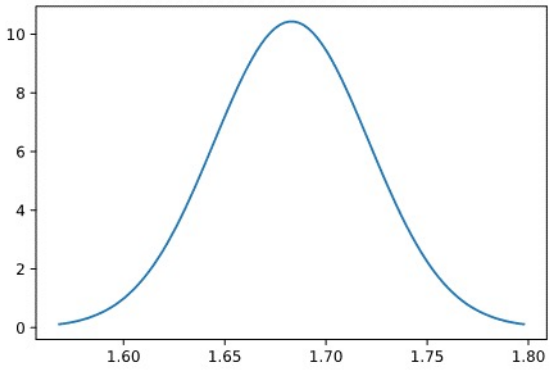
  - Mỗi cạnh đại diện cho mối quan hệ phụ thuộc có điều kiện giữa hai biến
  - Mô hình đồ thị có hướng: mỗi cạnh tuân theo một chiều
  - Mô hình đồ thị vô hướng: không có chiều trên các cạnh.
- **Mô hình biến ẩn** (latent variable model): một PGM có ít nhất 1 biến ẩn
- **Mô hình Bayes**: một PGM có xác suất tiên nghiệm trên các tham số mô hình.

Phân phối chuẩn đơn biến

- 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
- *Giả thuyết:* x tuân theo phân phối chuẩn (Gaussian) với hàm mật độ xác suất (PDF):

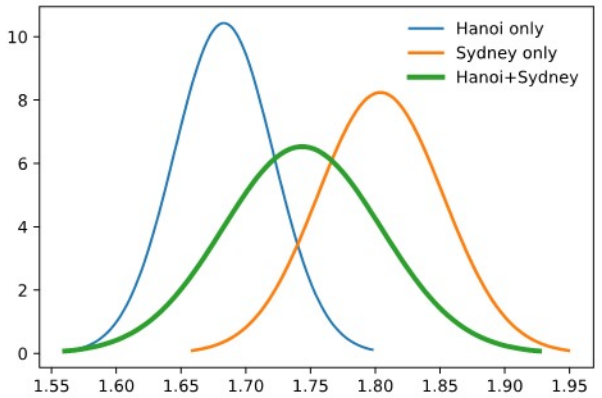
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Trong đó $\{\mu, \sigma^2\}$ là trung bình và phương sai
- Ghi chú:
  - $\mathcal{N}(x|\mu,\sigma^2)$ đại diện cho lớp phân phối chuẩn
  - Lớp này có tham số đại diện là $\boldsymbol{\theta} = (\mu, \sigma^2)$
- Bài toán học: chúng ta cần biết các giá trị cụ thể của $\{\mu, \sigma^2\}$

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

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

  Quá trình sinh:

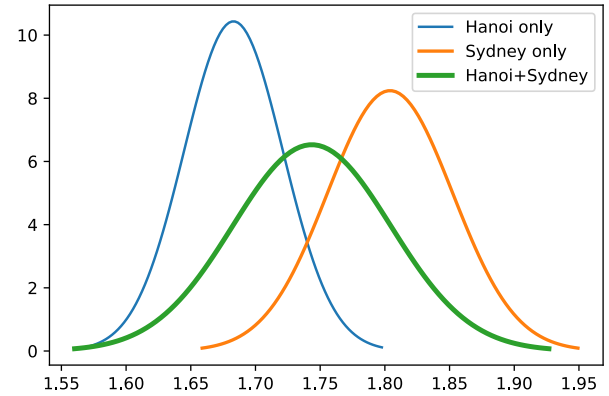
  - Chọn chỉ số $z \sim Multinomial(z|\phi)$
  - Sinh mẫu $x \sim Normal(x \mid \mu_z, \sigma_z^2)$

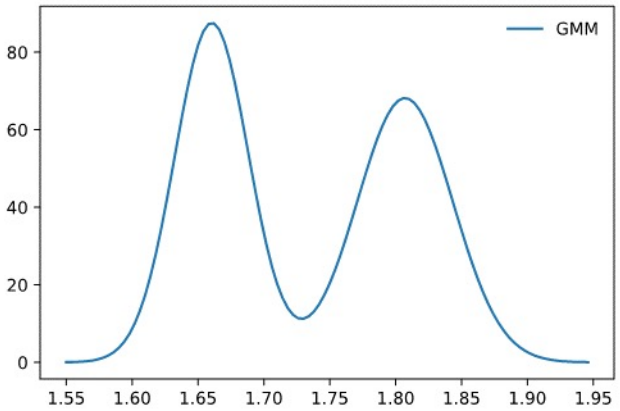
- *Đây là mô hình hỗn hợp Gauss* (Gaussian mixture model - GMM)
  - (μ_1, σ_1^2)đại diện cho phân phối Gaussian thứ nhất
  - (μ_2, σ_2^2) đại diện cho Gaussian thứ hai
  - $\phi \in [0,1]$ là tham số của phân phối Đa thức, và

  $P(z = 1|\ \phi) = \phi = 1 - P(z = 2|\ \phi)$

- Hàm mật độ:

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



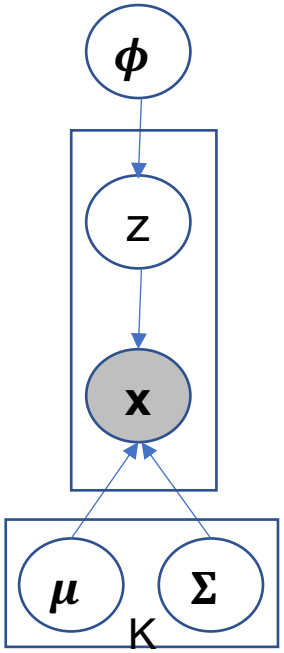


GMM nhiều biến

- Xét trường hợp mỗi **x** thuộc không gian n chiều.
- **GMM:** giả thuyết rằng dữ liệu là các mẫu được sinh ra từ K phân bố Gauss khác nhau.
- Mỗi **x** được tạo ra từ một trong K phân bố đó theo ***quá trình sinh*** như sau:
  - Sinh chỉ số $z \sim Multinomial(z|\boldsymbol{\phi})$
  - Sinh ra $\boldsymbol{x} \sim Normal(\boldsymbol{x} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- Hàm mật độ:

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

  - $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ chứa trọng số của từng phân bố con



- Mỗi Gaussian có hàm mật độ dạng:

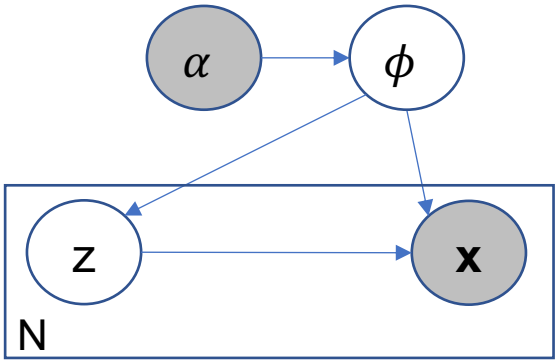
$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right]$$

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 quan sát cụ thể $\boldsymbol{x}_n$
  - Tìm biến cục bộ (ví dụ: z_n)
  - Tìm phân phối của các biến cục bộ (VD: $p(z_n, \boldsymbol{x}_n|\phi)$)
  - Ví dụ: đối với GMM, ta muốn biết z_n đại diện cho phân bố con nào đã tạo ra $\boldsymbol{x}_n$



- **Học (ước lượng):**
  Cho trước một tập dữ liệu, hãy ước lượng phân phối đồng thời của các biến
  - Ví dụ: ước lượng $p(\phi, z_1, \ldots, z_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\alpha)$
  - Ví dụ: ước lượng $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\alpha)$
  - Ví dụ: ước lượng α
  - Suy diễn của các biến cục bộ thường là cần thiết

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)

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

  - 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 ∝ 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)

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

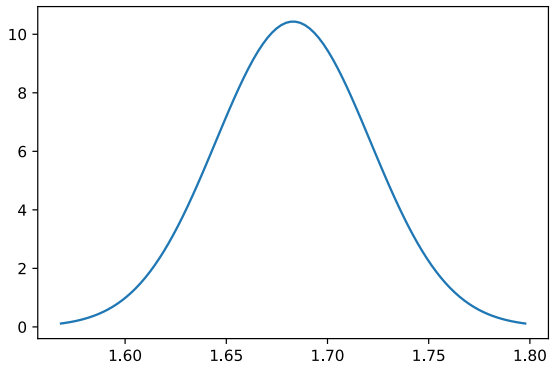
$$= \arg\max_{h \in \mathbf{H}} P(\mathbf{D}|h)\, P(h)$$

  - Tìm h^* tối đa hóa xác suất hậu nghiệm của h.
  - MAP tìm một điểm, không phải phân phối → 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 $P(h|D)$.
- *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
- **Quá trình học:** tìm (μ, σ) từ dữ liệu đã cho $\boldsymbol{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10}\}$.
- Gọi $f(\boldsymbol{x}|\mu,\sigma)$ là hàm mật độ của họ Gaussian, được tham số hóa bởi (μ,σ).
  - $f(\boldsymbol{x}_n|\mu,\sigma)$ là khả năng xảy ra của trường hợp $\boldsymbol{x}_n$
  - $f(\boldsymbol{D}|\mu,\sigma)$ là hàm khả năng xảy ra của **D**.
- Sử dụng MLE, chúng ta đi tìm

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



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 $P(\boldsymbol{D}|\mu,\sigma) = P(x_1,\dots,x_{10}|\mu,\sigma) = \prod_{i=1}^{10} P(x_i|\mu,\sigma)$
- Sử dụng giả thuyết này, MLE sẽ tìm

$$(\mu_*,\sigma_*) = \arg\max_{\mu,\sigma} \prod_{i=1}^{10} f(x_i|\mu,\sigma) = \arg\max_{\mu,\sigma} \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \arg\max_{\mu,\sigma} \log \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{10} \left(-\frac{1}{2\sigma^2}(x_i-\mu)^2 - \log\sqrt{2\pi\sigma^2}\right)$$

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

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

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

MAP: Gaussian Naïve Bayes

- Xét bài toán phân lớp
  - Dữ liệu huấn luyện **D** = {($\mathbf{x}_1$, y_1), ($\mathbf{x}_2$, y_2), …, ($\mathbf{x}_M$, y_M)} với M quan sát, C lớp.
  - Mỗi $\mathbf{x}_i$ là một vectơ trong không gian n chiều $\mathbb{R}^n$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$
- *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 phân phối Gauss có tham số $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
  - $\boldsymbol{\mu}_c$ là vectơ trung bình, $\boldsymbol{\Sigma}_c$ là ma trận hiệp phương sai kích thước $n \times n$.
- *Quá trình học:* ta xét $P(\boldsymbol{\mu}, \boldsymbol{\Sigma}, c|\boldsymbol{D})$, với $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C)$

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

Định lý Bayes, bỏ P(**D**), giả thuyết phân bố tiên nghiệm đều cho $\boldsymbol{\mu}, \boldsymbol{\Sigma}$

  - Ước lượng P(c) là tỷ lệ của lớp c trong **D**: $P(c) = |\boldsymbol{D}_c|/|\boldsymbol{D}|$ trong đó $\boldsymbol{D}_c$ 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

$$(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} P(\boldsymbol{D}_c|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) P(c) = \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} P(\boldsymbol{D}_c|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

MAP: Gaussian Naïve Bayes (2)

- Giả sử các mẫu là độc lập, ta có:

$$(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) = \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \prod_{\boldsymbol{x} \in \boldsymbol{D}_c} P(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \log P(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

$$= \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \log \left[\frac{1}{\sqrt{\det(2\pi \boldsymbol{\Sigma}_c)}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \right) \right]$$

$$= \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) - \log \sqrt{\det(2\pi \boldsymbol{\Sigma}_c)}$$

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

$$\boldsymbol{\mu}_{c*} = \frac{1}{|\boldsymbol{D}_c|} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \boldsymbol{x}, \qquad \boldsymbol{\Sigma}_{c*} = \frac{1}{|\boldsymbol{D}_c|} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} (\boldsymbol{x} - \boldsymbol{\mu}_{c*})(\boldsymbol{x} - \boldsymbol{\mu}_{c*})^T$$

- 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 sau huấn luyện: $(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, P(c))$ 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

$$c_z = \arg\max_{c\in\{1,\dots,C\}} P(c|\boldsymbol{z}, \boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) = \arg\max_{c\in\{1,\dots,C\}} P(\boldsymbol{z}|\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, c)P(c)$$

$$= \arg\max_{c\in\{1,\dots,C\}} \log P(\boldsymbol{z}|\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, c) + \log P(c)$$

$$= \arg\max_{c\in\{1,\dots,C\}} -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_{c*})^T \boldsymbol{\Sigma}_{c*}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_{c*}) - \log\sqrt{\det(2\pi\boldsymbol{\Sigma}_{c*})} + \log P(c)$$

- 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 (thuộc tính rời rạc)
  - Tập huấn luyện **D** = {($\mathbf{x}_1$, y_1), ($\mathbf{x}_2$, y_2), …, ($\mathbf{x}_M$, 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: $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$ 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 $\boldsymbol{\theta}_c$ và có hàm khối lượng xác suất

$$f(x_1, \ldots, x_V | \theta_{c1}, \ldots, \theta_{cV}) = \frac{\Gamma(\sum_{j=1}^{V} x_j + 1)}{\prod_{j=1}^{V} \Gamma(x_j + 1)} \prod_{k=1}^{V} \theta_{ck}^{x_k}$$

  - $\theta_{cj} = P(x = j | \theta_{cj})$ là xác suất mà từ $j \in \{1, \ldots, V\}$ xuất hiện, và thỏa mãn $\sum_{k=1}^{V} \theta_{ck} = 1$, và Γ là hàm gamma.
- *Quá trình học:* chúng ta có thể làm tương tự với Gaussian Naïve Bayes để ước lượng $\boldsymbol{\theta}_c = (\theta_{c1}, \ldots, \theta_{cV})$ và P(c).

MAP: Multinomial Naïve Bayes (2)

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

$$c_z = \arg\max_{c\in\{1,\ldots,C\}} P(c|\boldsymbol{z}, \boldsymbol{\theta}_{c*}) = \arg\max_{c\in\{1,\ldots,C\}} P(\boldsymbol{z}|\boldsymbol{\theta}_{c*}, c)P(c)$$

$$= \arg\max_{c\in\{1,\ldots,C\}} \log P(\boldsymbol{z}|\boldsymbol{\theta}_{c*}) + \log P(c) \quad \text{(MNB.1)}$$

$$= \arg\max_{c\in\{1,\ldots,C\}} \log \frac{\Gamma(\sum_{j=1}^{V} z_j + 1)}{\prod_{j=1}^{V} \Gamma(z_j + 1)} \prod_{k=1}^{V} \theta_{ck*}^{z_k} + \log P(c)$$

$$= \arg\max_{c\in\{1,\ldots,C\}} \log \prod_{k=1}^{V} \theta_{ck*}^{z_k} + \log P(c)$$

$$= \arg\max_{c\in\{1,\ldots,C\}} \log \prod_{k=1}^{V} P(z_k|\theta_{ck*}) + \log P(c) \quad \text{(MNB.2)}$$

  - Nhãn có xác suất hậu nghiệm cao nhất

- Lưu ý: về cơ bản chúng ta ngầm giả thuyết rằng *các thuộc tính là độc lập với nhau* (nhìn 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 $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$.
- Hàm mật độ là $p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \phi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ đại diện cho trọng số của các Gaussian
  - Mỗi Gaussian đa biến có mật độ:

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

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

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \sum_{i=1}^{M} \log \sum_{k=1}^{K} \phi_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- 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
  - Nhiều mô hình 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.

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions !

soict.hust.edu.vn/ fb.com/groups/soict