Buổi 8: Ước lượng tham số & Phân cụm dữ liệu

Trung Cấn
Trường Đại học Khoa học tự nhiên, ĐHQG-HCM

Ngày 13 tháng 5, 2025

Outlines

Các nội dung

Outlines

Gia đình phân phối

Gia đình phân phối $\mathcal{D}$ là tập hợp các phân phối có *cùng loại* công thức hàm khối/mật độ xác suất p_a, còn viết là $p(\cdot; a)$, với a là tham số *thay đổi*.

Ví dụ:

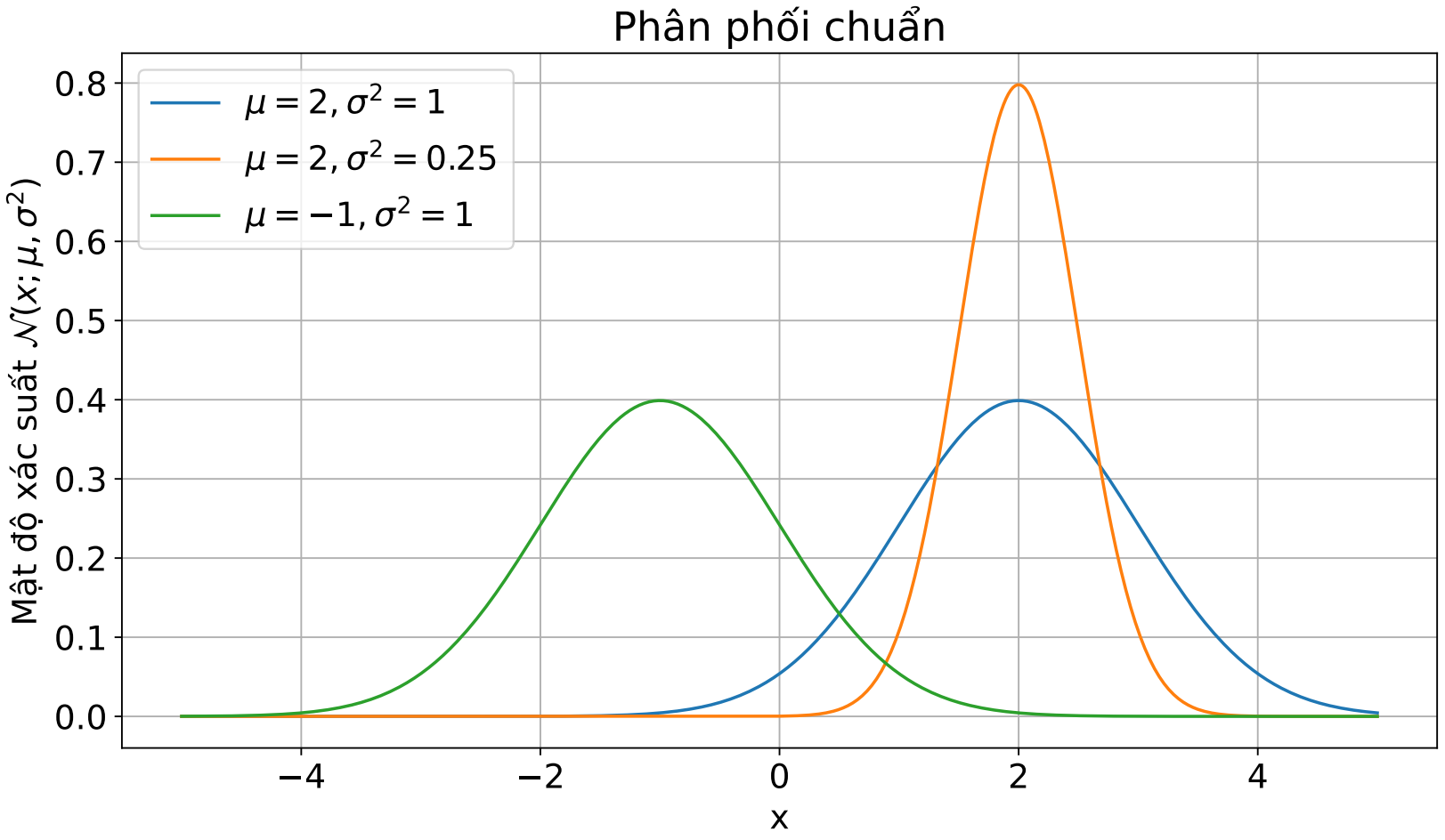
- Gia đình phân phối Poisson: $X \sim \text{Pois}(\lambda)$ có hàm khối là

$$p(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \lambda \in [0, \infty)$$

- Gia đình phân phối chuẩn: $X \sim \mathcal{N}(\mu, \sigma^2)$ có hàm mật độ là

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], (\mu, \sigma^2) \in \mathbb{R} \times [0, \infty)$$

Gia đình các phân phối chuẩn



Tình huống áp dụng

An nhặt được một đồng xu trên đường. An tung thử 10 lần thì thu được kết quả lần lượt là $N, S, N, N, N, N, S, S, S, N$. Tìm xác suất ra mặt ngửa khi tung đồng xu này?

Tình huống áp dụng

An nhặt được một đồng xu trên đường. An tung thử 10 lần thì thu được kết quả lần lượt là $N, S, N, N, N, N, S, S, S, N$. Tìm xác suất ra mặt ngửa khi tung đồng xu này?

Mô hình bài toán

- $X_1, \ldots, X_{10} \sim \text{Bernoulli}(a)$

$$p(x; a) = \begin{cases} a, & x = 1 \\ 1 - a, & x = 0 \\ 0, & \text{khác} \end{cases}$$

- $\Theta = (1, 0, 1, 1, 1, 1, 0, 0, 0, 1)$
- Tìm a như thế nào là *hợp lý nhất*?

Bài toán ước lượng tham số

Đầu vào

- $\mathcal{D} =$ gia đình phân phối p_a được tham số hóa bởi a (biến số);
- $\Theta = (x_1, x_2, \ldots, x_n)$ với $x_1, \ldots, x_n \in \mathbb{R}$ bộ dữ liệu;
- $\mathcal{L}(a, \Theta)$ hàm số theo a để đánh giá độ hợp lý;

Giả định: các giá trị trong Θ là các mẫu ngẫu nhiên được lấy *độc lập và cùng từ* p_a

Đầu ra: xác định được a để *tối ưu hóa* $\mathcal{L}(a, \Theta)$

Outlines

Hàm hợp lí ứng với điểm dữ liệu

Định nghĩa (hàm hợp lí ứng với điểm dữ liệu)

Xét họ phân phối p_a và $x \in \mathbb{R}$. **Hàm hợp lí** của p_a ứng với x, ký hiệu $\mathcal{L}(\cdot|a)$ hay $\mathcal{L}_x$, là một *hàm theo* a được định nghĩa là

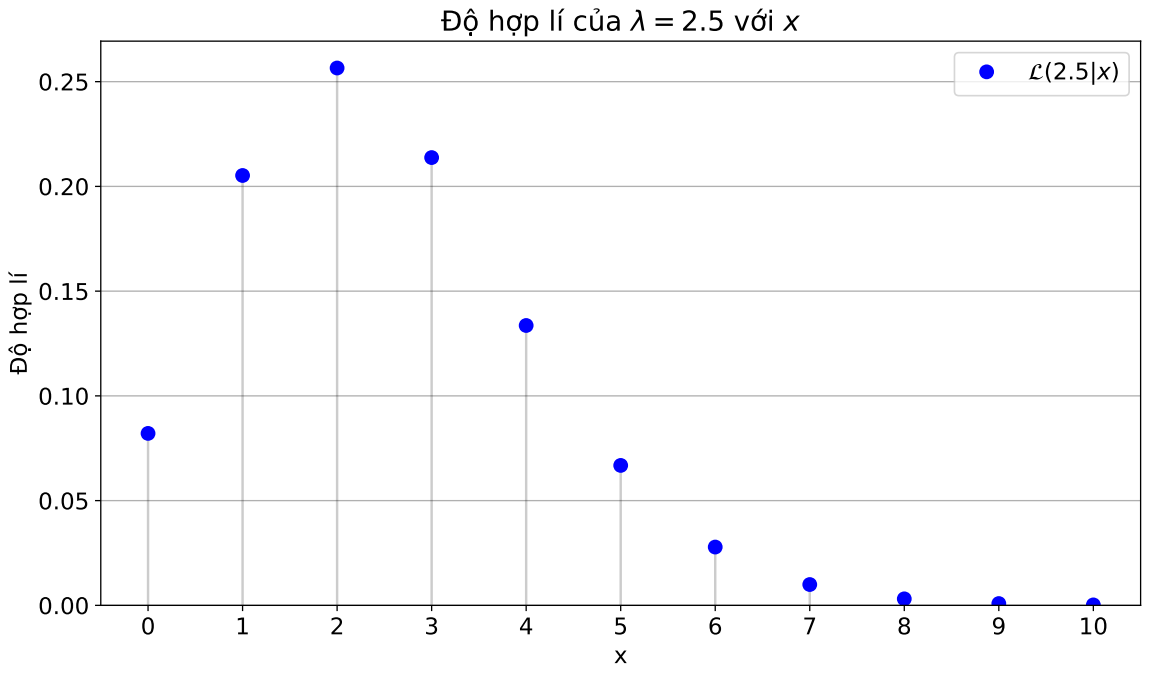
$$\mathcal{L}(a|x) := p_a(x) \text{ với mọi } a \in a$$

Ta gọi $\mathcal{L}(a|x)$ là **độ hợp lí** của a với x.

Nhận xét: Độ hợp lí phản ánh khả năng sinh ra giá trị x khi phân phối có tham số a.

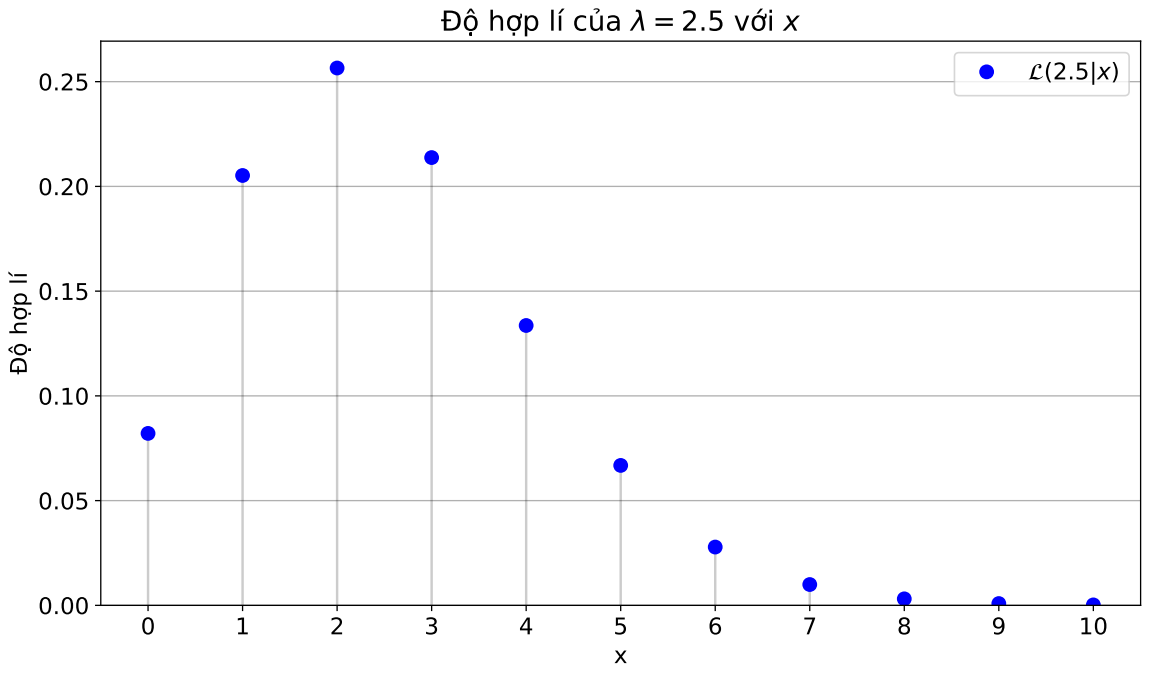
Độ hợp lý

Ví dụ: Xét bnn $X \sim \text{Poisson}(\lambda)$. Độ hợp lí của $\lambda = 2.5$ với x từ 0 đến 10 là



Độ hợp lý

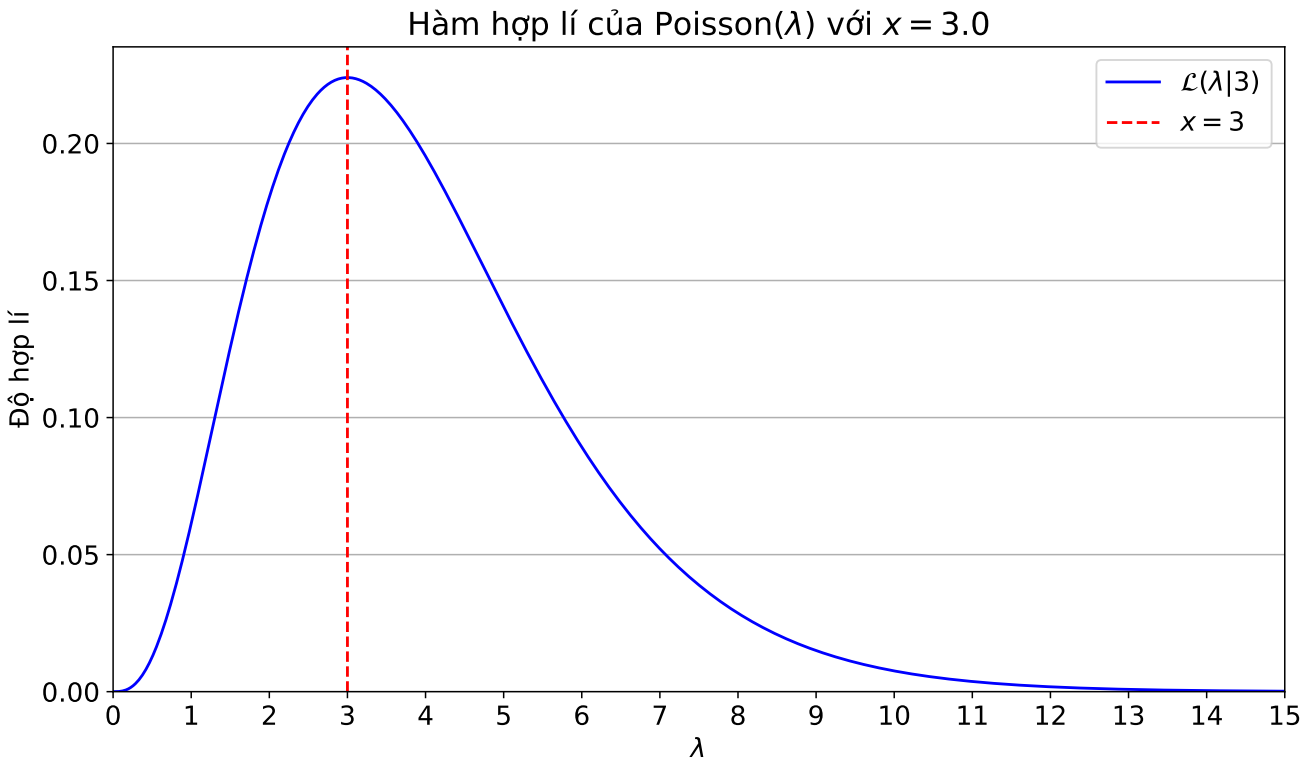
Ví dụ: Xét bnn $X \sim \text{Poisson}(\lambda)$. Độ hợp lí của $\lambda = 2.5$ với x từ 0 đến 10 là



Lưu ý: Là đồ thị hàm khối xác suất của $\text{Poisson}(2.5)$.

Độ hợp lý (tt)

Ví dụ: Xét bnn $X \sim \text{Poisson}(\lambda)$ và cố định giá trị cụ thể $x = 3$. Đồ thị của $\mathcal{L}_x$ là



Hàm hợp lí ứng với bộ dữ liệu

Nhắc lại: các mẫu trong $\Theta = (x_1, \ldots, x_n)$ là các mẫu ngẫu nhiên được lấy *độc lập* và *cùng từ* p_a

$$p(x_1, \ldots, x_n; a) = \prod_{i=1}^{n} p(x_i; a).$$

Định nghĩa (hàm hợp lí ứng với bộ dữ liệu)

Xét họ phân phối p_a và bộ dữ liệu Θ thỏa giả định trên. **Hàm hợp lí** của p_a ứng với $\Theta = (x_1, \ldots, x_n)$, ký hiệu $\mathcal{L}(\cdot|\Theta)$ hay $\mathcal{L}_\Theta$, là một hàm số của a được xác định bởi

$$\mathcal{L}(a|\Theta) = \prod_{i=1}^{n} \mathcal{L}(a|x_i) \text{ với mọi } a$$

Ta gọi $\mathcal{L}(a|\Theta)$ là **độ hợp lí** của a với Θ.

Ước lượng hợp lí cực đại

Còn gọi tắt là **MLE** (maximum likelihood estimation)

Ước lượng hợp lí cực đại

Còn gọi tắt là **MLE** (maximum likelihood estimation)

Bài toán tối ưu hóa cho MLE

Ta có:

- p_a: họ phân phối tham số hóa bởi $a \in D$;
- $\Theta = (x_1, \ldots, x_n)$: các giá trị được giả sử là lấy mẫu từ p_a;

Tham số ước lượng theo phương pháp MLE:

$$a_{MLE} = \arg\max_{a \in D} \mathcal{L}(a|\Theta)$$

Ước lượng hợp lí cực đại

Còn gọi tắt là **MLE** (maximum likelihood estimation)

Bài toán tối ưu hóa cho MLE

Ta có:

- p_a: họ phân phối tham số hóa bởi $a \in D$;
- $\Theta = (x_1, \ldots, x_n)$: các giá trị được giả sử là lấy mẫu từ p_a;

Tham số ước lượng theo phương pháp MLE:

$$a_{MLE} = \arg\max_{a \in D} \mathcal{L}(a|\Theta)$$

Trong nhiều trường hợp, ta sử dụng hàm *đồng biến* log,

$$a_{MLE} = \arg\max_{a \in a} \log \mathcal{L}(a|\Theta)$$

Ví dụ

An nhặt được một đồng xu trên đường. Anh tung thử 10 lần thì thu được kết quả lần lượt là $N, S, N, N, N, N, S, S, S, N$. Hỏi: xác suất ra mặt ngửa khi tung đồng xu này?

Trả lời

- $p_a \sim$ Bernoulli(a) với $a \in [0, 1]$

$$p(x; a) = \begin{cases} a, & x = 1 \\ 1 - a, & x = 0 \\ 0, & \text{khác} \end{cases}$$

- $\Theta = (1, 0, 1, 1, 1, 1, 0, 0, 0, 1)$
- Hàm hợp lí

Ví dụ

An nhặt được một đồng xu trên đường. Anh tung thử 10 lần thì thu được kết quả lần lượt là $N, S, N, N, N, N, S, S, S, N$. Hỏi: xác suất ra mặt ngửa khi tung đồng xu này?

Trả lời

- $p_a \sim$ Bernoulli(a) với $a \in [0, 1]$

$$p(x; a) = \begin{cases} a, & x = 1 \\ 1 - a, & x = 0 \\ 0, & \text{khác} \end{cases}$$

- $\Theta = (1, 0, 1, 1, 1, 1, 0, 0, 0, 1)$
- Hàm hợp lí

$$\mathcal{L}(a|\Theta) = \mathcal{L}(a|1) \times \cdots \times \mathcal{L}(a|1) =$$

Ví dụ

An nhặt được một đồng xu trên đường. Anh tung thử 10 lần thì thu được kết quả lần lượt là $N, S, N, N, N, N, S, S, S, N$. Hỏi: xác suất ra mặt ngửa khi tung đồng xu này?

Trả lời

- $p_a \sim \text{Bernoulli}(a)$ với $a \in [0, 1]$

$$p(x; a) = \begin{cases} a, & x = 1 \\ 1 - a, & x = 0 \\ 0, & \text{khác} \end{cases}$$

- $\Theta = (1, 0, 1, 1, 1, 1, 0, 0, 0, 1)$
- Hàm hợp lí

$$\mathcal{L}(a|\Theta) = \mathcal{L}(a|1) \times \cdots \times \mathcal{L}(a|1) = a^6(1 - a)^4$$

Ví dụ

- Hàm log-hợp lí

$$\log \mathcal{L}(p|\Theta) = 6 \log a + 4 \log(1 - a)$$

- Ước lượng hợp lí cực đại

$$a_{MLE} = \arg\max_{a \in [0,1]} \log \mathcal{L}(a|\Theta) = \arg\max_{a \in [0,1]} 6 \log a + 4 \log(1 - a)$$

- Kết quả

$$p_{MLE} =$$

Ví dụ

- Hàm log-hợp lí

$$\log \mathcal{L}(p|\Theta) = 6 \log a + 4 \log(1 - a)$$

- Ước lượng hợp lí cực đại

$$a_{MLE} = \arg\max_{a \in [0,1]} \log \mathcal{L}(a|\Theta) = \arg\max_{a \in [0,1]} 6 \log a + 4 \log(1 - a)$$

- Kết quả

$$p_{MLE} = \frac{6}{6 + 4} = 0.6$$

Bài tập tự luyện

1. Giải công thức tổng quát của ví dụ tung đồng xu ở trên: Xét phân phối Bernoulli với tham số p. Cho n mẫu $x_1, x_2, \ldots, x_n$ được biết là lấy từ phân phối này. Ước lượng p theo phương pháp MLE.
2. Xét phân phối Poisson với tham số λ. Cho n mẫu $x_1, x_2, \ldots, x_n$ được biết là lấy từ phân phối này. Ước lượng λ theo phương pháp MLE.
3. Xét phân phối chuẩn có 2 tham số là μ và σ^2. Cho n mẫu $e_1, e_2, \ldots, e_n$ được biết là lấy từ phân phối này. Lần lượt cố định giá trị của μ và σ^2 rồi ước lượng tham số còn lại.
   Nhận xét: Trường hợp $\mu = 0$ tương đương với mô hình hồi quy tuyến tính $y = ax + b + e$ với $e \sim \mathcal{N}(0, \sigma^2)$

Outlines

Tình huống áp dụng

Dựa vào kinh nghiệm, An tin là phân bố điểm thi tuân theo một phân bố chuẩn $\mathcal{N}(\mu_0, \tau^2)$ với trung bình μ và phương sai τ^2 cho trước. Thực tế, các dữ liệu quan sát được là $\Theta = \{x_1, \ldots, x_n\}$ có thể tuân theo phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ với biến số μ có thể khác với niềm tin ban đầu.

Mục tiêu: Ước lượng μ hợp lý nhất

Bài toán ước lượng cực đại hậu nghiệm

Còn gọi là **MAP** (maximum a posterior estimation).

Ta có:

- Dữ liệu $\Theta = x_1, x_2, \ldots, x_n \sim p_a | a \in D$;
- Phân phối tiên nghiệm: $\mu \sim p_{a_0}$;
- Hàm hợp lý $\mathcal{L}(a|a_0, \Theta) = \prod_i p_a(x_i|a_0)$

Tìm

$$a_{\text{MAP}} = \arg\max_{a \in D} \mathcal{L}(a|a_0, \Theta).$$

MAP cho phân phối chuẩn

Phân phối tiền nghiệm:

$$P(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau^2}\right)$$

Hợp lý hậu nghiệm cho điểm dữ liệu: là hàm mật độ xác suất có điều kiện

$$P(x_i \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Hợp lý hậu nghiệm cho bộ dữ liệu:

$$P(\Theta \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Tính toán MAP

Hàm log-hậu nghiệm:

$$\log P(\mu \mid x) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\tau^2}$$

Lấy đạo hàm và giải phương trình:

$$\mu_{\text{MAP}} = \frac{n\bar{x}/\sigma^2 + \mu_0/\tau^2}{n/\sigma^2 + 1/\tau^2}$$

Nhận xét: Đây là công thức tổng quát cho MAP với tiền nghiệm và phỏng đoán đều là phân phối chuẩn

Ví dụ cụ thể

- Dữ liệu: $\Theta = \{6, 7, 8\} \sim \mathcal{N}(\mu, 1)$;
- Tiền nghiệm: $\mu_0 = 5, \tau^2 = 4$
- Áp dụng công thức:

$$\mu_{\text{MAP}} = \frac{3 \cdot 7 + 5/4}{3 + 1/4} = \frac{21 + 1.25}{3.25} = \frac{22.25}{3.25} \approx 6.846$$

So sánh với MLE

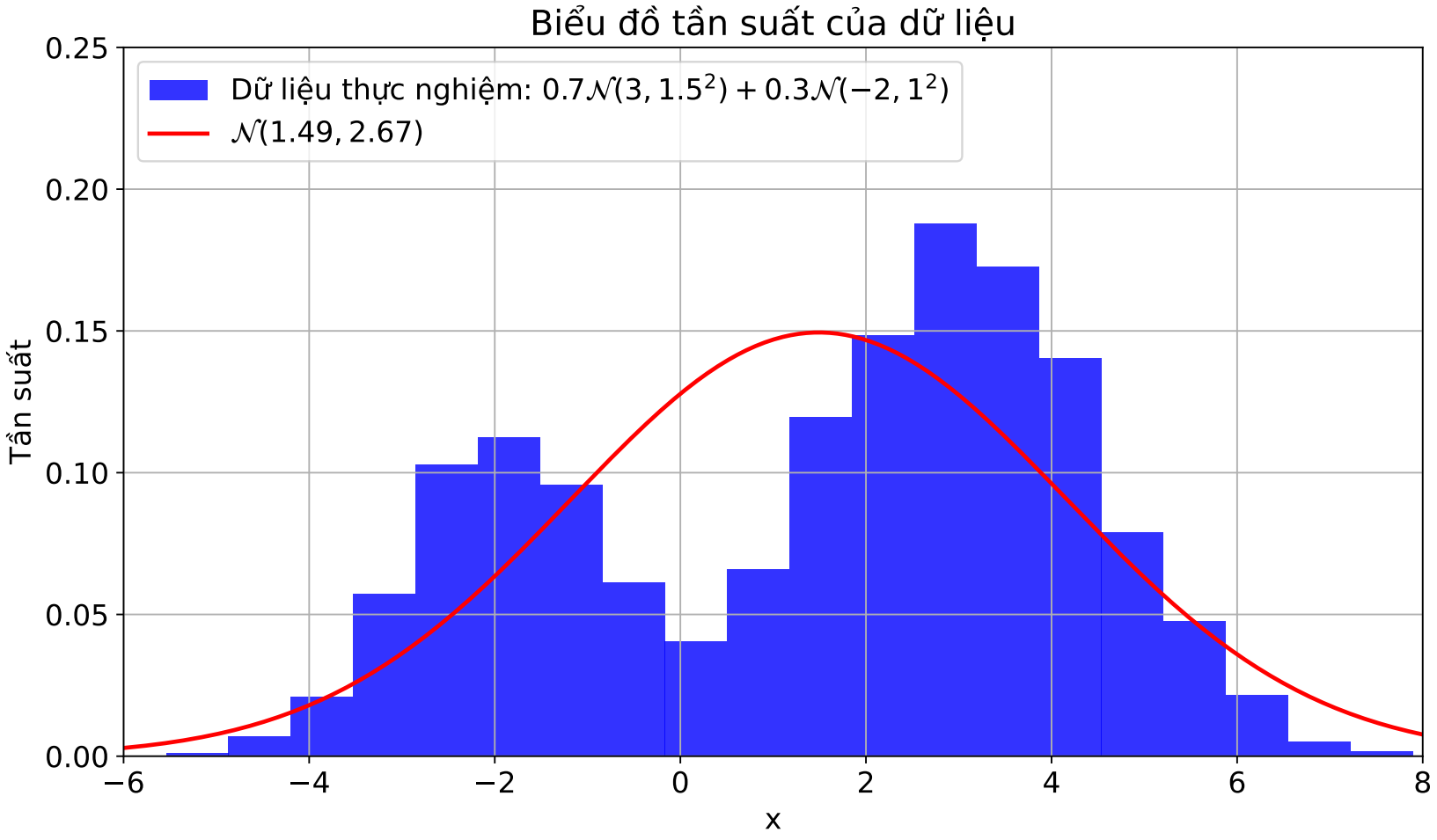
- MLE (chỉ dùng dữ liệu): $\mu_{\text{MLE}} = \bar{x} = 7$
- MAP: $\mu_{\text{MAP}} \approx 6.846$
- MAP bị kéo về giá trị tiền nghiệm $\mu_0 = 5$
- MAP giúp *làm mượt* ước lượng khi dữ liệu quan sát được ít

Tóm tắt

- MAP kết hợp *dữ liệu quan sát được* và *niềm tin tiền nghiệm*.
- Phương pháp này phổ biến trong thống kê Bayesian và xử lý tín hiệu.

Outlines

Ví dụ: hỗn hợp Gaussian



Phân bố hỗn hợp Gaussian

Đề xuất: Gia đình phân phối $\mathcal{D}$ là hỗn hợp của 2 phân phối Gaussian

$$Z \sim \text{Bernoulli}(\pi)$$
$$X|Z = z \sim \mathcal{N}(\mu_z, \sigma_z^2)$$

- Tham số:

$$a = (\pi, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2) \in a = [0,1] \times \mathbb{R} \times (0,\infty) \times \mathbb{R} \times (0,\infty)$$

- Hàm mật độ:

$$p(x; a) = \sum_{k=0}^{1} \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2)$$

với $\pi_0 = 1 - \pi$ và $\pi_1 = \pi$.

Ước lượng phấn bố hỗn hợp Gaussian

Hàm hợp lí của p_a với bộ dữ liệu $\Theta = (x_1, \ldots, x_n)$

$$\mathcal{L}(a|\Theta) = \prod_{i=1}^{n} \left[\sum_{k=0}^{1} \pi_k \mathcal{N}(x_i; \mu_k, \sigma_k^2) \right]$$

Ước lượng phấn bố hỗn hợp Gaussian

Hàm hợp lí của p_a với bộ dữ liệu $\Theta = (x_1, \ldots, x_n)$

$$\mathcal{L}(a|\Theta) = \prod_{i=1}^{n} \left[\sum_{k=0}^{1} \pi_k \mathcal{N}(x_i; \mu_k, \sigma_k^2) \right]$$

Hàm log-hợp lí của p_a với bộ dữ liệu $\Theta = (x_1, \ldots, x_n)$

$$\log \mathcal{L}(a|\Theta) = \sum_{i=1}^{n} \log \left[\sum_{k=0}^{1} \pi_k \mathcal{N}(x_i; \mu_k, \sigma_k^2) \right]$$

Vấn đề Khó tối ưu (không có công thức nghiệm chính xác)

Cách giải quyết 1

Hàm cần tối ưu

$$\log \mathcal{L}(a|\Theta) = \sum_{i=1}^{n} \log \left[\sum_{k=0}^{1} \pi_k \mathcal{N}(x_i; \mu_k, \sigma_k^2) \right]$$

Đề xuất

- Tối ưu theo độ dốc (gradient descent)

Cách giải quyết 2: Thuật toán Kì vọng - Cực đại hóa

Hàm cần tối ưu

$$\begin{aligned}
\log \mathcal{L}(a|\Theta) &= \sum_{i=1}^{n} \log p_{X_i}(x_i; a) \\
&= \sum_{i=1}^{n} \log \left[\sum_{z_i} p_{X_i,Z_i}(x_i, z_i; a) \right] \\
&= \sum_{i=1}^{n} \log \left[\sum_{z_i} p_{X_i|Z_i=z_i}(x_i; a) p_{Z_i}(z_i; a) \right] \\
&= \sum_{i=1}^{n} \log \mathbb{E}_{\tilde{z}_i \sim p_{Z_i;a}} \left[p_{X_i|Z_i=\tilde{z}_i}(x_i; a) \right]
\end{aligned}$$

Cách giải quyết 2: Thuật toán Kì vọng - Cực đại hóa

Xét một phân phối $q_{Z_i;\phi}$.

$$\begin{aligned}
\log \mathcal{L}(a|\Theta) &= \sum_{i=1}^{n} \log \mathbb{E}_{\tilde{z}_i \sim p_{Z_i;a}} \left[p_{X_i|Z_i=\tilde{z}_i}(x_i; a) \right] \\
&= \sum_{i=1}^{n} \log \mathbb{E}_{\tilde{z}_i \sim q_{Z_i;\phi}} \left[p_{X_i|Z_i=\tilde{z}_i}(x_i; a) \frac{p_{Z_i}(\tilde{Z}_i; a)}{q_{Z_i}(\tilde{Z}_i; \phi)} \right] \\
&\geq \sum_{i=1}^{n} \mathbb{E}_{\tilde{z}_i \sim q_{Z_i;\phi}} \log \left[p_{X_i|Z_i=\tilde{z}_i}(x_i; a) \frac{p_{Z_i}(\tilde{Z}_i; a)}{q_{Z_i}(\tilde{Z}_i; \phi)} \right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{\tilde{z}_i \sim q_{Z_i;\phi}} \left[\log p_{X_i,Z_i}(x_i, \tilde{Z}_i; a) - \log q_{Z_i}(\tilde{Z}_i; \phi) \right]
\end{aligned}$$

Mục tiêu mới tối ưu chặn dưới này

Thuật toán Kì vọng - Tối ưu hóa (EM)

Hàm mục tiêu

$$F\left((q_{Z_i;\phi})_{i=1}^n, a\right) = \sum_{i=1}^n \mathbb{E}_{Z_i \sim q_{Z_i;\phi}} \left[\log p_{X_i,Z_i}(x_i, Z_i; a) - \log q_{Z_i}(Z_i; \phi)\right]$$

Tối ưu từng thành phần (coordinate descent)

1. Khởi tạo $a^{(0)}$ hợp lệ bất kì
2. Lặp lại với $t = 1, 2, \ldots$ đến khi hội tụ
   1. Bước kì vọng (E): cố định $a^{(t)}$ và tìm các $q_{Z_i;\phi}$ tối ưu

$$q_{Z_i;\phi}^{(t+1)} = \arg\max_{q_1,\ldots,q_n} F\left((q_i)_{i=1}^n, a^{(t)}\right)$$

   2. Bước cực đại hóa (M): cố định các $q_{Z_i;\phi}^{(t)}$ và tìm a tối ưu

$$a^{(t+1)} = \arg\max_a F\left(\left(q_{Z_i;\phi}^{(t+1)}\right)_{i=1}^n, a\right)$$

Bước kì vọng thứ t

$$
\begin{aligned}
& F\left(\left(q_{Z_i;\phi}\right)_{i=1}^{n}, a^{(t)}\right) \\
&= \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim q_{Z_i;\phi}} \left[\log p_{X_i,Z_i}\left(x_i, Z_i; a^{(t)}\right) - \log q_{Z_i}(Z_i;\phi)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim q_{Z_i;\phi}} \left[\log p_{X_i}(x_i; a) + \log p_{Z_i|X_i=x_i}\left(z_i; a^{(t)}\right) - \log q_{Z_i}(Z_i;\phi)\right] \\
&= \log \mathcal{L}(a|\Theta) + \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim q_{Z_i;\phi}} \log \left[\frac{p_{Z_i|X_i=x_i}\left(z_i; a^{(t)}\right)}{q_{Z_i}(Z_i;\phi)}\right] \\
&= \log \mathcal{L}(a|\Theta) - \sum_{i=1}^{n} D_{\mathrm{KL}}\left(q_{Z_i;\phi} \| p_{Z_i|X_i=x_i;a^{(t)}}\right)
\end{aligned}
$$

Kết luận $q_{Z_i;\phi} \equiv p_{Z_i|X_i=x_i;a^{(t)}}$ với mọi $i \in \{1, \ldots, n\}$.

Bước cực đại hóa thứ t

$$F\left(p_{Z_i|X_i=x_i;a^{(t)}}, a\right) = \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim p_{Z_i|X_i=x_i;a^{(t)}}} \log p_{X_i,Z_i}(x_i, Z_i; a) + C$$
$$=: Q(a, a^{(t)}) + C$$

$\log p_{X,Z}(x, z; a)$ còn được gọi là độ *hợp lí đầy đủ* của tham số a và điểm dữ liệu (x, z) (giả định biết z).

Kết luận $a^{(t+1)} = \arg\max_a(a, a^{(t)})$

Tại sao thuật toán kì vọng - cực đại hóa hoạt động?

Ta chứng minh được

$$\log \mathcal{L}(a|\Theta) - \log \mathcal{L}(a^{(t)}|\Theta) \geq Q(a, a^{(t)}) - Q(a^{(t)}, a^{(t)})$$

với mọi a. (Tại sao?)

Giá trị tham số a tối ưu Q cũng sẽ tối ưu $\log \mathcal{L}(a|\Theta)$!

Ví dụ: mô hình hỗn hợp Gaussian

Nhắc lại mô hình phân cấp

$$Z \sim \text{Bernoulli}(\pi)$$
$$X|Z = z \sim \mathcal{N}(\mu_z, \sigma_z^2)$$

Hàm khối của biến ẩn

$$p_Z(z; a) = \text{Bernoulli}(z; \pi) = \begin{cases} \pi, & z = 1 \\ 1 - \pi, & z = 0 \end{cases} = \pi^{\mathbb{I}_{z=1}}(1 - \pi)^{\mathbb{I}_{z=0}}$$
$$\log p_Z(z; a) = \mathbb{I}_{z=1} \log \pi + \mathbb{I}_{z=0} \log(1 - \pi)$$

Hàm mật độ của biến quan sát được khi biết giá trị biến ẩn

$$p_{X|Z=z}(x; a) = \mathcal{N}(x; \mu_1, \sigma_1^2)^{\mathbb{I}_{z=1}} \mathcal{N}(x; \mu_0, \sigma_0^2)^{\mathbb{I}_{z=0}}$$
$$\log p_{X|Z=z}(x; a) = \mathbb{I}_{z=1} \log \mathcal{N}(x; \mu_1, \sigma_1^2) + \mathbb{I}_{z=0} \log \mathcal{N}(x; \mu_0, \sigma_0^2)$$

Ví dụ: mô hình hỗn hợp Gaussian

Hàm log-hợp lí đầy đủ

$$\log p_{X,Z}(x, z; a) = \log p_{X|Z=z}(x; a) + \log p_Z(z; a)$$
$$= \sum_{k=0}^{1} \mathbb{I}_{z=k} \left(\log \pi_k + \log \mathcal{N}(x; \mu_k, \sigma_k^2) \right)$$

Ở đây, $\pi_0 = 1 - \pi$ và $\pi_1 = \pi$.

$$Q(a, a^{(t)})$$
$$= \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim p_{Z_i|X_i=x_i;a^{(t)}}} \log p_{X_i,Z_i}(x_i, Z_i; a)$$
$$= \sum_{i=1}^{n} \sum_{k=0}^{1} \mathbb{E}_{Z_i \sim p_{Z_i|X_i=x_i;a^{(t)}}} (\mathbb{I}_{Z_i=k}) [\log \pi_k^{(t)} + \log \mathcal{N}(x_i; \mu_k^{(t)}, (\sigma_k^{(t)})^2)]$$

Ví dụ: mô hình hỗn hợp Gaussian

Với $i \in \{1, \ldots, n\}, k \in \{0, 1\}$,

$$\begin{aligned}
\gamma_{i,k}^{(t)} &:= \mathbb{E}_{Z_i \sim p_{Z_i|X_i=x_i;a^{(t)}}} \left(\mathbb{I}_{Z_i=k}\right) \\
&= p_{Z_i|X_i=x_i}(k; a^{(t)}) \\
&= \frac{p_{Z_i}(k; a^{(t)}) p_{X_i|Z_i=k}(x_i; a^{(t)})}{p_{X_i}(x_i; a^{(t)})} \\
&= \frac{\pi_k^{(t)} \mathcal{N}(x_i; \mu_k^{(t)}, (\sigma_k^{(t)})^2)}{\sum_{h=0}^{1} \left[\pi_h^{(t)} \mathcal{N}(x_i; \mu_h^{(t)}, (\sigma_h^{(t)})^2)\right]}
\end{aligned}$$

Ví dụ: mô hình hỗn hợp Gaussian

Cần tìm $a = (\pi, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$ để tối ưu

$$Q(a, a^{(t)}) = \sum_{i=1}^{n} \sum_{k=0}^{1} \gamma_{i,k}^{(t)} [\log \pi^{(t)} + \log \mathcal{N}(x_i; \mu_k^{(t)}, (\sigma_k^{(t)})^2)]$$

Với $k \in \{0, 1\}$,

$$\mu_k^{(t+1)} = \frac{1}{\sum_{i=1}^{n} \gamma_{i,k}^{(t)}} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} x_i$$

(Tại sao?)

Ví dụ: mô hình hỗn hợp Gaussian

Cần tìm $a = (\pi, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$ để tối ưu

$$Q(a, a^{(t)}) = \sum_{i=1}^{n} \sum_{k=0}^{1} \gamma_{i,k}^{(t)} [\log \pi^{(t)} + \log \mathcal{N}(x_i; \mu_k^{(t)}, (\sigma_k^{(t)})^2)]$$

Với $k \in \{0, 1\}$,

$$(\sigma_k^{(t+1)})^2 = \frac{1}{\sum_{i=1}^{n} \gamma_{i,k}^{(t)}} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} (x_i - \hat{\mu}_k)^2$$

(Tại sao?)

Ví dụ: mô hình hỗn hợp Gaussian

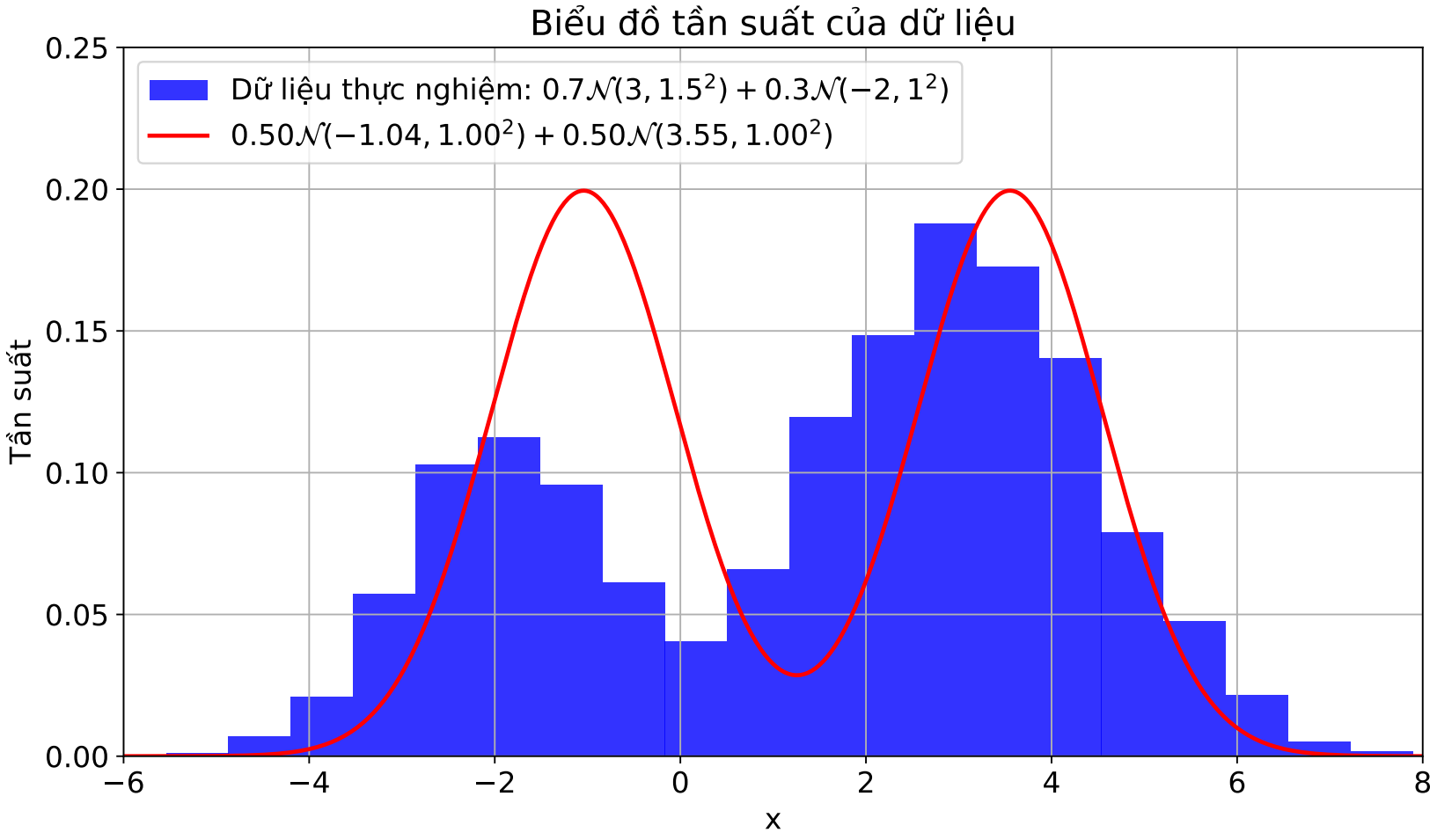
Cần tìm $a = (\pi, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$ để tối ưu

$$Q(a, a^{(t)}) = \sum_{i=1}^{n} \sum_{k=0}^{1} \gamma_{i,k}^{(t)} [\log \pi^{(t)} + \log \mathcal{N}(x_i; \mu_k^{(t)}, (\sigma_k^{(t)})^2)]$$

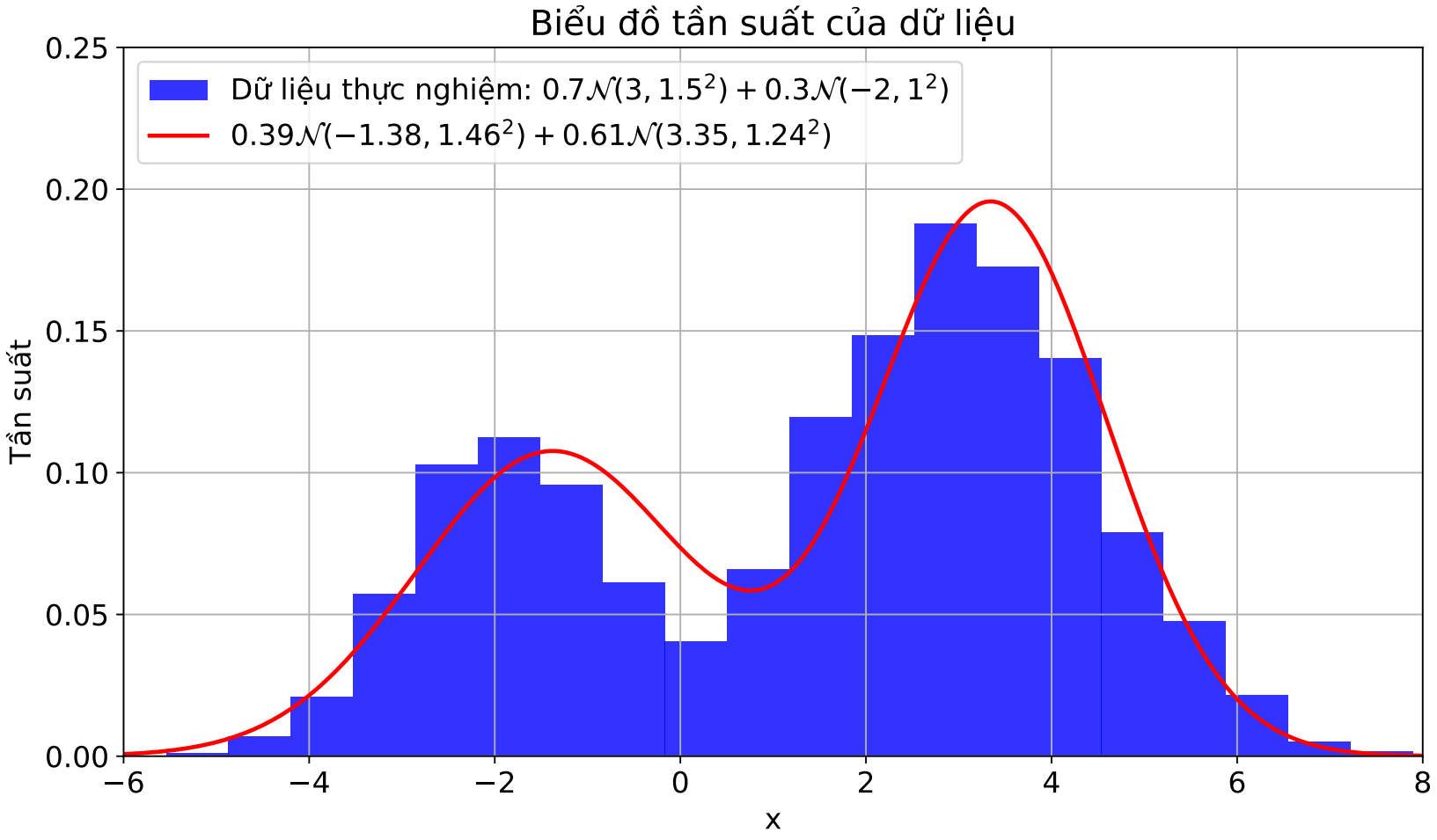
$$\pi^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,1}^{(t)}$$

(Tại sao?)

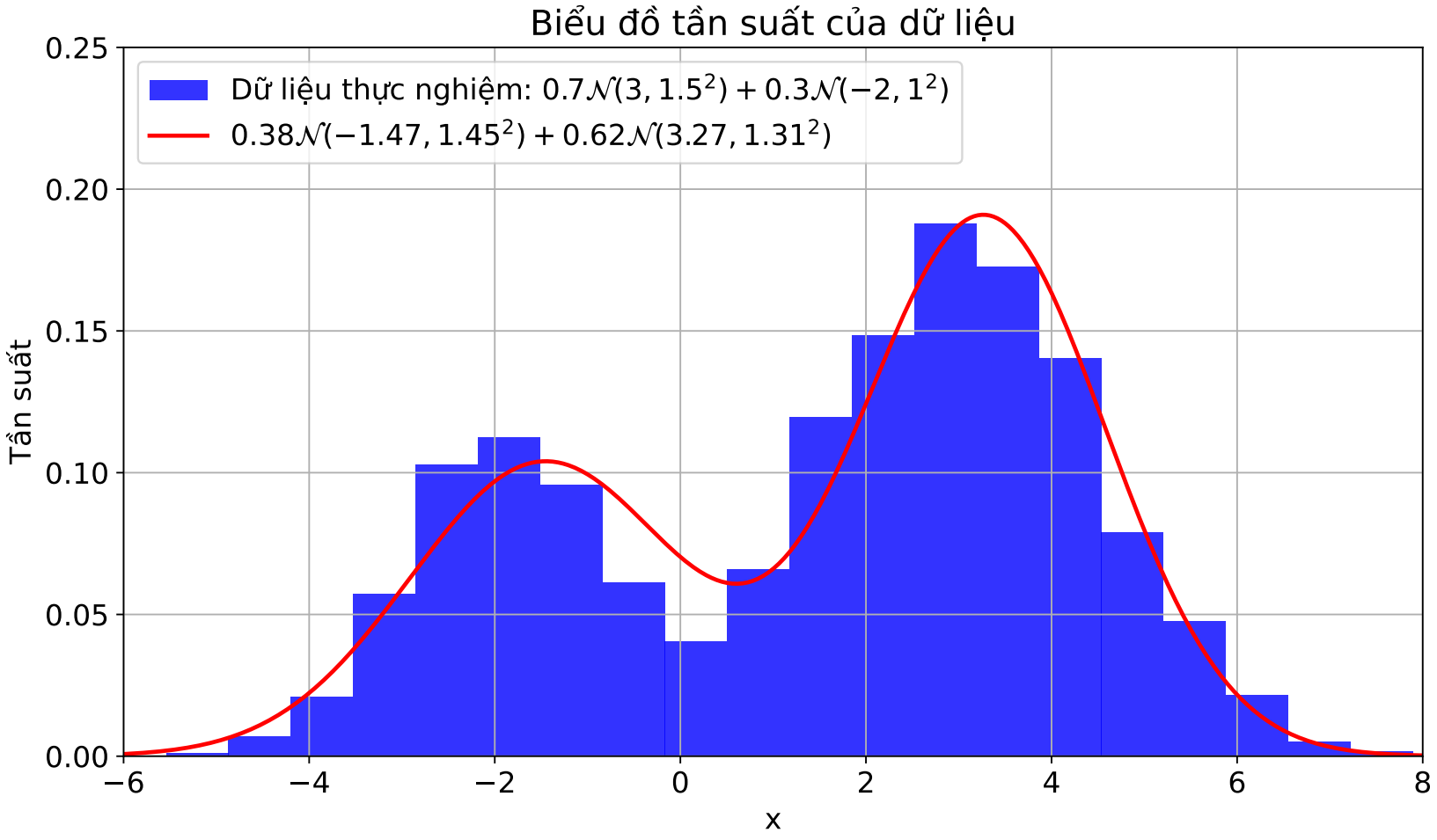
Ví dụ: mô hình hỗn hợp Gaussian



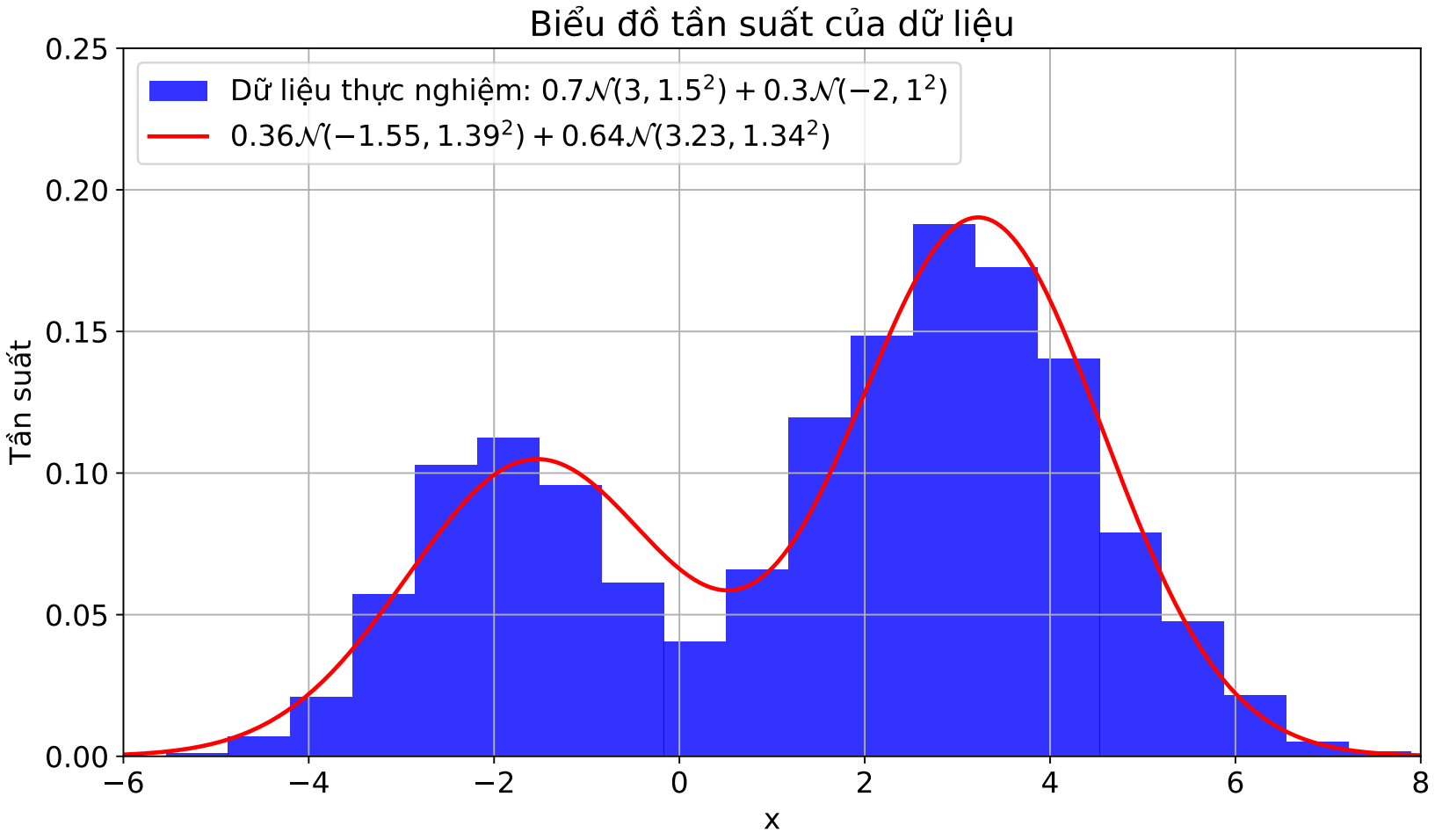
Ví dụ: mô hình hỗn hợp Gaussian



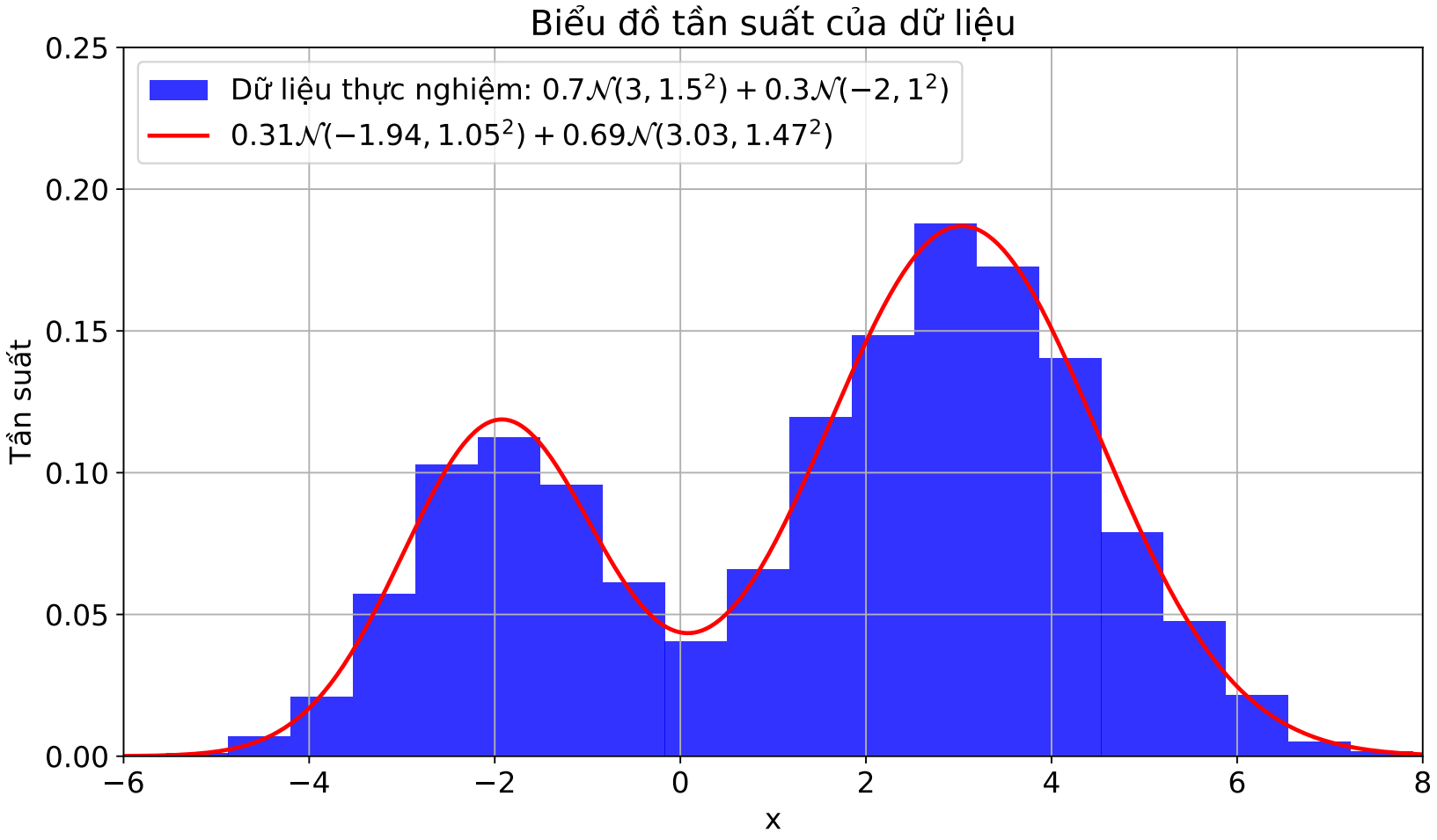
Ví dụ: mô hình hỗn hợp Gaussian



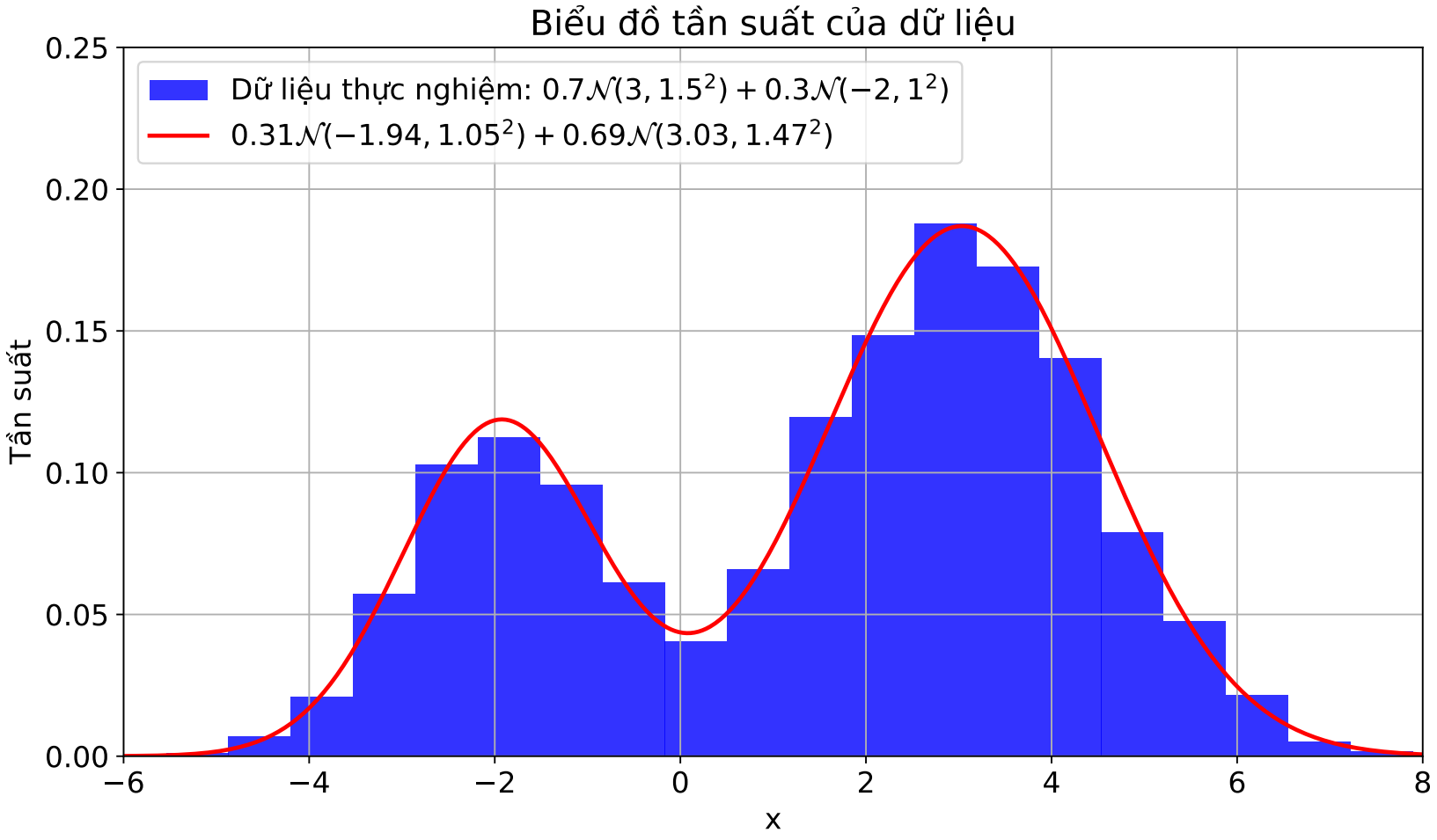
Ví dụ: mô hình hỗn hợp Gaussian



Ví dụ: mô hình hỗn hợp Gaussian



Ví dụ: mô hình hỗn hợp Gaussian



Câu hỏi thêm

Mở rộng về hướng các mô hình hỗn hợp nhiều thành phần

- Liệu có thể áp dụng với mô hình hỗn hợp nhiều thành phần Gaussian hơn?
- Liệu có thể áp dụng với mô hình hỗn hợp nhiều thành phần cùng họ phân phối?
- Liệu có thể áp dụng với mô hình hỗn hợp nhiều thành phần với họ phân phối khác nhau?
- Liệu có thể xác định số lượng thành phần từ dữ liệu?

Mở rộng về hướng mô hình biến ẩn

- Liệu có thể áp dụng với mô hình có biến ẩn liên tục?
- Liệu có thể áp dụng với mô hình phân nhiều cấp biến ẩn hơn và/hoặc cấu trúc phụ thuộc phức tạp hơn?

Các nội dung

Outlines

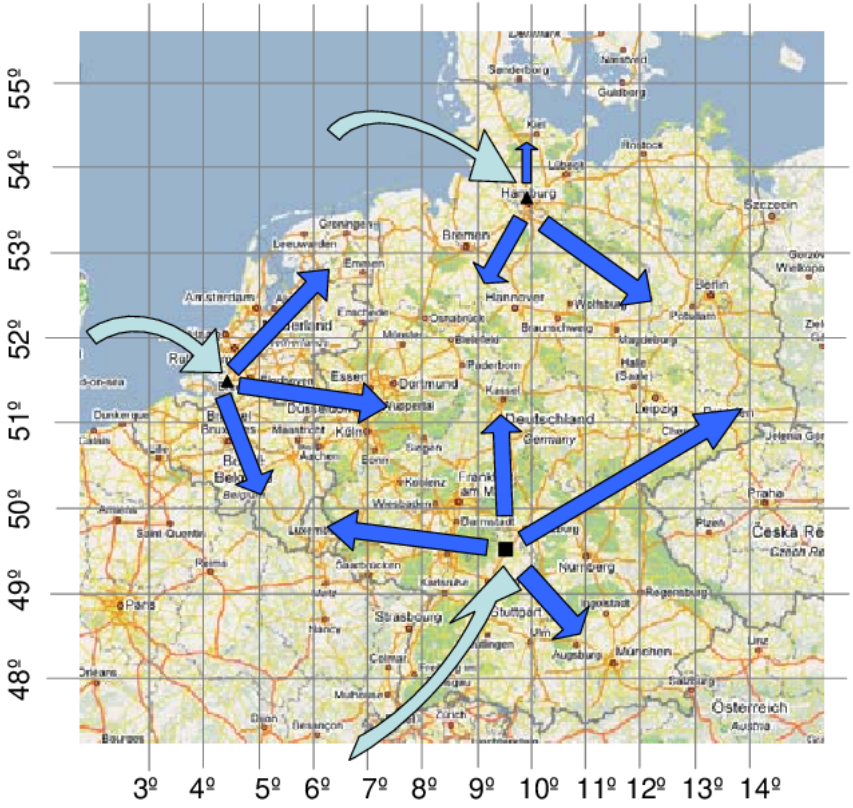
(Nhắc lại) Loại bài toán học

Phân cụm (clustering) là **bài toán học không giám sát (unsupervised learning)**. Nói cách khác, a là tập dữ liệu chưa dán nhãn (unlabelled data).

Ví dụ: Phần vùng dịch Covid từ vị trí của các ca bệnh

Một số ứng dụng

Ví dụ: Bài toán đặt nhà kho chứa hàng hóa (warehouse location problem)



Tổng quát: Bài toán đặt cơ sở bất kì (facility location problem)

Warehouse location problem

Bài toán không đơn giản do nhiều dữ liệu thông tin:

- rất nhiều đại lý bán lẻ,
- số lượng, vị trí đặt nhà kho,
- chi phí giao hàng đến các đại lý bán, v.v.

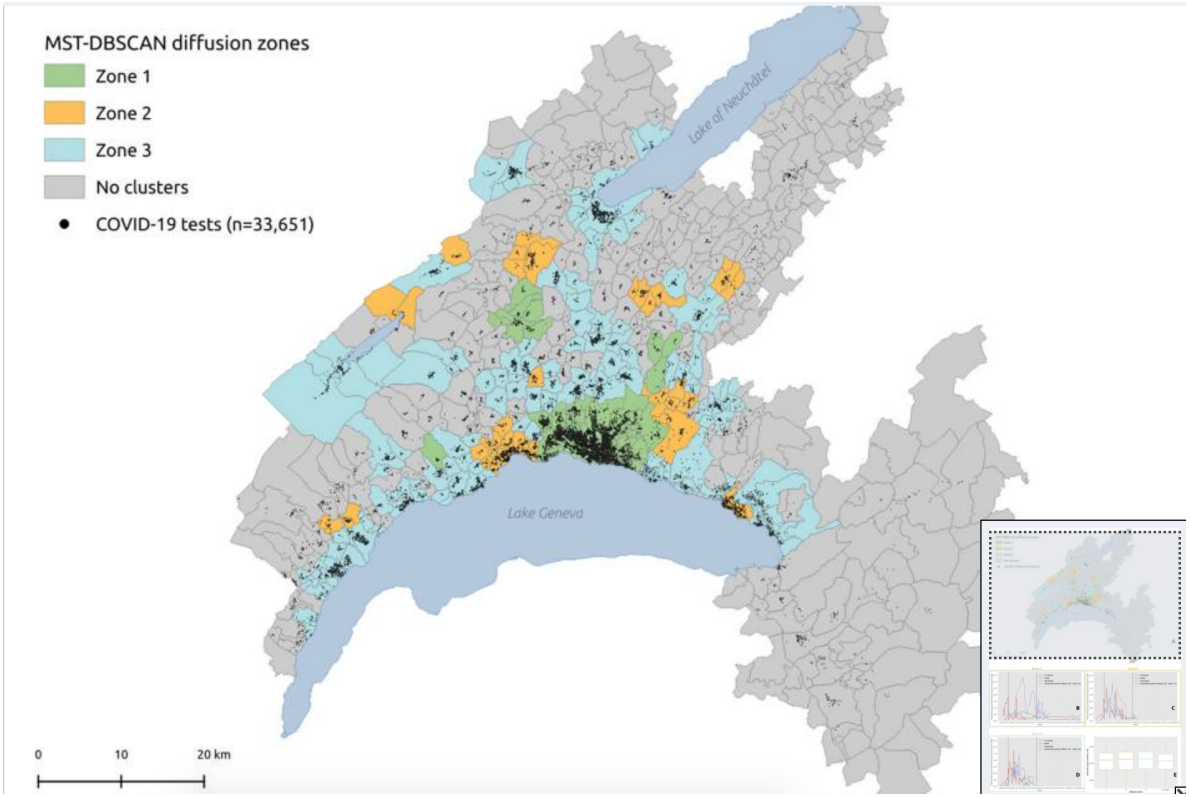
Một phương án là chia bài toán làm 2 bước:

1. phân cụm các đại lý một cách hợp lý, mỗi cụm sẽ được phân phối bởi một nhà kho (**clustering**)
2. trong từng cụm, sắp xếp lịch phân phối hàng một cách hợp lý nhất (bài toán travelling salesman, bài toán vehicle routing)

Link tham khảo

Ứng dụng của phân cụm

Ví dụ: Phân vùng dịch bệnh để phòng tránh và đối phó



Một phương án đề xuất xuất gồm 2 bước:

1. Phân cụm các điểm bệnh (**clustering**) để xác định các vùng chứa các cụm bệnh
2. Phân loại các vùng bệnh để xử lý cho phù hợp: Zone 1, Zone 2, Zone 3 (**classification**)

Link tham khảo

Bài toán phân cụm

Định nghĩa vắn tắt

Bài toán data clustering là bài toán phân nhóm cho tập dữ liệu đầu vào một cách hợp lý nhất.

Tập dữ liệu đầu vào: $\Theta = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$
Tập các cụm dữ liệu đầu ra: $C = \{C_1, C_2, \ldots, C_k\}$

Mục tiêu: Tìm hàm số $f : \Theta \to C$ hợp lý nhất

Giống như bài toán ML tổng quát hay ước lượng tham số, cần thêm các giả định về tính hợp lý để đưa về bài toán tối ưu hóa.

Outlines

Bài toán k-means

Mô hình k-means: Cho trước $\Theta = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$ và số tự nhiên k. Tìm một phân hoạch $S_1, S_2, \ldots, S_k$ của Θ sao cho hàm số

$$\mathcal{L} = \sum_j \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$$

đạt min với $\mu_j = \dfrac{\sum\limits_{x_i \in S_j} x_i}{|S_j|}$.

Nhận xét:

- Ý nghĩa của $\mathcal{L}$ là muốn tối thiểu phương sai trong từng cụm;
- Là bài toán tối ưu tổ hợp *khó*: Không gian hàm $\mathcal{H} = \{f : \Theta \to \{1, \ldots, k\}\}$ rời rạc, có độ lớn lũy thừa k^n;
- Hàm mất mát L chưa được tham số hóa.

Giải pháp: Kỹ thuật để tham số hóa các cấu trúc rời rạc

Biến chỉ thị (indicator variable):
Đặt $w_{ij} = 1$ nếu $x_i \in C_j$ (ký hiệu chuẩn là biến chỉ thị $\mathbb{I}(x_i \in C_j)$)

Xem L là hàm số $L(w_{ij}, \mu_j)$ và áp dụng thuật toán kỳ vọng-cực đại (expectation-maximization) gồm 2 bước:

1. (Expectation) $w_{ij} := 1$ nếu $j = \operatorname{argmin}_k \|x_i - \mu_k\|$
2. (Maximization) $\mu_j := (\sum_i w_{ij} x_i)/(\sum_i w_{ij})$

Nhận xét: Đây chính là thuật toán k-means cơ bản
Visualization

Outlines

Hard clustering

Output: Mỗi điểm dữ x_i liệu thuộc đúng một cụm duy nhất C_j hoặc không thuộc cụm nào, gọi là các điểm ngoài lề (outliers).

Ví dụ: k-means cơ bản, thuật toán dịch chuyển trung bình (mean-shift), thuật toán phân cụm dựa trên phổ của đồ thị (spectral clustering)

Soft clustering

Output: Mỗi điểm dữ liệu x_i thuộc cụm C_j với xác suất p_{ij}.

Ví dụ: k-means có trọng số (weighted), mô hình hỗn hợp Gaussian (Gaussian mixture model)

Câu hỏi trên lớp

Các giá trị $\{p_{ij}\}$ phải thỏa mãn điều kiện gì?

Các giả sử về dữ liệu và cụm hợp lý sẽ xác định không gian tìm kiếm $\mathcal{H}$. Một số giả định phổ biến:

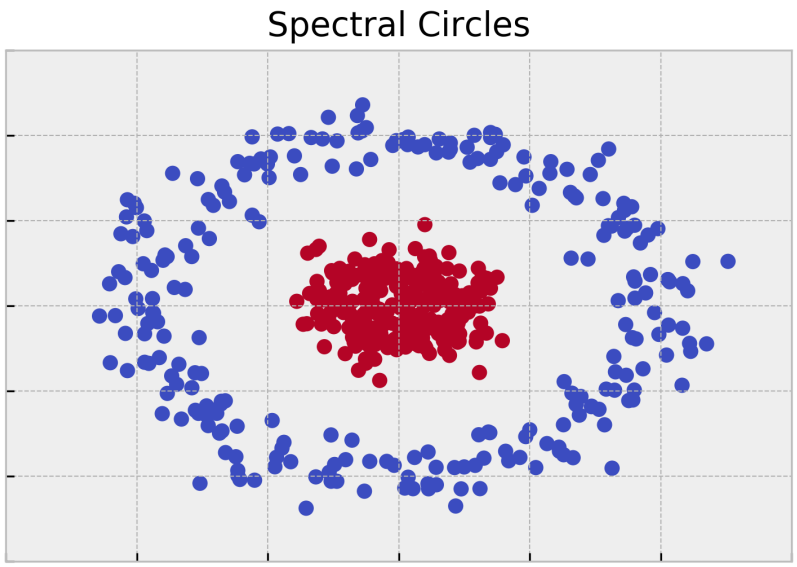
1. cụm dựa trên trọng tâm (centroid): k-means
2. cụm dựa trên tính liên thông (connectivity): spectral clustering
3. cụm dựa trên hàm nhân mật độ (kernel density): mean-shift
4. cụm dựa trên mật độ và phân phối xác suất (probability density distribution): mô hình hỗn hợp Gaussian

Cụm trọng tâm

Giả sử: Mỗi cụm được đại diện bởi một trọng tâm và các điểm dữ liệu được phân cụm dựa trên các trọng tâm này. Ví dụ: k-means

Cụm liên thông

Giả sử: Các điểm giống nhau sẽ ở chung cụm và sự chung cụm có tính bắc cầu (transitive). Ví dụ: spectral clustering (spectral ở đây là từ spectral graph theory)

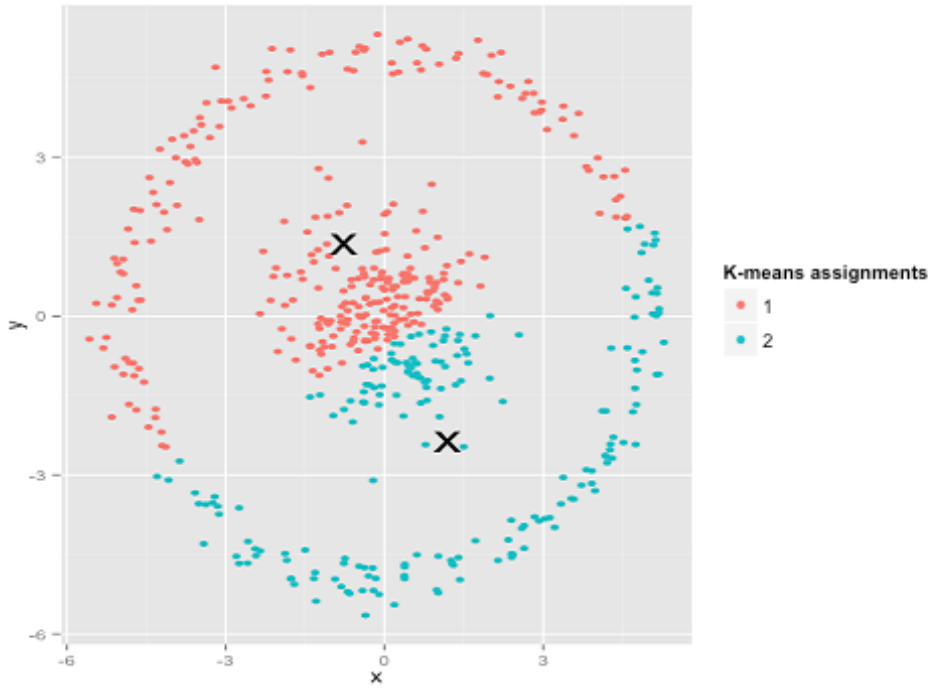


Visualization

Cụm liên thông

Câu hỏi trên lớp

Tại sao k-means với $k = 2$ không phân cụm được vòng tròn xanh bên ngoài trong dữ liệu ở dưới?



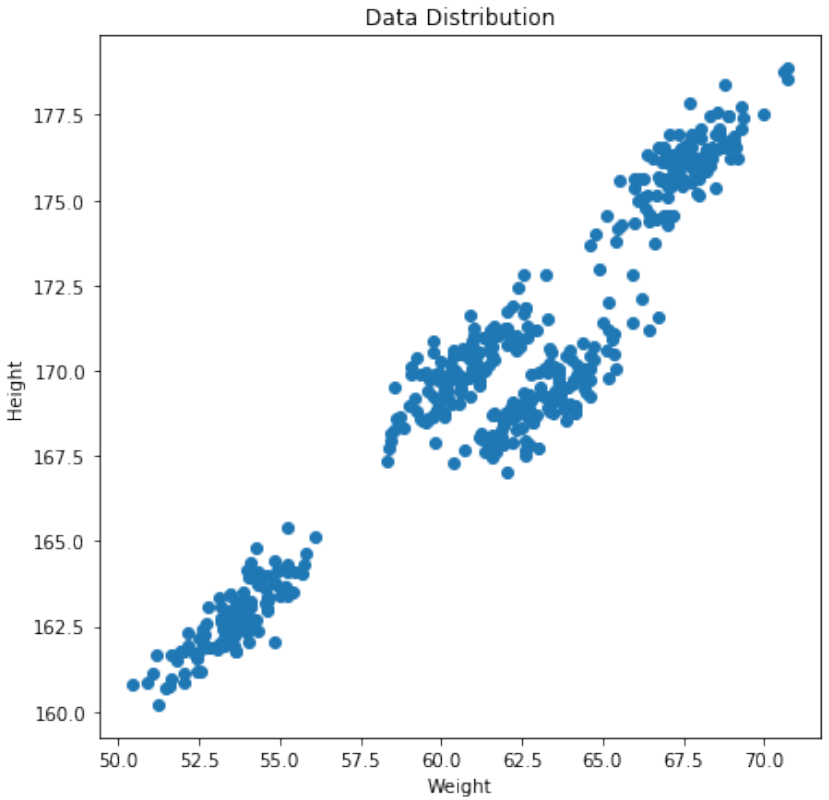
Cụm mật độ

Giả sử: Các cụm được tạo thành từ các vùng có mật độ điểm dữ liệu dày đặc. Ví dụ: mean-shift

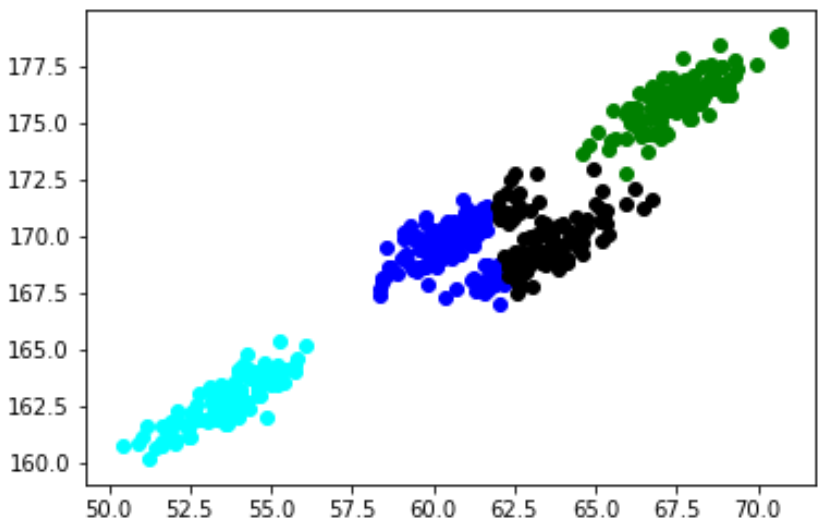
Visualization

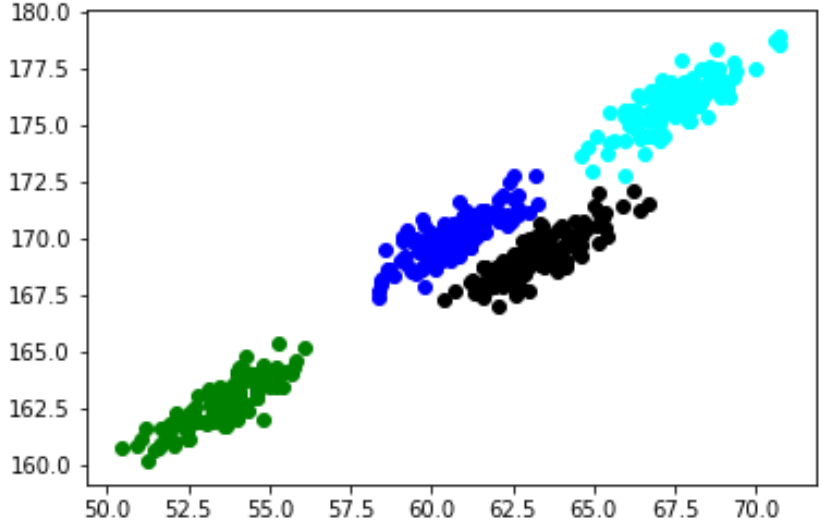
Cụm phân phối

Giả sử: Các điểm dữ liệu được lấy ngẫu nhiên (random sampling) theo một phối xác suất nào đó. Ví dụ: Gaussian mixture model



Cụm phân phối





Tóm tắt

- Phân cụm dữ liệu thuộc nhóm mô hình học không giám sát
- Các mô hình phân cụm khác nhau dựa trên những giả sử và đầu ra khác nhau
- Đưa về một bài toán tối ưu hóa và xử lý

Một số vấn đề chưa được cover

- Khởi tạo các tham số một cách hợp lý. Ví dụ: chọn k và vị trí k trọng tâm ban đầu cho k-means.
- Đánh giá độ hiệu quả của thuật toán cho một dữ liệu cụ thể. Ví dụ: homogeneity score, completeness score.

Q & A

Cảm ơn mọi người đã theo dõi!

Câu hỏi & Thảo luận.