

Gaussian Mixture Models

Tiến Dũng, Minh Tú, Đăng Minh, Đức Triều

PiMA 2021



Trình bày: Nhóm 4, GMM

August 8, 2021

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- Tối ưu hàm mục tiêu
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận

Contents

1 Giới thiệu

■ Định nghĩa

■ GMM trong bài toán phân cụm

2 Bài toán GMM

■ Mô hình toán học

■ Hàm mục tiêu

■ Tối ưu hàm mục tiêu

■ Thuật toán Expectation–maximization (EM)

■ EM trong GMM

3 Thực hành áp dụng

4 Kết luận



Bài toán phân cụm

Cho các điểm dữ liệu \rightarrow Chia thành các cụm khác nhau phù hợp nhất với tính chất của từng điểm dữ liệu.



Phân phối Gauss

- 1 Là phân phối xác suất thường dùng cho những tập dữ liệu đối xứng qua giá trị trung bình
- 2 Các đại lượng liên quan
 - Giá trị trung bình μ
 - Ma trận hiệp phương sai Σ
 - Hàm mật độ $p(x|\mu, \Sigma) = \frac{\exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))}{\sqrt{(2\pi)^d \det(\Sigma)}}$



Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- Tối ưu hàm mục tiêu
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận



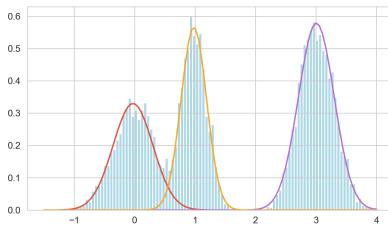
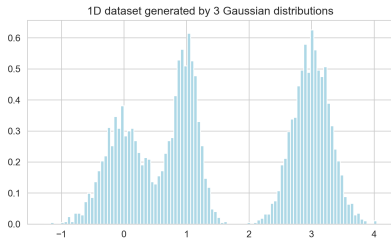
Ý tưởng

Một phân phối Gauss \rightarrow một tập dữ liệu đối xứng qua giá trị trung bình

Trường hợp tập dữ liệu có nhiều hơn một đỉnh \rightarrow nhiều phân phối Gauss \rightarrow GMM



Ví dụ



Ý tưởng

- **Input:** Số nguyên k các cụm và các điểm dữ liệu



Ý tưởng

- **Input:** Số nguyên k các cụm và các điểm dữ liệu
- **Output:** k phân phối Gauss hợp lí nhất và xác suất thuộc về các phân phối trên của mỗi điểm dữ liệu



Ý tưởng

- **Input:** Số nguyên k các cụm và các điểm dữ liệu
- **Output:** k phân phối Gauss hợp lí nhất và xác suất thuộc về các phân phối trên của mỗi điểm dữ liệu

→ *GMM trong bài toán clustering thuộc kiểu soft-clustering*

Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- Tối ưu hàm mục tiêu
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận



Thành phần của GMM

Tập hợp Θ chứa k tham số $\theta_i(\mu_i, \Sigma_i, \pi_i), i = \overline{1, k}$

Trong đó các π_i là độ ưu tiên (mixture coefficient) của phân phối thứ i



Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- **Hàm mục tiêu**
- Tối ưu hàm mục tiêu
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận



Hàm mục tiêu

Cần tối ưu hàm likelihood của mẫu:

$$L(\Theta|\mathbf{X}) = \prod_{i=1}^n \sum_{j=1}^k \pi_j P(x_i|\theta_j)$$

Trong thực hành ta thường dùng hàm log-likelihood

$$\ln L(\Theta|\mathbf{X}) = \sum_{i=1}^n \ln \left(\sum_{j=1}^k \pi_j P(x_i|\theta_j) \right)$$



Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- **Tối ưu hàm mục tiêu**
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận



Tối ưu hàm mục tiêu

- Tối ưu trực tiếp bằng Gradient ascent
- Tối ưu thông qua cập nhật bằng EM và k-means.



Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- Tối ưu hàm mục tiêu
- Thuật toán Expectation-maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận



Lịch sử

- Lần đầu được đề cập đến trong "Maximum Likelihood from Incomplete Data via the EM Algorithm" Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977).
- Một công cụ mới trong phân tích thống kê.



Biến ẩn (Latent variable)

- Biến không được cho biết, là thông tin đi kèm với dữ liệu quan sát được
- Ví dụ
 - Nhãn bị mất của lọ hóa học.
 - Cụm đúng của một điểm dữ liệu trong bài toán phân cụm.



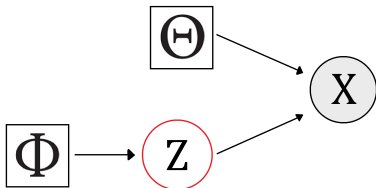
Tổng quan

- Thuật toán EM là thuật toán ước lượng tham số của mô hình thống kê.
- Iterative algorithm
- Được dùng khi xuất hiện biến ẩn trong hàm độ hợp lí (Likelihood function) .
- Khó khăn trong việc lấy đạo hàm của hàm hợp lí và trực tiếp giải bằng không.



Cách dữ liệu được tạo ra trong EM

- Gồm 2 bước:
 - 1 Biến ẩn Z được lấy ngẫu nhiên từ phân phối tham số bởi Φ
 - 2 Với một giá trị Z tương ứng ta sẽ lấy ngẫu nhiên được X từ phân phối tham số bởi Θ
- Ta chỉ quan sát được X và không có thông tin về Z . Mục tiêu là tìm lại được Φ và Θ .



Cách dữ liệu được tạo ra trong EM



Các dữ liệu được tạo ra trong EM

- Ví dụ cụ thể, cách dữ liệu được tạo ra trong GMM
 - 1 Một điểm dữ liệu chọn một trong K phân phối.
 - 2 Điểm dữ liệu được lấy ngẫu nhiên theo phân phối đó.
- Ta chỉ quan sát được tập dữ liệu gồm các điểm dữ liệu mà không có thông tin về các cụm.



Ước lượng hợp lí cực đại (MLE)

- Ta sẽ thử tiếp cận bài toán ước lượng tham số mô hình thống kê bằng ước lượng hợp lí cực đại.
- Ta có hàm Log-Likelihood:

$$\begin{aligned}l(\Phi, \Theta; X) &= \sum_x \log(p(x; \Phi, \Theta)) \\ &= \sum_x \log \sum_z p(\mathbf{z}; \Phi) p(x|\mathbf{z}; \Theta)\end{aligned}$$

- Chi tiết về ước lượng hợp lí cực đại trên GMM được trình bày cụ thể trong bài báo cáo.



Biến đổi hàm Log-Likelihood

- Để thuận tiện trong việc biến đổi, ta gọi Φ và Θ là θ .
- Ta biến đổi hàm Log-Likelihood:

$$\begin{aligned}
 l(\theta; x) &= \sum_z \log p(x; \theta) \\
 &= \log p(x; \theta) \sum_z p(z|x; \theta) \\
 &= \sum_z p(z|x; \theta) \log p(x; \theta) \\
 &= \sum_z p(z|x; \theta) \log \frac{p(z|x; \theta)p(x; \theta)}{p(z|x; \theta)} \\
 &= \sum_z p(z|x; \theta) \log \frac{p(x, z; \theta)}{p(z|x; \theta)} = \mathbb{E}_{z \sim p(z|x; \theta)} \left[\log \frac{p(x, z; \theta)}{p(z|x; \theta)} \right]
 \end{aligned}$$

Hàm mục tiêu mới

- Log-Likelihood: $l(\theta; X) = \mathbb{E}_{z \sim p(z|x; \theta)} \left[\log \frac{p(x, z; \theta)}{p(z|x; \theta)} \right]$

$$\begin{aligned} Q(\theta | \theta^{(t)}) &= \sum_z p(z|x; \theta^{(t)}) \log p(x, z; \theta) \\ &= \mathbb{E}_{z \sim p(z|x; \theta^{(t)})} [\log p(x, z; \theta)] \end{aligned}$$

- Ta có thể chứng minh:

$$l(\theta; X) \geq Q(\theta | \theta^{(t)})$$

- Giá trị θ tối đa hàm Q đồng thời sẽ tối đa hàm Log-Likelihood



Thuật toán EM

- Khởi tạo $\theta^{(0)}$ ngẫu nhiên
- Repeat until convergence
 - Bước E: Xây dựng lại hàm mục tiêu

$$Q(\theta|\theta^{(t)}) := \mathbb{E}_{z \sim p(z|x; \theta^{(t)})} [\log p(x, z; \theta)]$$

Bằng cách $p(z|x; \theta^{(t)}) := p(z|x; \theta^{(t-1)})$

- Bước M: Tối đa hàm mục tiêu

$$\theta^{(t+1)} := \operatorname{argmax}_{\theta} Q(\theta|\theta^{(t)})$$

- Convergence criteria: $|\theta^{(t+1)} - \theta^{(t)}| \leq \epsilon$



Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- Tối ưu hàm mục tiêu
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

3 Thực hành áp dụng

4 Kết luận



Trình tự

Khởi đầu với k phân phối Gauss có tham số $\theta_1, \theta_2, \dots, \theta_k$ bất kì, có thể có độ ưu tiên π_1, \dots, π_k .



Trình tự

Bước E

- Tính xác suất xuất hiện mỗi x_i khi có phân phối θ_j :

$$P(x_i|\theta_j) = \frac{\exp(-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu))}{\sqrt{(2\pi)^n \det(\Sigma_j)}}$$



Trình tự

Bước E

- Tính xác suất xuất hiện mỗi x_i khi có phân phối θ_j :

$$P(x_i|\theta_j) = \frac{\exp(-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu))}{\sqrt{(2\pi)^n \det(\Sigma_j)}}$$

- Tính xác suất để phần tử x_i thuộc phân phối θ_j :

$$P(\theta_j|x_i) = \frac{P(x_i|\theta_j)\pi_j}{\sum_{j=1}^k P(x_i|\theta_j)\pi_j}$$



Trình tự

Bước M

Cập nhật lại các θ_j

Đặt $N_j = \sum_{i=1}^n P(\theta_j | x_i)$. Khi đó:

$$\blacksquare \mu_j = \frac{1}{N_j} \sum_{i=1}^n P(\theta_j | x_i) x_i$$



Trình tự

Bước M

Cập nhật lại các θ_j

Đặt $N_j = \sum_{i=1}^n P(\theta_j | x_i)$. Khi đó:

- $\mu_j = \frac{1}{N_j} \sum_{i=1}^n P(\theta_j | x_i) x_i$
- $\Sigma_j = \frac{1}{N_j} \sum_{i=1}^n P(\theta_j | x_i) (x_i - \mu_j)(x_i - \mu_j)^T$



Trình tự

Bước M

Cập nhật lại các θ_j

Đặt $N_j = \sum_{i=1}^n P(\theta_j | x_i)$. Khi đó:

- $\mu_j = \frac{1}{N_j} \sum_{i=1}^n P(\theta_j | x_i) x_i$
- $\Sigma_j = \frac{1}{N_j} \sum_{i=1}^n P(\theta_j | x_i) (x_i - \mu_j)(x_i - \mu_j)^T$
- $\pi_j = \frac{1}{n} N_j$



Trình tự

Lặp lại hai bước trên

Thuật toán dừng khi sai số của hàm likelihood ở 2 bước lặp liên tiếp nhỏ hơn một số ϵ cho trước.



Tính hội tụ

Tại sao lại làm những bước trên?

Việc cập nhật lại các tham số thực chất là đang tối ưu hàm ELBO (các tham số tìm được tại một bước lặp là nghiệm của đạo hàm hàm ELBO ở thời điểm đó)

→ Tối ưu được hàm likelihood.



Contents

1 Giới thiệu

- Định nghĩa
- GMM trong bài toán phân cụm

2 Bài toán GMM

- Mô hình toán học
- Hàm mục tiêu
- Tối ưu hàm mục tiêu
- Thuật toán Expectation–maximization (EM)
- EM trong GMM

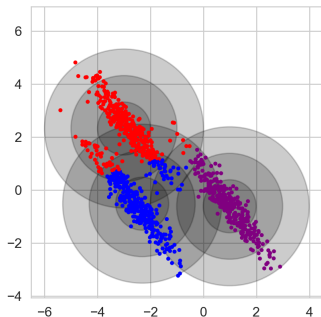
3 Thực hành áp dụng

4 Kết luận

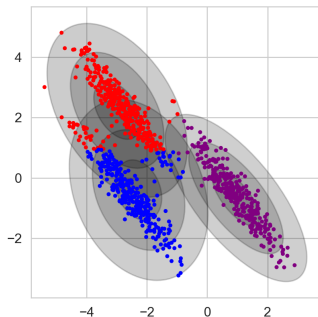


Kết quả thực hành

Visualize các Gaussian distribution thuật toán tìm được sau một số bước EM



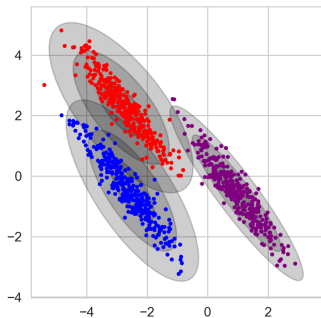
Step 0



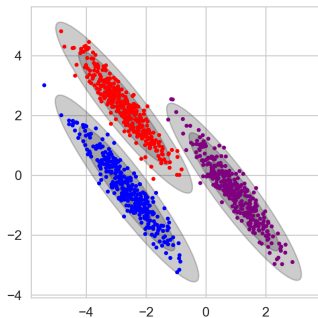
Step 1



Kết quả thực hành



Step 3



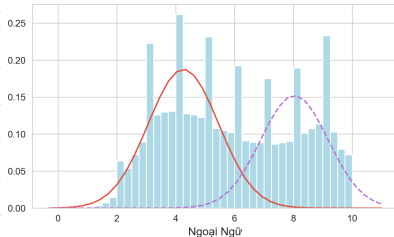
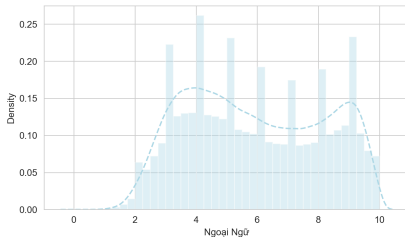
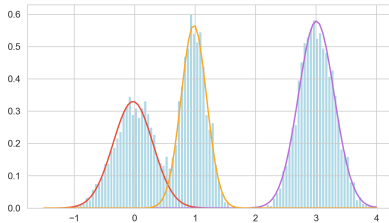
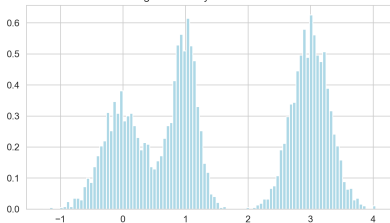
Step 11



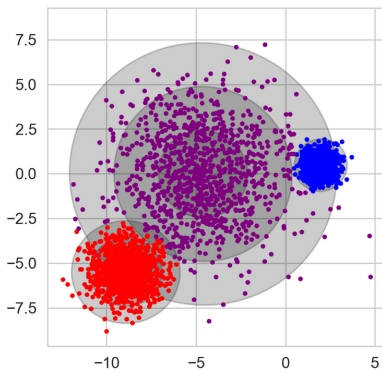
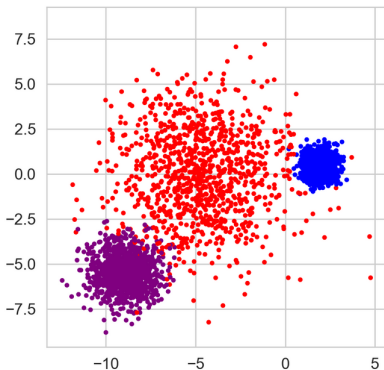
Kết quả thực hành

Một số ví dụ khác sử dụng GMM

1D dataset generated by 3 Gaussian distributions

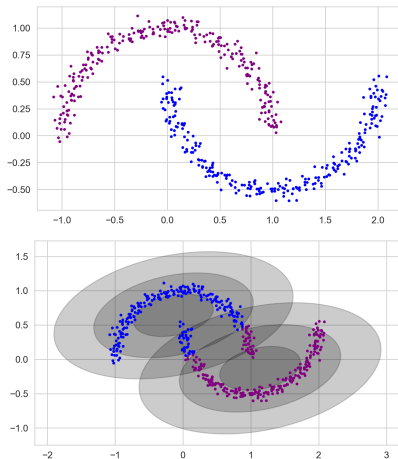


Kết quả thực hành

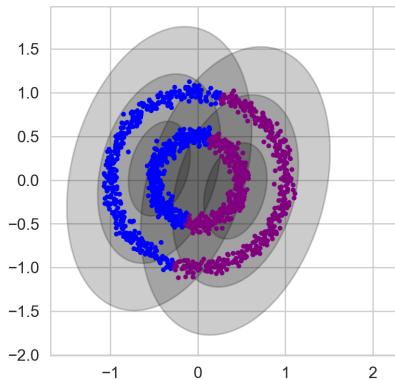
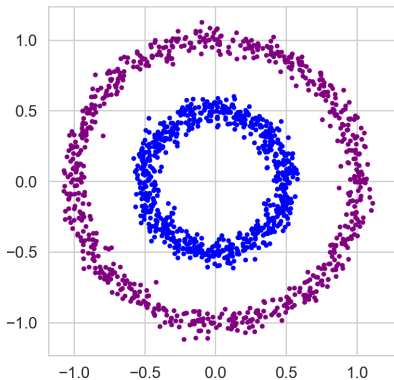


Kết quả thực hành

Một số ví dụ GMM phân cụm không chính xác

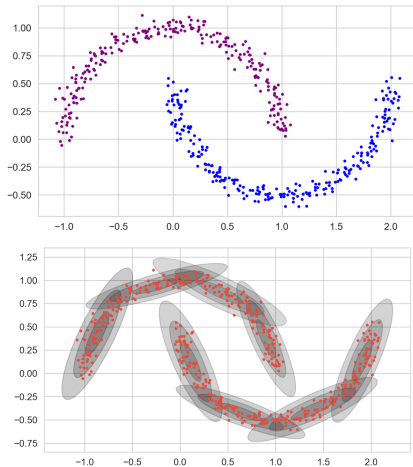


Kết quả thực hành



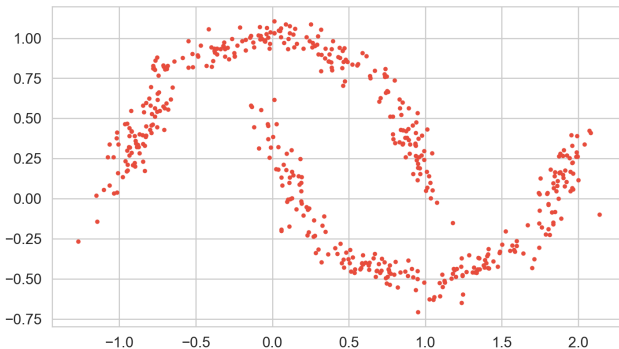
Kết quả thực hành

Sử dụng GMM như một generative model



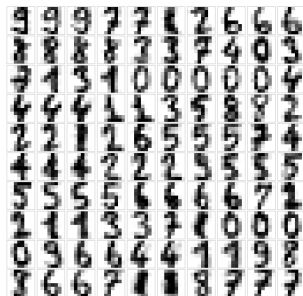
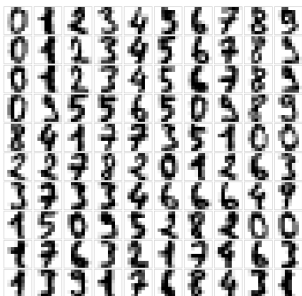
Kết quả thực hành

Dùng model trên sinh ra các data mới



Kết quả thực hành

Ví dụ trên handwritten digits. Bên trái là một phần nhỏ của datasets digits lấy từ sklearn, bên phải là việc sử dụng GMM để tạo ra các digits mới.



- Gaussian Mixture Model là mô hình xác suất miêu tả các quần thể con của một tập dữ liệu không gắn nhãn.
- GMM được giải quyết bằng thuật toán EM.
- Thuật toán EM gồm 2 bước: Expectation, Maximization.
- Trong ứng dụng, GMM được dùng để tìm lại phân phối của các quần thể con và có thể tạo thêm nhiều điểm dữ liệu mới từ phân phối đó.



References

- 1 Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- 2 Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1), 95–103.
- 3 gregorygundersen.com/blog/2019/11/10/em

