4.5

Trần Bửu Ân

Ngày 25 tháng 12 năm 2021

Phân cụm liên quan đến việc nhóm các vectơ đặc trưng không được gắn nhãn thành các cụm, sao cho các mẫu trong một cụm giống với nhau hơn các mẫu thuộc các cụm khác nhau. Thông thường, người ta cho rằng số lượng cụm được biết trước, nhưng nếu không thì không có thông tin trước nào được đưa ra về dữ liệu. Các ứng dụng của phân cụm có thể được tìm thấy trong các lĩnh vực truyền thông, nén và lưu trữ dữ liệu, tìm kiếm cơ sở dữ liệu, đối sánh mẫu và nhận dạng đối tượng.
Một cách tiếp cận phổ biến để phân tích phân cụm là giả định rằng dữ liệu đến từ một tổ hợp các phân phối (thường là Gaussian), và do đó mục tiêu là ước tính các tham số của mô hình hỗn hợp bằng cách tối đa hóa hàm khả năng cho dữ liệu. Tối ưu hóa trực tiếp hàm khả năng trong trường hợp này không phải là một nhiệm vụ đơn giản, do những ràng buộc cần thiết đốivới các tham số (sẽ nói thêm về điều này sau) và tính chất phức tạp của hàm khả năng, nói chung có rất nhiều địa phương cực đại và điểm yên ngựa.
Một phương pháp phổ biến để ước lượng các tham số của mô hình hỗn hợp là thuật toán EM, đã được thảo luận trong một cài đặt tổng quát hơn trong Phần 4.3. Trong phần này, chúng tôi giải thích những điều cơ bản về mô hình hỗn hợp(trong active 128)và giải thích hoạt động của phương pháp EM trong bối cảnh này.
Ngoài ra, chúng ta chỉ ra cách các phương pháp tối ưu hóa trực tiếp có thể được sử dụng để tối đa hóa khả năng xảy ra.

4.5.1 Mô hình hỗn hợp.

Cho $\tau = \{X_1, X_2, ..., X_n\}$ là các vectơ ngẫu nhiên nhận các giá trị trong một số tập $\chi \subseteq R^d$, mỗi X_i được phân phối theo mật độ hỗn hợp

$$g(x|\theta) = w_1\phi_1(x) + ... + w_n\phi_n(x) \tag{4.31}$$

trong đó $\phi_1, ..., \phi_K$ là mật độ xác suất (rời rạc hoặc liên tục) trên χ và trọng số dương $w_1, ..., w_K$ tổng lên đến 1. Bản pdf hỗn hợp này có thể được diễn giải theo cách sau. Gọi Z là một biến ngẫu nhiên rời rạc nhận các giá trị $1, 2, ..., K$ với các xác suất $w_1, ..., w_K$ và cho X là vectơ ngẫu nhiên có pdf có điều kiện, cho trước $Z = z$, là ϕ_z .
Theo quy tắc tích số (C.17), joint pdf của Z và X được đưa ra bởi $\phi_{Z,X}(z, x) = \phi_Z(z)\phi_{Z|X}(x|z) = w_z\phi_z(x)$ và pdf cận biên của X được tìm thấy bằng cách tính tổng joint pdf với các giá trị của z,cho (4.31). Do đó, một vectơ ngẫu nhiên

$X \sim g$ có thể được mô phỏng theo hai bước:
1. Đầu tiên, vẽ Z theo các xác suất

$$P[Z = z] = w_z, z = 1, ..., K.$$

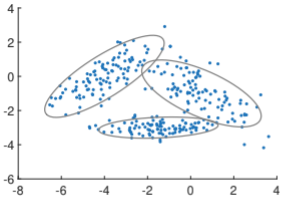
2. Sau đó vẽ X theo pdf ϕ_z
Vì τ chỉ chứa các biến $\{X_i\}$, nên $\{Z_i\}$ được xem như là các biến tiềm ẩn. Chúng ta có thể gọi Z_i là nhãn ẩn của cụm mà X_i thuộc về
Thông thường, mỗi ϕ_k trong (4.31) được giả định là đã biết đến một số vector tham số η_k. Thông thường 1 trong phân tích phân cụm là làm việc với các hỗn hợp Gaussian; nghĩa là, mỗi mật độ ϕ_k là Gaussian với một số vectơ kỳ vọng chưa biết μ_k và ma trận hiệp phương sai Σ_k. Chúng ta tập hợp tất cả các tham số chưa biết, bao gồm cả trọng số w_k, vào một vectơ tham số θ Như thường lệ, $\tau = x_1, ..., x_n$ biểu thị kết quả của τ . Vì các thành phần của τ là iid, joint pdf của chúng được đưa ra bởi

$$g(\tau|\theta) = \prod_{i=1}^{n} g(x_i|\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} w_k \phi_k(x_i|\mu_k, |\Sigma_k) \tag{4.32}$$

Theo lý luận tương tự như đối với (4.5), chúng ta có thể ước tính θ từ một kết quả τ bằng cách tối ưu hóa hàm log-khả năng

$$l(\theta|\tau) = \sum_{i=1}^{n} \ln g(x_i|\theta) = \sum_{i=1}^{n} \ln(\sum_{k=1}^{K} w_k \phi_k(x_i|\mu_k, \Sigma_k)) \tag{4.33}$$

Tuy nhiên, việc tìm giá trị cực đại của $L(\theta|\tau)$ nói chung là không dễ dàng, vì hàm thường là đa chiều
Ví dụ 4.4 (Phân cụm thông qua các mô hình hỗn hợp) Dữ liệu được mô tả trong Hình 4.4 bao gồm 300 điểm dữ liệu được tạo độc lập từ ba phân phối chuẩn hai biến , có các tham số được đưa ra trong cùng một hình. Đối với mỗi trong số ba giải thưởng này , chính xác 100 điểm đã được tạo ra. Tốt nhất, chúng tôi muốn phân cụm dữ liệu thành ba cụm tương ứng với ba trường hợp.
Để phân cụm dữ liệu thành ba nhóm, một mô hình khả thi cho dữ liệu là giả định rằng các điểm được lấy từ hỗn hợp (chưa biết) của ba phân biệt Gaussian 2 chiều. Đây là một cách tiếp cận hợp lý, mặc dù trên thực tế, dữ liệu không được mô phỏng theo cách này. Đó là hướng dẫn để hiểu sự khác biệt giữa hai mô hình.
Trong mô hình hỗn hợp, mỗi nhãn cụm Z nhận giá trị $1, 2, 3$ với xác suất bằng nhau và do đó, vẽ các nhãn một cách độc lập, tổng số điểm trong mỗi cụm sẽ được $B(300, 1/3)$ (phân phối nhị thức). Tuy nhiên, trong mô phỏng thực tế, số điểm trong mỗi cụm chính xác là 100. Tuy nhiên, mô hình hỗn hợp sẽ là một mô hình chính xác (mặc dù không chính xác) cho những dữ liệu này.
Hình 4.5 hiển thị mật độ hỗn hợp Gaussian "mục tiêu" cho dữ liệu trong Hình 4.4; nghĩa là hỗn hợp có khối lượng bằng nhau và với các thông số chính xác như quy định trong Hình 4.4



cluster	mean vector	covariance matrix
1	$\begin{bmatrix}-4\\0\end{bmatrix}$	$\begin{bmatrix}2 & 1.4\\1.4 & 1.5\end{bmatrix}$
2	$\begin{bmatrix}0.5\\-1\end{bmatrix}$	$\begin{bmatrix}2 & -0.95\\-0.95 & 1\end{bmatrix}$
3	$\begin{bmatrix}-1.5\\-3\end{bmatrix}$	$\begin{bmatrix}2 & 0.1\\0.1 & 0.1\end{bmatrix}$

Figure 4.4: Cluster the 300 data points (left) into three clusters, without making any assumptions about the probability distribution of the data. In fact, the data were generated from three bivariate normal distributions, whose parameters are listed on the right.

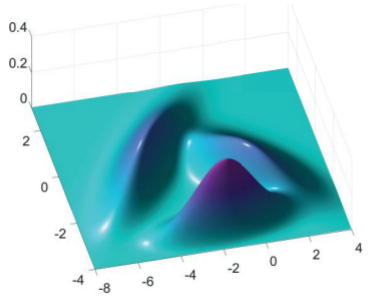


Figure 4.5: The target mixture density for the data in Figure 4.4.

Trong phần tiếp theo, chúng ta sẽ thực hiện phân cụm bằng thuật toán EM

4.5.2 Thuật Toán EM Cho Các Mô Hình Hỗn Hợp.

Như chúng ta đã thấy trong Phần 4.3, thay vì tối đa hóa hàm khả năng log trong (4.33) trực tiếp từ dữ liệu $\tau = \{x_1, ..., x_n\}$, thuật toán EM trước tiên tăng cường dữ liệu bằng vectơ của các biến tiềm ẩn - trong trường hợp này là các nhãn cụm ẩn $z = \{z_1, ..., z_n\}$.

Ý tưởng là τ là chỉ phần quan sát được của dữ liệu ngẫu nhiên hoàn chỉnh (τ, Z) , được tạo ra thông qua quy trình hai bước được mô tả ở trên. Nghĩa là, đối với mỗi điểm dữ liệu X, trước tiên hãy vẽ nhãn cụm $Z \in \{1, ..., K\}$ theo xác suất $\{w_1, ..., w_K\}$ và sau đó, cho $Z = z$, vẽ X từ ϕ_z. Hàm joint pdf của τ và Z là

$$g(\tau, z|\theta) = \prod_{i=1}^{n} w_{zi}\phi_{zi}(x_i)$$

có dạng đơn giản hơn nhiều so với (4.32).

Tiếp theo là complete-data log-likelihood function

$$\widetilde{l}(\theta|\tau, z) = \sum_{i=1}^{n} \ln(w_{zi}\phi_{zi}(x_i)) \tag{4.34}$$

thường dễ tối đa hóa hơn khả năng log ban đầu (4.33), đối với bất kỳ (τ, z) đã cho. Tuy nhiên, tất nhiên các biến tiềm ẩn z không được quan sát và $\widetilde{l}(\theta|\tau, z)$ không thể được đánh giá.

Trong bước E của thuật toán EM, the complete-data log-likelihood được thay thế bằng kỳ vọng $E_p\widetilde{l}(\theta|\tau, z)$ trong đó chỉ số con p trong kỳ vọng chỉ ra rằng Z được phân phối theo pdf có điều kiện của Z cho trước $T = \tau$; nghĩa là, với pdf

$$p(z) = g(z|\tau, \theta) \propto g(\tau, z|\theta) \tag{4.35}$$

Chú ý rằng $p(z)$ có dạng $p_1(z_1)....p_n(z_n)$ sao cho $T = \tau$ cho trước, các thành phần của Z là độc lập với nhau. Thuật toán EM cho các mô hình hỗn hợp hiện có thể được xây dựng như sau.

Algorithm 4.5.1: EM Algorithm for Mixture Models

input: Data τ, initial guess $\boldsymbol{\theta}^{(0)}$.
output: Approximation of the maximum likelihood estimate.

$t \leftarrow 1$
while a stopping criterion is not met **do**
  Expectation Step: Find $p^{(t)}(\mathbf{z}) := g(\mathbf{z}\,|\,\tau, \boldsymbol{\theta}^{(t-1)})$ and $Q^{(t)}(\boldsymbol{\theta}) := \mathbb{E}_{p^{(t)}}\widetilde{l}(\boldsymbol{\theta}\,|\,\tau, \mathbf{Z})$.
  Maximization Step: Let $\boldsymbol{\theta}^{(t)} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} Q^{(t)}(\boldsymbol{\theta})$.
  $t \leftarrow t + 1$
return $\boldsymbol{\theta}^{(t)}$

Điều kiện kết thúc có thể xảy ra là dừng khi

$$|l(\theta^{(t)}|\tau) - l(\theta^{(t-1)}|\tau)|/|l(\theta^t|\tau)| < \epsilon$$

đối với một số dung sai nhỏ $\epsilon > 0$

Như đã được đề cập trong Phần 4.3, chuỗi các giá trị khả năng xảy ra log không giảm theo mỗi lần lặp. Theo những điều kiện liên tục nhất định , chuỗi $\{\theta^{(t)}\}$ được đảm bảo hội tụ tới một cực đại cục bộ của khả năng l. Sự hội tụ thành công cụ tối đa hóa toàn cầu (nếu nó tồn tại) phụ thuộc vào lựa chọn thích hợp cho giá trị bắt đầu. Thông thường, thuật toán được chạy từ các điểm bắt đầu ngẫu nhiên khác nhau .

Đối với trường hợp hỗn hợp Gauss, mỗi $\phi_k = \phi(\cdot|\mu_k, \Sigma_k)$, k=1,..,K là mật độ của phân bố Gaussian d-chiều. Gọi $\theta^{(t-1)}$ là dự đoán hiện tại cho vectơ tham số tối ưu , bao gồm trọng số $\{w_k^{(t-1)}\}$, vectơ trung bình $\{\mu_k^{(t-1)}\}$ và ma trận hiệp phương sai $\{\Sigma_k^{(t-1)}\}$.

Đầu tiên chúng ta xác định $p^{(t)}$ - pdf của Z với điều kiện $T = \tau$ – đối với dự đoán $\theta^{(t-1)}$ đã cho. Như đã đề cập trước đây, các thành phần của Z cho trước $T = \tau$ là độc lập, vì vậy nó đủ để chỉ định pdf rời rạc, $p_i^{(t)}$ mà tôi nói, của mỗi Z_i cho điểm quan sát $X_i = x_i$. Cái sau có thể được tìm thấy từ công thức của Bayes:

$$p_i^t(k) \propto w_k^{(t-1)}\phi_k(x_i|\mu_k^{(t-1)}, \Sigma_k^{(t-1)}), k = 1, .., K \tag{4.36}$$

Tiếp theo, theo quan điểm của (4.34) , hàm $Q^{(t)}(\theta)$ có thể được viết dưới dạng

$$Q^{(t)}(\theta) = E_{p^{(t)}} \sum_{i=1}^{n} (\ln w_{Z_i} + (\ln \phi_{Z_i}(x_i|\mu_{Z_i}, \Sigma_{Z_i}))$$

$$= \sum_{i=1}^{n} (E_{p^{(t)}}(\ln w_{Z_i} + \ln(\phi_{Z_i}(x_i|\mu_{Z_i}, \Sigma_{Z_i})))$$

trong đó $\{Z_i\}$ là độc lập và Z_i được phân phối theo $p_i^{(t)}$ trong (4.36). Điều này làm giảm tốc độ của E-step.

Trong bước M, chúng ta cực đại $Q^{(t)}$ đối với tham số θ; nghĩa là đối với $\{w_k\}$, $\{\mu_k\}$ và $\{\Sigma_k\}$.Đặc biệt, chúng ta tối đa hóa

$$\sum_{i=1}^{n} \sum_{k=1}^{K} p_i^k(t)(\ln w_k + \ln \phi_k(x_i|\mu_k, \Sigma_k))$$

với điều kiện $\sum_k w_k = 1$. Sử dụng nhân Lagrange và thực tế là $\sum_{k=1}^{K} p_i^{(t)}(k) = 1$ đưa ra nghiệm cho $\{w_k\}$:

$$w_k = \frac{1}{n} \sum_{i=1}^{n} p_i^{(t)}(k), k = 1, 2, .., K \tag{4.37}$$

Các giải pháp cho μ_k và Σ_k bây giờ tuân theo từ việc tối đa hóa $\sum_{i=1}^{n} p_i^{(t)}(k) \ln \phi_k(x_i|\mu_k, \Sigma_k)$

$$\mu_k = \frac{\sum_{i=1}^{n} p_i^{(t)}(k) x_i}{\sum_{i=1}^{n} p_i^{(t)}(k)}, k = 1, .., K \tag{4.38}$$

and

$$\Sigma_k = \frac{\sum_{i=1}^{n} p_i^{(t)}(k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} p_i^{(t)}(k)}, k = 1, .., K \tag{4.39}$$

những công thức rất giống với các công thức nổi tiếng cho MLE của các tham số của phân phối Gauss. Sau khi gán các tham số nghiệm cho $\theta^{(t)}$ và tăng bộ đếm lặp t lên 1, các bước (4.36), (4.37), (4.38) và (4.39) được lặp lại cho đến khi đạt được sự kết hợp. Sự hội tụ của thuật toán EM rất nhạy cảm với việc lựa chọn các tham số ban đầu. Do đó, chúng tôi đề nghị thử nhiều điều kiện bắt đầu khác nhau. Để thảo luận thêm về các khía cạnh lý thuyết và thực tiễn của thuật toán EM, chúng tôi đề cập đến [85]. .

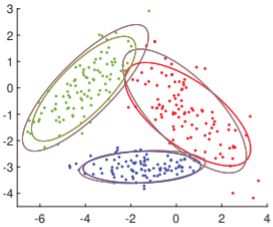
Ví dụ 4.5 (Phân cụm qua EM) Chúng tôi quay lại dữ liệu trong Ví dụ 4.4, được mô tả trong Hình 4.4 và áp dụng mô hình mà dữ liệu đến từ hỗn hợp của ba phân phối Gaussian hai biến.
Mã Python bên dưới thực hiện thủ tục EM được mô tả trong Thuật toán 4.5.1.

Các vectơ trung bình ban đầu $\{\mu_k\}$ của phân bố Gaussian hai biến được chọn (từ việc kiểm tra bằng mắt) để nằm gần đúng ở giữa mỗi cụm, trong trường hợp này là $[-2, -3]^T, [-4, 1]^T, [0, -1]^T$.
Các ma trận hiệp phương sai tương ứng ban đầu được chọn làm ma trận nhận dạng, điều này thích hợp với sự trải rộng dữ liệu quan sát được trong Hình 4.4. Cuối cùng, các trọng lượng ban đầu là 1/3, 1/3, 1/3. Để đơn giản, thuật toán dừng sau 100 lần lặp, trong trường hợp này là quá đủ để đảm bảo sự hội tụ.

Mã và dữ liệu có sẵn từ trang web của cuốn sách trong thư mục GitHub Chương 4

```
import numpy as np
from scipy.stats import multivariate_normal

Xmat = np.genfromtxt('clusterdata.csv', delimiter=',')
K = 3
n, D = Xmat.shape

W = np.array([[1/3,1/3,1/3]])
M  = np.array([[-2.0,-4,0],[-3,1,-1]], dtype=np.float32)
# Note that if above *all* entries were written as integers, M would
# be defined to be of integer type, which will give the wrong answer

C = np.zeros((3,2,2))

C[:,0,0] = 1
C[:,1,1] = 1

p = np.zeros((3,300))

for i in range(0,100):

#E-step
    for k in range(0,K):
        mvn = multivariate_normal( M[:,k].T, C[k,:,:] )
        p[k,:] = W[0,k]*mvn.pdf(Xmat)

# M-Step
    p = (p/sum(p,0))   #normalize
    W = np.mean(p,1).reshape(1,3)

    for k in range(0,K):
        M[:,k] = (Xmat.T @ p[k,:].T)/sum(p[k,:])
        xm = Xmat.T - M[:,k].reshape(2,1)
        C[k,:,:] = xm @ (xm*p[k,:]).T/sum(p[k,:])
```

Các thông số ước lượng của sự phân bố hỗn hợp được cho ở bên phải của Hình 4.6. Sau khi gắn nhãn lại cho các cụm, chúng ta có thể quan sát thấy sự trùng khớp chặt chẽ với các thông số trong Hình 4.4.Các hình elip ở phía bên trái của Hình 4.6 cho thấy sự phù hợp chặt chẽ giữa các hình elip xác suất 95% của các phân phối Gaussian ban đầu (màu xám) và các phân bố ước tính.
Một cách tự nhiên để phân cụm từng điểm x_i là gán nó vào cụm k mà xác suất có điều kiện $p_i(k)$ là cực đại (với các mối quan hệ được giải quyết tùy ý). Điều này cho phép nhóm các điểm thành các cụm màu đỏ, xanh lục và xanh lam trong hình.



weight	mean vector	covariance matrix
0.33	$\begin{bmatrix}-1.51\\-3.01\end{bmatrix}$	$\begin{bmatrix}1.75 & 0.03\\0.03 & 0.095\end{bmatrix}$
0.32	$\begin{bmatrix}-4.08\\-0.033\end{bmatrix}$	$\begin{bmatrix}1.37 & 0.92\\0.92 & 1.03\end{bmatrix}$
0.35	$\begin{bmatrix}0.36\\-0.88\end{bmatrix}$	$\begin{bmatrix}1.93 & -1.20\\-1.20 & 1.44\end{bmatrix}$

Figure 4.6: The results of the EM clustering algorithm applied to the data depicted in Figure 4.4.

Để thay thế cho thuật toán EM, tất nhiên người ta có thể sử dụng các thuật toán tối ưu hóa đa văn bản liên tục để tối ưu hóa trực tiếp hàm khả năng đăng nhập $l(\theta|\tau) = \ln g(\tau|\theta)$ trong (4.33) trên tập Θ tất cả những gì có thể θ. Điều này được thực hiện chẳng hạn trong [15], chứng tỏ kết quả vượt trội so với EM khi có ít điểm dữ liệu .
Điều tra kỹ hơn về hàm khả năng cho thấy rằng có một vấn đề tiềm ẩn với bất kỳ cách tiếp cận khả năng tối đa nào để phân cụm nếu Θ được chọn càng lớn càng tốt – tức là, bất kỳ phân phối hỗn hợp nào cũng có thể xảy ra.

Để chứng minh vấn đề này, hãy xem xét Hình 4.7, mô tả hàm mật độ xác suất, $g(\cdot|\theta)$ của hỗn hợp hai phân bố Gaussian, trong đó $\theta = [w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2]^T$ là véc tơ tham số của phân bố hỗn hợp.
Hàm log-khả năng xảy ra được cho bởi

$$l(\theta|\tau) = \sum_{i=1}^{4} \ln g(x_i|\theta)$$

trong đó $x_1, ..., x_4$ là dữ liệu (được biểu thị bằng các dấu chấm trong hình)

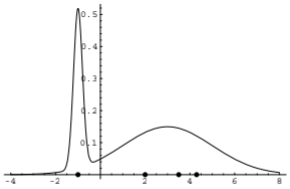


Figure 4.7: Mixture of two Gaussian distributions.

Rõ ràng là bằng cách cố định hằng số trộn w ở $0,25$ (giả sử) và căn giữa cụm đầu tiên ở x_1, người ta có thể thu được giá trị khả năng lớn tùy ý bằng cách lấy phương sai của cụm đầu tiên nhỏ tùy ý.
Tương tự, đối với dữ liệu có chiều cao hơn, bằng cách chọn các cụm "điểm" hoặc "dòng" hoặc các cụm "suy giảm" nói chung, người ta có thể làm cho giá trị của khả năng là vô hạn.

Đây là biểu hiện của vấn đề trang bị quá mức quen thuộc đối với mất đào tạo mà chúng ta đã gặp trong Chương 2.
Do đó, cực đại không bị giới hạn của hàm khả năng log-khả năng là một bài toán khó, bất kể việc lựa chọn thuật toán tối ưu hóa là gì!

Hai giải pháp khả thi cho vấn đề "overfitting" này là:
1. Hạn chế tập tham số Θ theo cách không cho phép các cụm suy biến (đôi khi được gọi là cụm giả).
2. Chạy thuật toán đã cho và nếu giải pháp bị suy biến, hãy loại bỏ nó và chạy thuật toán mới. Tiếp tục khởi động lại thuật toán cho đến khi thu được giải pháp không suy biến .
Cách tiếp cận đầu tiên thường được áp dụng cho các thuật toán tối ưu hóa đa chiều và cách tiếp cận thứ hai được sử dụng cho thuật toán EM.