# Thống kê tính toán - Computational Statistics

## Bài tập 02: Thuật toán EM

**Bài tập 1.** Phân phối $t$-student là một trong những phân phối được sử dụng rộng rãi trong phân tích dữ liệu thống kê. Gọi $Y$ là biến tuân theo phân phối $t$-student với bậc tự do $\nu$, khi đó, hàm mật độ xác suất của $Y$ là

$$f(y;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{y^2}{\nu}\right)^{-(\nu+1)/2},$$

trong đó, hàm $\Gamma(u)$ được xác định là

$$\Gamma(u) = \int_0^{+\infty} s^{u-1}\exp(-s)\,\mathrm{d}s.$$

Bên cạnh đó, ta cũng có biến thể là phân phối location-scale $t$-student, tức là

$$f(y;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(y-\mu)^2}{\sigma\nu}\right)^{-(\nu+1)/2},$$

trong đó, $-\infty < \mu < \infty$ là tham số cho location (vị trí của vùng dữ liệu), $\sigma > 0$ là tham số cho scale (biên độ rộng của phân phối), và $\nu > 0$ là bậc tự do. Khi đó, biến $Y$ có thể được miêu tả bởi mô hình location-scale:

$$Y = \mu + \sqrt{\sigma}T,$$

với $T \sim t_\nu$, phân phối $t$-student với bậc tự do $\nu$. Mục tiêu là ước lượng $\mu, \sigma$ và $\nu$.

(a) Áp dụng thuật toán EM xây dựng quy trình ước lượng cho các tham số $\mu, \sigma$ và $\nu$. Tham khảo thêm tại https://zouyuxin.github.io/Note/EMtDistribution.pdf

(b) Áp dụng quy trình EM ở câu (a) cho dữ liệu mô phỏng từ mô hình location-scale, với $\mu = 0$, $\sigma = 1$ và $\nu = 5$. Thực hiện 1 quy trình mô phỏng Monte Carlo với 1000 lần lặp để đánh giá độ chính xác của ước lượng EM.

(c) Áp dụng quy trình EM ở câu (a) cho dữ liệu sự khác biệt về bức xạ mặt trời tối đa giữa hai vùng địa lý theo thời gian đã tạo ra dữ liệu (đã sắp xếp) sau:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| −26.8 | −3.6 | −3.4 | −1.2 | 0.4 | 1.3 | 2.3 | 2.7 | 3.0 | 3.2 | 3.2 | 3.6 | 3.6 |
| 3.9 | 4.2 | 4.4 | 6.0 | 6.6 | 6.7 | 7.1 | 8.1 | 10.6 | 10.7 | 24.0 | 32.8 | |

**Bài tập 2.** Xét dữ liệu `geyser` được cung cấp bởi thư viện `MASS`, về thời gian chờ (tính bằng phút) của $n = 299$ lần phun trào liên tiếp của mạch nước nóng phun trào Old Faithful tại Công viên quốc gia Yellowstone, Hoa Kỳ.

(a) Sử dụng phương pháp Monte Carlo EM để xây dựng quy trình xác định các tham số của hàm mật độ normal-mixture của dữ liệu.

(b) Thi hành quy trình ở câu (a). Chú ý, khảo sát sự ảnh hưởng của việc lựa chọn giá trị $m_t$.

(c) Thi hành phương pháp EM Gradient để xác định các tham số của hàm mật độ normal-mixture của dữ liệu.

(d) Thi hành phương pháp quasi-Newton EM để xác định các tham số của hàm mật độ normal-mixture của dữ liệu.

**Bài tập 3.** Xét ví dụ về di truyền nổi tiếng từ Rao (1973, trang 369) cho rằng dữ liệu kiểu hình

$$\boldsymbol{x} = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$$

được phân phối theo phân phối đa thức (multinomial distribution[1]) với xác suất

$$\boldsymbol{p} = (p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right),$$

với $\theta \in (0, 1)$. Hàm log-likelihood dựa trên $\boldsymbol{x}$ và phân phối đa thức là

$$\ell(\theta; \boldsymbol{x}) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta).$$

Giả sử rằng $\boldsymbol{x}$ là thành phần quan sát được của một vector $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)$ được phân phối theo phân phối đa thức (multinomial distribution) với xác suất

$$\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5) = \left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right),$$

Khi đó, $x_1 = y_1 + y_2$, $x_2 = y_3$, $x_3 = y_4$, $x_4 = y_5$, tức là $y_1$ hoặc $y_2$ là bị khuyết.

(a) Áp dụng thuật toán EM để ước lượng $\theta$ dựa trên $\boldsymbol{y}$.

(b) Thi hành thuật toán ở câu (a), để tìm ước lượng $\widehat{\theta}$.

(c) Thi hành phương pháp EM Gradient để tìm ước lượng $\widehat{\theta}$ dựa trên $\boldsymbol{y}$.

(d) Thi hành phương pháp quasi-Newton EM tìm ước lượng $\widehat{\theta}$ dựa trên $\boldsymbol{y}$.

**Bài tập 4.** Tệp dữ liệu `trivariatenormal.dat` cung cấp thông tin quan sát của $X_1, X_2$ và $X_3$, trong đó, có một số giá trị bị khuyết (ký hiệu là `NA`). Đặt $\mathbf{X} = (X_1, X_2, X_3)$. Giả sử rằng $\mathbf{X}$ tuân theo phân phối chuẩn 3 chiều[2], $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, với vec tơ trung bình $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ và ma trận hiệp phương sai

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^3 \end{pmatrix}$$

Ta mong muốn ước lượng $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$ dựa vào thông tin của dữ liệu.

[1] https://en.wikipedia.org/wiki/Multinomial_distribution
[2] https://en.wikipedia.org/wiki/Multivariate_normal_distribution

(a) Giả sử $\boldsymbol{\Sigma}$ đã biết

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 & 1.2 \\ 0.6 & 0.5 & 0.5 \\ 1.2 & 0.5 & 3 \end{pmatrix}$$

Áp dụng thuật toán EM để ước lượng $\boldsymbol{\mu}$. Chú ý rằng phân phối chuẩn nhiều chiều thuộc họ phân phối mũ nhiều chiều.

(b) Thi hành thuật toán ở câu (a) với điểm bắt đầu phù hợp. Điều tra hiệu suất của thuật toán.

(c) Áp dụng thuật toán EM để ước lượng $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$.

(d) Thi hành thuật toán ở câu (c) với điểm bắt đầu phù hợp. Điều tra hiệu suất của thuật toán. So sánh với quá trình câu (b).

**Bài tập 5.** Xét dữ liệu sau về tuổi thọ (lifetimes) của 14 khớp nối bánh răng (gear coupling)

| | | | | | | |
|---|---|---|---|---|---|---|
| (6.94) | 5.50 | 4.54 | 2.14 | (3.65) | (3.40) | (4.38) |
| 10.24 | 4.56 | 9.42 | (4.55) | (4.15) | 5.64 | (10.23) |

trong đó, các giá trị nằm trong dấu ngoặc ám chỉ giá trị tuổi thọ của khớp nối bị che khuất bên phải (right censoring) bởi vì thiết bị đã được thay thế trước khi khớp nối bánh răng bị hỏng.

Dữ liệu này thường được mô hình với phân phối Weibull có hàm mật độ xác suất

$$f(x) = abx^{b-1} \exp(-ax^b),$$

với $x > 0$, và hai tham số $a$ và $b$. (Xem thêm bài tập ..., Bài tập số 1 về chi tiết mô hình). Đặt $\boldsymbol{\theta} = (a, b)$.

(a) Xác định hàm $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ bằng phương pháp phù hợp.

(b) Theo dạng của hàm $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ xác định được ở câu (a), hãy đề xuất quy trình thuật toán EM phù hợp.

(c) Thi hành thuật toán ở câu (b) với điểm bắt đầu phù hợp. So sánh kết quả với điểm bắt đầu $(a^{(0)}, b^{(0)}) = (0.003, 2.5)$.