Thống kê tính toán - Computational Statistics

Bài tập 01: Tối ưu trong Thống kê

Bài tập 1. Dempster, Laird và Rubin (1977) [1] đã xem xét phân phối đa thức với bốn nhóm danh mục, tức là multinomial có hàm xác suất,

$$p(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}$$

trong đó $n = x_1 + x_2 + x_3 + x_4$ và $p_1 + p_2 + p_3 + p_4 = 1$. Họ cho rằng xác suất có liên quan đến một tham số duy nhất, θ:

$$\begin{aligned} p_1 &= \frac{1}{2} + \frac{1}{4}\theta \\ p_2 &= \frac{1}{4} - \frac{1}{4}\theta \\ p_3 &= \frac{1}{4} - \frac{1}{4}\theta \\ p_4 &= \frac{1}{4}\theta, \end{aligned}$$

với $0 \leq \theta \leq 1$. (Mô hình này quay trở lại ví dụ được Fisher thảo luận năm 1925 trong Statistical Methods for Research Workers.) Với một quan sát (x_1, x_2, x_3, x_4), hàm log-likelihood là:

$$\ell(\theta) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta) + c$$

và

$$\frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = \frac{x_1}{2 + \theta} - \frac{x_2 + x_3}{1 - \theta} + \frac{x_4}{\theta}.$$

Mục tiêu là ước lượng θ.

(a) Xác định MLE cho θ. (Chỉ cần giải một phương trình đa thức đơn giản.) Đánh giá ước lượng bằng cách sử dụng dữ liệu mà Dempster, Laird và Rubin đã sử dụng: $n = 197$ và $x = (125, 18, 20, 34)$.

(b) Mặc dù dễ dàng tìm thấy nghiệm tối ưu như trong phần trước của bài tập này, nhưng việc sử dụng phương pháp của Newton là rất hữu ích. Viết các bước giải thuật để xác định ước lượng theo phương pháp của Newton. Viết một chương trình thi hành giải thuật này, bắt đầu với $\theta^{(0)} = 0.5$.

(c) Viết các bước giải thuật để xác định ước lượng theo phương pháp Fisher scoring. Viết một chương trình thi hành giải thuật này, bắt đầu với $\theta^{(0)} = 0.5$.

(d) Viết các bước giải thuật để xác định ước lượng theo phương pháp quasi-Newton. Viết một chương trình thi hành giải thuật này, bắt đầu với $\theta^{(0)} = 0.5$.

(e) So sánh các phương pháp này như thế nào?

Bài tập 2. Giả sử thời gian sống sót t cho các cá thể trong quần thể có hàm mật độ f và hàm phân phối tích lũy F. Khi đó hàm sống sót (*survival function*) là $S(t) = 1 - F(t)$. Hàm nguy cơ (*hazard function*) là $h(t) = \dfrac{f(t)}{1 - F(t)}$, hàm này đo lường rủi ro tử vong tức thời tại thời điểm t khi sống sót đến thời điểm t. Một mô hình nguy cơ tỷ lệ đặt ra rằng hàm nguy cơ phụ thuộc vào cả thời gian và một vectơ gồm các hiệp biến, $\boldsymbol{x}$, thông qua mô hình

$$h(t; x) = \lambda(t) \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right),$$

trong đó $\boldsymbol{\beta}$ là vectơ tham số. Đặt $\Lambda(t) = \displaystyle\int_{-\infty}^{t} \lambda(u)\, \mathrm{d}u$, ta chứng minh được rằng

$$S(t) = \exp\left\{-\Lambda(t) \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right)\right\}$$

và $f(t) = \lambda(t) \exp\left\{\boldsymbol{x}^\top \boldsymbol{\beta} - \Lambda(t) \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right)\right\}$.

(a) Giả sử dữ liệu của chúng ta là thời gian sống sót đã bị che khuất (censored survival time) t_i đối với $i = 1, \ldots, n$. Vào cuối nghiên cứu, một bệnh nhân hoặc đã chết (thời gian sống sót đã biết) hoặc vẫn còn sống (thời gian bị che khuất; biết là sống sót ít nhất cho đến khi kết thúc nghiên cứu). Đặt w_i là 1 nếu t_i là thời gian không bị che khuất và 0 nếu t_i là thời gian bị che khuất. Chứng minh rằng log-likelihood có dạng

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(w_i \log(\mu_i) - \mu_i\right) + \sum_{i=1}^{n} w_i \log\left(\frac{\lambda(t_i)}{\Lambda(t_i)}\right),$$

trong đó, $\mu_i = \Lambda(t_i) \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$.

(b) Hãy xem xét một mô hình về thời gian thuyên giảm cho bệnh nhân mắc bệnh bạch cầu cấp tính trong một thử nghiệm lâm sàng. Bệnh nhân được điều trị bằng thuốc 6-mercaptopurine (6-MP) hoặc giả dược [2]. Một năm sau khi bắt đầu nghiên cứu, thời gian (đơn vị: tuần) của thời gian thuyên giảm cho mỗi bệnh nhân đã được ghi lại (xem file `mercaptopurine.csv`). Một số kết quả đã bị che khuất vì thời gian thuyên giảm kéo dài sau thời gian nghiên cứu. Mục tiêu là xác định liệu phương pháp điều trị có kéo dài thời gian thuyên giảm hay không. Giả sử chúng ta đặt $\Lambda(t) = t^\alpha$ với $\alpha > 0$, khi đó, ta xác định được $\lambda(t) = \alpha t^{\alpha-1}$ và $f(t) = \alpha t^{\alpha-1} \exp\left\{\boldsymbol{x}^\top \boldsymbol{\beta} - t^\alpha \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right)\right\}$ (hàm mật độ Weibull). Áp dụng tham số hóa biến phụ thuộc: $\boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \delta_i \beta_1$, trong đó δ_i là 1 nếu bệnh nhân thứ i nằm trong nhóm điều trị (treatment) và bằng 0 nếu bệnh nhân đó nằm trong nhóm đối chứng (nhóm control). Viết các bước giải thuật để xác định ước lượng MLE của α, β_0 và β_1 bằng phương pháp Newton và Fisher Scoring. Viết chương trình thi hành các giải thuật này.

(c) Viết các bước giải thuật để xác định ước lượng MLE của α, β_0 và β_1 bằng phương pháp quasi-Newton. Viết một chương trình thi hành giải thuật này.

(d) Ước lượng sai số chuẩn (standard errors) của ước lượng MLE. Có bất kỳ ước lượng MLE nào có tương quan mạnh hay không?

Bài tập 3. Có 46 vụ tràn dầu thô với khối lượng ít nhất là 1000 thùng từ tàu chở dầu trên vùng biển Hoa Kỳ trong giai đoạn 1974-1999. Tập tin `oilspills.dat` chứa dữ liệu sau:

- `year` – năm thứ i;
- `spills` – số vụ tràn dầu N_i trong năm thứ i;
- `importexport` – lượng dầu ước tính được vận chuyển qua vùng biển Hoa Kỳ như một phần của hoạt động xuất nhập khẩu của Hoa Kỳ trong năm thứ i, được điều chỉnh cho sự cố tràn ở vùng biển quốc tế hoặc nước ngoài, b_{i1};
- `domestic` – lượng dầu được vận chuyển qua vùng biển Hoa Kỳ trong các chuyến hàng trong nước trong năm thứ i, b_{i2}.

Dữ liệu được điều chỉnh từ [3]. Lượng dầu được vận chuyển được đo bằng tỷ thùng (Bbbl).

Khối lượng dầu được vận chuyển là thước đo mức độ tiếp xúc với rủi ro tràn dầu. Giả sử chúng ta sử dụng giả định quá trình Poisson được đưa ra bởi $N_i|b_{i1}, b_{i2} \sim \mathcal{P}(\lambda_i)$ trong đó $\lambda_i = \alpha_1 b_{i1} + \alpha_2 b_{i2}$. Các tham số của mô hình này là α_1 và α_2, biểu thị tỷ lệ xảy ra tràn dầu trên mỗi Bbbl dầu được vận chuyển trong quá trình xuất nhập khẩu và trong nước.

(a) Xác hàm log-likelihood cho các hệ số α_1 và α_2.

(b) Viết các bước giải thuật để xác định ước lượng MLE của α_1 và α_2 bằng phương pháp Newton. Viết một chương trình thi hành giải thuật này.

(c) Viết các bước giải thuật để xác định ước lượng MLE của α_1 và α_2 bằng phương pháp Fisher Scoring. Viết một chương trình thi hành giải thuật này. So sánh với kết quả câu (b).

(d) Ước lượng sai số chuẩn (standard errors) của ước lượng MLE.

(e) Áp dụng phương pháp quasi-Newton với hai cách chọn $\mathbf{M}^{(0)}$: (1) ma trận đơn vị âm và (2) ma trận thông tin Fisher, $-\mathcal{I}\left(\boldsymbol{\alpha}^{(0)}\right)$. So sánh tính ổn định, tốc độ của hai cách chọn này.

(f) Xây dựng một đồ thị biểu diễn vùng tối ưu của hàm log-likelihood và để so sánh các đường đi được thực hiện bởi các phương pháp được sử dụng trong (b), (c) và (e). Chọn vùng vẽ đồ thị và điểm bắt đầu để minh họa tốt nhất các tính năng về hiệu suất của thuật toán.

Bài tập 4. Tệp dữ liệu `beetles.csv` cung cấp số lượng quần thể bọ cánh cứng bột (*Tribolium confusum*) tại nhiều thời điểm khác nhau [4] sau 154 ngày. Bọ cánh cứng ở mọi giai đoạn

phát triển đều được đếm và nguồn cung cấp thức ăn được kiểm soát cẩn thận. Một mô hình cơ bản cho sự tăng trưởng quần thể là mô hình logistic được đưa ra bởi

$$\frac{\mathrm{d}N}{\mathrm{d}t} = rN\left(1 - \frac{N}{K}\right),$$

trong đó, N là kích thước quần thể, t là thời gian, r là một tham số biểu diễn tốc độ tăng trưởng, K là tham số biểu diễn sức chứa dân số của môi trường sống. Nghiệm của phương trình vi phân này là

$$N_t = g(t) = \frac{KN_0}{N_0 + (K - N_0)\exp(-rt)},$$

với N_t là kích thước quần thể tại thời điểm t.

(a) Xây dựng hàm mục tiêu để ước lượng cho các tham số r và K của mô hình, sao cho sai số của mô hình là nhỏ nhất.

(b) Viết các bước giải thuật để xác định ước lượng của r và K bằng phương pháp Newton. Viết một chương trình thi hành giải thuật này.

(c) Viết các bước giải thuật để xác định ước lượng của r và K bằng phương pháp quasi-Newton. Viết một chương trình thi hành giải thuật này. So sánh với kết quả câu (b).

(d) Trong nhiều mô hình hóa dân số, người ta áp dụng giả định về log-normality. Giả định đơn giản nhất là $\log(N_t)$ độc lập và tuân theo phân phối chuẩn với trung bình $\log\{g(t)\}$ và phương sai σ^2. Tìm MLE cho r, K và σ^2 theo giả định này, sử dụng cả 3 phương pháp Newton, Fisher Scoring và quasi-Newton. Cung cấp các sai số chuẩn của ước lượng MLE và ước tính hệ số tương quan giữa chúng. Nhận xét về kết quả.

Tài liệu tham khảo

[1] A. P. Dempster, ; N. M. Laird; and D. B. Rubin (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, 45, 1–37.

[2] E. J. Freireich, E. Gehan, E. Frei III, L. R. Schroeder, I. J. Wolman, R. Anabari, E. O. Burgert, S. D. Mills, D. Pinkel, O. S. Selawry, J. H. Moon, B. R. Gendel, C. L. Spurr, R. Storrs, F. Haurani, B. Hoogstraten, and S. Lee (1963). The effect of 6-mercaptopurine on the duraction of steriod-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, 21(6):699-716.

[3] C. M. Anderson and R. P. Labelle (2000). Update of comparative occurrence rates for offshore oil spills. *Spill Science and Technology Bulletin*, 6:303-321.

[4] R. N. Chapman (1928). The quantitative analysis of environmental factors. *Ecology*, 9:111-122.