Gabarito - Quarta Prova
Fundamentos Estatísticos para Ciência dos Dados

04/07/2016

Adaptado de questão de exame do curso *EE511 - Introduction to Statistical Learning*, da *University of Washington*, 2008.

Seja o tempo de espera T até que um comentário seja postado por um certo usuário sobre um vídeo no YouTube. Este tempo de espera pode ser de dois tipos: 1 e 2. O primeiro tipo é um tempo de espera de usuários que escrevem pouco, e segue uma distribuição exponencial com parâmetro λ_1. O segundo tipo é um tempo de espera de usuários mais ativos e também segue uma distribuição exponencial com parâmetro λ_2. A proporção de usuários do tipo 1 é π e a proporção do segundo tipo é $1-\pi$.

1. Você recebe uma amostra rotulada $(t_1, r_1), \ldots, (t_n, r_n)$ de dados i.i.d. onde $r_i = 1$ ou $r_i = 2$ indica a população do usuário i, do tipo 1 ou tipo 2. O MLE de π é simplesmente a proporção de indivíduos da população do primeiro tipo. Obtenha o MLE de λ_1 e λ_2

   Solução: Podemos pensar em duas maneiras de resolver e você recebeu créditos se tiver feito de qualquer uma dessas meneiras. Na primeira, mais simples e estimulada pelo enunciado do problema, ignoramos o parâmetro π e focamos na estimação apenas de λ_1 e λ_2. Seja n_1 o número de elementos (t_i, r_i) com $r_i = 1$ e

   $$\bar{t}_1 = \frac{1}{n_1} \sum_i t_i I[r_i = 1]$$

   a média aritmética de seus tempos de espera. De maneira análoga, defina $n_2 = n - n_1$ e $\bar{t}_2$.

   A verossimilhaça de λ_1 e λ_2 é dada por:

   $$\begin{aligned} L(\lambda_1, \lambda_2) &= \prod_{i=1}^{n} f(x_i, r_i; \lambda_1, \lambda_2, \pi) \\ &= \prod_{\substack{i=1 \\ r_i=1}}^{n} f(x_i, r_i; \lambda_1, \lambda_2, \pi) \quad \prod_{\substack{i=1 \\ r_i=2}}^{n} f(x_i, r_i; \lambda_1, \lambda_2, \pi) \\ &= \prod_{\substack{i=1 \\ r_i=1}}^{n} \lambda_1 \exp(-\lambda_1 t_i) \quad \prod_{\substack{i=1 \\ r_i=2}}^{n} \lambda_2 \exp(-\lambda_2 t_i) \\ &= \lambda_1^{n_1} \exp(-\lambda_1 n_1 \bar{t}_1) \;\; \lambda_2^{n_2} \exp(-\lambda_2 n_2 \bar{t}_2) \end{aligned}$$

   A log-verossmilhança (que é muito mais fácil de derivar) é dada por

   $$\ell(\lambda_1, \lambda_2) = n_1 \log(\lambda_1) - \lambda_1 n_1 \bar{t}_1 + n_2 \log(\lambda_2) - \lambda_2 n_2 \bar{t}_2 \,.$$

   A sua derivada em relaąão a λ_1 é

   $$\frac{\partial \ell}{\partial \lambda_1} = \frac{n_1}{\lambda_1} - n_1 \bar{t}_1$$

   o que implica no MLE $\hat{\lambda}_1 = 1/\bar{t}_1$, o inverso da média aritmética dos tempos de espera. De maneira análoga, encontra-se $\hat{\lambda}_2 = 1/\bar{t}_2$.

Na segunda soluç ao, mais apropriada, o parãmetro π é incorporado e a verossimilhança é

$$L(\lambda_1, \lambda_2, \pi) = \prod_{i=1}^{n} \left(\pi\lambda_1 \exp(-\lambda_1 t_i)\right)^{2-r_i} \left((1-\pi)\lambda_2 \exp(-\lambda_2 t_i)\right)^{r_i - 1} .$$

Você pode ter trocado r_i por $z_i = r_i - 1$ para ficar com rótulos valendo 0 ou 1 e assim terminar com fórmulas mais fáceis. Vou seguir com a notação r_i do problema.

A log-verossmilhança é igual a

$$\ell(\lambda_1, \lambda_2, \pi) = \sum_{i=1}^{n} (2 - r_i) \log\left(\pi\lambda_1 \exp(-\lambda_1 t_i)\right) + (r_i - 1) \log\left((1-\pi)\lambda_2 \exp(-\lambda_2 t_i)\right)$$

Notando que $\sum_i (2 - r_i) = n_1$ e que $\sum_i (2 - r_i) t_i = \bar{t}_1$ temos

$$\ell(\lambda_1, \lambda_2, \pi) = n_1 \log(\pi) + n_1 \log(\lambda_1) - \lambda_1 \bar{t}_1 + n_2 \log(1-\pi) + n_2 \log(\lambda_2) - \lambda_2 \bar{t}_2$$

Tomando as derivadas parciais, obtemos

$$\frac{\partial \ell}{\partial \pi} = \frac{n_1}{\pi} - \frac{n_1}{1-\pi}$$

o que implica em $\hat{\pi} = n_1/(n_1 + n_2)$.

Como antes,

$$\frac{\partial \ell}{\partial \lambda_1} = \frac{n_1}{\lambda_1} - n_1 \bar{t}_1$$

produzindo $\hat{\lambda}_1 = 1/\bar{t}_1$. Analogamente, para λ_2.

---

2. Você recebe uma amostra $t_1, \ldots, t_n$ sem os rótulos da população a qual pertence o usuário. Obtenha os dois passos do algoritmo EM para os parâmetros λ_1, λ_2 e π.

   Solução: Para o passo E do algoritmo *EM*, você precisa da log-verossmilhança completa, como se todos os dados estivessem disponíveis. Esta log-verossimilhança completa já foi obtida no item anterior:

$$\ell^c(\lambda_1, \lambda_2, \pi) = \sum_{i=1}^{n} (2 - r_i) \log\left(\pi\lambda_1 \exp(-\lambda_1 t_i)\right) + (r_i - 1) \log\left((1-\pi)\lambda_2 \exp(-\lambda_2 t_i)\right)$$

   Agora, substitua os valores r_i que, de fato, não observados pelas variáveis aleatórias R_i e tome a sua esperança condicionada nos dados $t_1, \ldots t_n$:

$$\mathbb{E}\left[\ell^c(\lambda_1, \lambda_2, \pi) | \mathbf{T} = \mathbf{t}\right] = \sum_{i=1}^{n} \mathbb{E}[(2 - R_i)|t_i] \log\left(\pi\lambda_1 \exp(-\lambda_1 t_i)\right) + \mathbb{E}[(R_i - 1)|t_i] \log\left((1-\pi)\lambda_2 \exp(-\lambda_2 t_i)\right)$$

   Como $2 - R_i$ é uma v.a. binária, temos

$$\hat{\pi}_i = \mathbb{E}[(2 - R_i)|t_i] = \mathbb{P}[R_i = 1|t_i] = \frac{\pi\lambda_1 \exp(-\lambda_1 t_i)}{\pi\lambda_1 \exp(-\lambda_1 t_i) + (1-\pi)\lambda_2 \exp(-\lambda_2 t_i)} \tag{1}$$

   onde $\hat{\pi}_i$ é avaliado com os valores correntes $\boldsymbol{\theta}_0$ de $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \pi)$.

   Para o passo M, temos de maximizar em $\boldsymbol{\theta}$ a função

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \mathbf{t}) = \sum_{i=1}^{n} \hat{\pi}_i \log\left(\pi\lambda_1 \exp(-\lambda_1 t_i)\right) + (1 - \hat{\pi}_i) \log\left((1-\pi)\lambda_2 \exp(-\lambda_2 t_i)\right)$$

o que resulta em

$$\hat{\pi} = \frac{1}{n}\sum_i \hat{\pi}_i \tag{2}$$

$$\hat{\lambda}_1 = \frac{\sum_i \hat{\pi}_i}{\sum_i \hat{\pi}_i t_i} \tag{3}$$

$$\hat{\lambda}_2 = \frac{\sum_i (1-\hat{\pi}_i)}{\sum_i (1-\hat{\pi}_i) t_i} \tag{4}$$

Com estes novos valores para $\boldsymbol{\theta}$, novas estimativas para $\hat{\pi}_i$ são obtidas com (1). Com estes novos valores de $\hat{\pi}_i$, novas estimativas para $\boldsymbol{\theta}$ são obtidas usando (2), (3), (4). O algoritmo EM itera estes dois passos até convergência.