Quarto Trabalho Prático de Estatística Computacional
- Algoritmo EM e Amostrador de Gibbs -

Este TP consiste na implementação (em linguagem R) do algoritmo EM e do amostrador de Gibbs para uma mistura de binomiais. Vamos considerar o exemplo estudado na aula.

Duas máquinas, A e B, produzem as peças de uma fábrica. Cada peça produzido pela máquina A tem probabilidade θ_A de ser defeituosa, ao passo que para a máquina B essa probabilidade é θ_B. Suponha que N lotes, de n peças cada um, foram inspecionados. As peças de um lote são produzidas pela mesma máquina, mas desconhecemos a máquina que produziu as peças de cada lote. Queremos estimar as probabilidades θ_A e θ_B, bem como a proporção de peças produzidas pelas máquinas A e B.

Sejam $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$ em que Y_i representa o número de peças do i-ésimo lote, dentre as n, que são defeituosas e p a proporção de peças produzidas pela máquina B (logo, $(1-p)$ é a proporção de peças produzidas pela máquina A).

Sejam, ainda, $\delta = \{\delta_1, \ldots, \delta_N\}$ em que $\delta_i = 0$ se o i-ésimo lote foi produzido pela máquina A e $\delta_i = 1$ se foi produzido pela máquina B.

Algoritmo EM

Sejam $P_A(Y|\theta_A) = \theta_A^Y(1-\theta_A)^{n-Y}$ e $P_B(Y|\theta_B) = \theta_B^Y(1-\theta_B)^{n-Y}$. O algoritmo EM pode ser escrito nos seguintes passos:

- Inicializar $\theta_A^{(0)}$, $\theta_B^{(0)}$ e $p^{(0)}$
- E-step: $\gamma_i^{(k)} = E(\delta_i|\theta_A^{(k)}, \theta_B^{(k)}, p^{(k)}) = \frac{p^{(k)}P_B(Y_i|\theta_B^{(k)})}{(1-p^{(k)})P_A(Y_i|\theta_A^{(k)}) + p^{(k)}P_B(Y_i|\theta_B^{(k)})}$, $i = 1, \ldots, N$
- M-step: $\theta_A^{(k+1)} = \frac{\sum_{i=1}^N Y_i(1-\gamma_i^{(k)})}{n\sum_{i=1}^N(1-\gamma_i^{(k)})}$, $\theta_B^{(k+1)} = \frac{\sum_{i=1}^N Y_i\gamma_i^{(k)}}{n\sum_{i=1}^N\gamma_i^{(k)}}$, $p = \frac{\sum_{i=1}^N \gamma_i^{(k)}}{N}$
- Iterar os passos E e M para $k = 0, 1, \ldots$, até covergência.

Inicializar θ_A e θ_B tomando elemento da amostra ao acaso e dividindo por n, e p usando uma distribuição uniforme(0,1). Parar o algoritmo quando, para todo i, $|\gamma_i^{(k+1)} - \gamma_i^{(k)}| < 10^{-9}$.

Amostrador de Gibbs

Se considerarmos as prioris

- $\theta_A \sim Beta(\alpha_A, \beta_A)$
- $\theta_B \sim Beta(\alpha_B, \beta_B)$
- $p \sim Beta(\alpha_p, \beta_p)$

então as posterioris são dadas por

- $\theta_A|Y, \delta \sim Beta(\alpha_A + \sum Y_i(1-\delta_i), \beta_A + \sum(n-Y_i)(1-\delta_i))$
- $\theta_B|Y, \delta \sim Beta(\alpha_B + \sum Y_i\delta_i, \beta_B + \sum(n-Y_i)\delta_i)$
- $p|Y, \delta \sim Beta(\alpha_p + \sum \delta_i, \beta_p + \sum(1-\delta_i))$

Assim, para estimar θ_A, θ_B e p, o amostrador de Gibbs pode ser resumido nos passos:

1. Inicializar os parâmetros, gerando
   - $\theta_A^{(0)} \sim Beta(\alpha_A, \beta_A)$
   - $\theta_B^{(0)} \sim Beta(\alpha_B, \beta_B)$
   - $p^{(0)} \sim Beta(\alpha_p, \beta_p)$
2. Para $k = 0, \ldots, B+S$

(a) Computar $\delta_i^{(k)} = \frac{p^{(k)} P_B(Y_i|\theta_B^{(k)})}{(1-p^{(k)})P_A(Y_i|\theta_A^{(k)})+p^{(k)}P_B(Y_i|\theta_B^{(k)})}$, $i = 1, \ldots, N$

(b) Gerar $\theta_A^{(k+1)}|Y, \delta^{(k)} \sim Beta(\alpha_A + \sum Y_i(1-\delta_i^{(k)}), \beta_A + \sum(n-Y_i)(1-\delta_i^k))$

(c) Gerar $\theta_B^{(k+1)}|Y, \delta^{(k)} \sim Beta(\alpha_B + \sum Y_i\delta_i^{(k)}, \beta_B + \sum(n-Y_i)\delta_i^{(k)})$

(d) Gerar $p^{(k+1)}|Y, \delta^{(k)} \sim Beta(\alpha_p + \sum \delta_i^{(k)}, \beta_p + \sum(1-\delta_i^{(k)}))$
(em que todos os somatórios são de $i = 1$ a N)

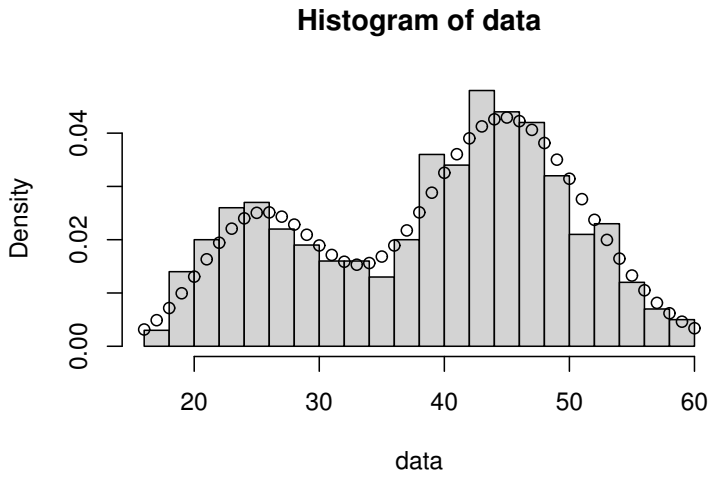
3. Obter as estimativas de θ_A, θ_B e p calculando as respectivas médias das S últimas amostras (isto é, ignorando as B primeiras amostras).

Utilizar o arquivo de dados fornecido contendo o número de peças defeituosas em $N = 500$ lotes de $n = 500$ peças cada um. Considerar prioris não informativas ($\alpha_A = \beta_A = \alpha_B = \beta_B = \alpha_p = \beta_p = 1$), $B = 10.000$ e $S = 10.000$.

Para cada algoritmo, após obter as estimativas para p, θ_A e θ_B, plotar no mesmo gráfico o histograma dos dados juntamente com a função de probabilidade estimada para a mistura das binomiais, dada por

$$P(y) = (1-p)P(y|\theta_A) + pP(y|\theta_B)$$

em que $P(y|\theta_A)$ e $P(y|\theta_B)$ representam a função de probabilidade de uma binomial com probabilidades de sucesso dadas por θ_A e θ_B, respectivamente, obtendo um gráfico conforme o exemplo abaixo.



Todos os algoritmos devem ser implementados em linguagem R. O relatório do TP deve ser entregue em formato de artigo **em arquivo PDF**, através de submissão eletrônica no site. Um modelo de relatório está disponível no site da disciplina.
Incluir todos os códigos em um anexo ao fim do relatório.