MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA EM ALGORITHM

Dempster, A. P., Laird, N.M. e Rubin, D.B. (1977).

Rodrigo Matheus

UFRN - Universidade Federal do Rio Grande do Norte
CCET - Centro de Ciências Exatas e da Terra
Pet - Estatística

01 de agosto de 2016

Sumário

Introdução

É um algoritmo amplamente aplicável para calcular estimativas de máxima verossimilhança de dados incompletos.

Algumas aplicações incluem: dados censurados, modelos de mistura finita e componente de variância.

O exemplo abordado é um exemplo clássico que consta no artigo original, e mostra uma das aplicações particulares do algoritmo EM, calcular a moda a posteriori em abordagens Bayesianas.

Introdução

É um algoritmo amplamente aplicável para calcular estimativas de máxima verossimilhança de dados incompletos.

Algumas aplicações incluem: dados censurados, modelos de mistura finita e componente de variância.

O exemplo abordado é um exemplo clássico que consta no artigo original, e mostra uma das aplicações particulares do algoritmo EM, calcular a moda a posteriori em abordagens Bayesianas.

Introdução

É um algoritmo amplamente aplicável para calcular estimativas de máxima verossimilhança de dados incompletos.

Algumas aplicações incluem: dados censurados, modelos de mistura finita e componente de variância.

O exemplo abordado é um exemplo clássico que consta no artigo original, e mostra uma das aplicações particulares do algoritmo EM, calcular a moda a posteriori em abordagens Bayesianas.

Introdução

Para motivar o algoritmo EM, considere a situação retratada na figura. Os o's representam o tempo de falha de um evento, enquanto os x's representam o tempo de falha até a censura.

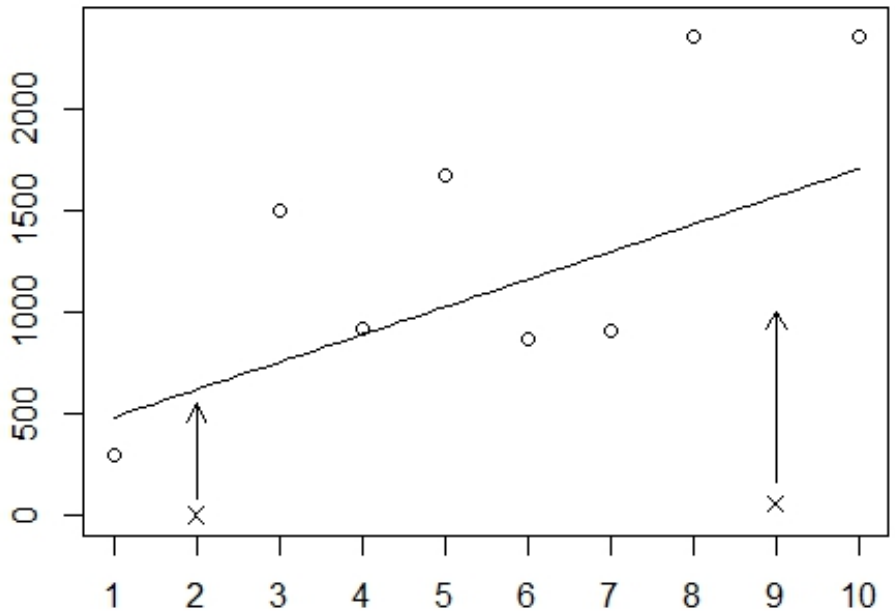

Figura: Tempos de falhas e censuras.

A idéia é preencher os valores faltantes com valores estimados e então atualizar a estimativa do parâmetro via algoritmo EM.

Mais especificamente, o algoritmo EM é um método iterativo que ao invés de fazer uma complicada maximização, executa uma série de maximizações mais simples.

Toda iteração consiste em dois passos: O passo E (esperança) e o passo M (maximização).

A idéia é preencher os valores faltantes com valores estimados e então atualizar a estimativa do parâmetro via algoritmo EM.

Mais especificamente, o algoritmo EM é um método iterativo que ao invés de fazer uma complicada maximização, executa uma série de maximizações mais simples.

Toda iteração consiste em dois passos: O passo E (esperança) e o passo M (maximização).

A idéia é preencher os valores faltantes com valores estimados e então atualizar a estimativa do parâmetro via algoritmo EM.

Mais especificamente, o algoritmo EM é um método iterativo que ao invés de fazer uma complicada maximização, executa uma série de maximizações mais simples.

Toda iteração consiste em dois passos: O passo E (esperança) e o passo M (maximização).

Definições

Seja $\mathbf{X} = (X_1, ..., X_n)$, amostra aleatória (a.a.) observada independente e identicamente distribuída (iid), com função de densidade (ou de probabilidade) $g(\mathbf{x}|\theta)$. Desejamos computar

$$\hat{\theta} = \arg\max g(\mathbf{x}|\theta) = \prod_{i=1}^{n} g(x_i|\theta).$$

A amostra completa é obtida com o aumento de $\mathbf{z} = (z_1, ..., z_m)$, onde $\mathbf{X},\mathbf{Z} \sim f(\mathbf{x},\mathbf{z}|\theta)$.

Definições

Seja $\mathbf{X} = (X_1, ..., X_n)$, amostra aleatória (a.a.) observada independente e identicamente distribuída (iid), com função de densidade (ou de probabilidade) $g(\mathbf{x}|\theta)$. Desejamos computar

$$\hat{\theta} = \arg\max g(\mathbf{x}|\theta) = \prod_{i=1}^{n} g(x_i|\theta).$$

A amostra completa é obtida com o aumento de $\mathbf{z} = (z_1, ..., z_m)$, onde $\mathbf{X}$,$\mathbf{Z} \sim f(\mathbf{x}$,$\mathbf{z}|\theta)$.

Definições

Essas funções se relacionam da seguinte forma:

$$g(\mathbf{x}|\theta) = \int_{\mathbf{Z}} f(\mathbf{x},\mathbf{z}|\theta)d\mathbf{z}$$

Definições

A distribuição condicional de $\mathbf{Z}$ dado $\mathbf{X}$ é expressa a seguir,

$$k(\mathbf{z}|\mathbf{x},\theta) = \frac{f(\mathbf{x},\mathbf{z}|\theta)}{g(\mathbf{x}|\theta)} \tag{1}$$

Demonstração:

$$P(\mathbf{z}|\mathbf{x},\theta) = \frac{P(\mathbf{z},\mathbf{x},\theta)}{P(\mathbf{x},\theta)} = \frac{P(\mathbf{z},\mathbf{x}|\theta)P(\theta)}{P(\mathbf{x}|\theta)P(\theta)} = \frac{P(\mathbf{x},\mathbf{z}|\theta)}{P(\mathbf{x}|\theta)}$$

Definições

Maximizar g é equivalente a maximizar $\log(g)$. Da relação (1), implica que g pode ser escrita como:

$$g(\mathbf{x}|\theta) = \frac{f(\mathbf{x},\mathbf{z}|\theta)}{k(\mathbf{z}|\mathbf{x},\theta)},$$

assim,

$$\log[g(\mathbf{x}|\theta)] = \log[f(\mathbf{x},\mathbf{z}|\theta)] - \log[k(\mathbf{z}|\mathbf{x},\theta)] \ .$$

Definições

Maximizar g é equivalente a maximizar $\log(g)$. Da relação (1), implica que g pode ser escrita como:

$$g(\mathbf{x}|\theta) = \frac{f(\mathbf{x},\mathbf{z}|\theta)}{k(\mathbf{z}|\mathbf{x},\theta)},$$

assim,

$$\log[g(\mathbf{x}|\theta)] = \log[f(\mathbf{x},\mathbf{z}|\theta)] - \log[k(\mathbf{z}|\mathbf{x},\theta)] \ .$$

Definições

Agora integre ambos os lados da equação com respeito a distribuição condicional $k(\mathbf{z}|\mathbf{x},\theta)$ obtendo assim:

$$\log[g(\mathbf{x}|\theta)] = \int_{\mathbf{Z}} \log[f(\mathbf{x},\mathbf{z}|\theta)]k(\mathbf{z}|\mathbf{x},\theta)d\mathbf{z} - \int_{\mathbf{Z}} \log[k(\mathbf{z}|\mathbf{x},\theta)]k(\mathbf{z}|\mathbf{x},\theta)d\mathbf{z}.$$

Dempster, Laird e Rubin (1977) definem a função Q por:

$$Q(\theta,\theta^*) = \int_{\mathbf{Z}} \log[f(\mathbf{x},\mathbf{z}|\theta)]k(\mathbf{z}|\mathbf{x},\theta)d\mathbf{z} = E\left\{\log[f(\mathbf{x},\mathbf{Z}|\theta)]|\theta,\mathbf{x}\right\}.$$

Definições

Agora integre ambos os lados da equação com respeito a distribuição condicional $k(\mathbf{z}|\mathbf{x},\theta)$ obtendo assim:

$$\log[g(\mathbf{x}|\theta)] = \int_{\mathbf{Z}} \log[f(\mathbf{x},\mathbf{z}|\theta)]k(\mathbf{z}|\mathbf{x},\theta)d\mathbf{z} - \int_{\mathbf{Z}} \log[k(\mathbf{z}|\mathbf{x},\theta)]k(\mathbf{z}|\mathbf{x},\theta)d\mathbf{z}.$$

Dempster, Laird e Rubin (1977) definem a função Q por:

$$Q(\theta,\theta^*) = \int_{\mathbf{Z}} \log[f(\mathbf{x},\mathbf{z}|\theta)]k(\mathbf{z}|\mathbf{x},\theta)d\mathbf{z} = E\left\{\log[f(\mathbf{x},\mathbf{Z}|\theta)]|\theta,\mathbf{x}\right\}.$$

Definições

O algoritmo EM é implementado então pelos seguintes passos.

Passo E:
Calcular $Q(\theta, \theta_i)$.

Passo M:
Encontrar θ_{i+1} que maximiza $Q(\theta, \theta_i)$.

Definições

Teorema 1. Todo algoritmo EM incrementa $g(\mathbf{x}|\theta)$ a cada iteração, i.e., $g(\mathbf{x}|\theta_{i+1}) \geq g(\mathbf{x}|\theta_i)$.

Teorema 2. Suponha que uma sequência de iterações EM, θ_i, satisfaz:
1. $\frac{\partial Q(\theta,\theta_i)}{\partial \theta_i} \mid_{\theta=\theta_{i+1}} = 0$;
2. θ_i converge para algum θ^* e $k(\mathbf{z}|\mathbf{x},\theta)$ é "suficientemente" suave.

Então segue que:

$$\frac{\partial \log g(\mathbf{x}|\theta)}{\partial \theta_i} \mid_{\theta=\theta^*} = 0$$

Exemplo - Rao (1973)

Em uma certa população de animais sabe-se que cada animal pode pertencer a uma dentre 4 linhagens genéticas com probabilidades,

$$p_1 = \frac{1}{2} + \frac{\theta}{4}, p_2 = \frac{1-\theta}{4}, p_3 = \frac{1-\theta}{4}, p_4 = \frac{\theta}{4}, 0 < \theta < 1.$$

Observando-se $n = 197$ animais dentre os quais x_i pertencem à linhagem i então o vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, X_4)$ tem distribuição multinomial com parâmetros n, p_1, p_2, p_3, p_4.A função de probabilidade a posteriori fica,

$$g(\mathbf{x}|\theta) = \frac{n!}{x_1!x_2!x_3!x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \propto (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}.$$

Exemplo - Rao (1973)

Em uma certa população de animais sabe-se que cada animal pode pertencer a uma dentre 4 linhagens genéticas com probabilidades,

$$p_1 = \frac{1}{2} + \frac{\theta}{4}, p_2 = \frac{1-\theta}{4}, p_3 = \frac{1-\theta}{4}, p_4 = \frac{\theta}{4}, 0 < \theta < 1.$$

Observando-se $n = 197$ animais dentre os quais x_i pertencem à linhagem i então o vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, X_4)$ tem distribuição multinomial com parâmetros n, p_1, p_2, p_3, p_4. A função de probabilidade a posteriori fica,

$$g(\mathbf{x}|\theta) = \frac{n!}{x_1!x_2!x_3!x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \propto (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}.$$

Exemplo - Rao (1973)

Em uma certa população de animais sabe-se que cada animal pode pertencer a uma dentre 4 linhagens genéticas com probabilidades,

$$p_1 = \frac{1}{2} + \frac{\theta}{4}, p_2 = \frac{1-\theta}{4}, p_3 = \frac{1-\theta}{4}, p_4 = \frac{\theta}{4}, 0 < \theta < 1.$$

Observando-se $n = 197$ animais dentre os quais x_i pertencem à linhagem i então o vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, X_4)$ tem distribuição multinomial com parâmetros n, p_1, p_2, p_3, p_4.A função de probabilidade a posteriori fica,

$$g(\mathbf{x}|\theta) = \frac{n!}{x_1!x_2!x_3!x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \propto (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}.$$

Exemplo - Rao (1973)

Considere agora outro vetor aleatório $\mathbf{Y} = (Z_1, Z_2, X_2, X_3, X_4)$, tal que $Z_1 + Z_2 = X_1$, com probabilidades,

$$\pi_1 = \frac{1}{2}, \pi_2 = \frac{\theta}{4}, \pi_3 = p_2, \pi_4 = p_3, \pi_5 = p_4.$$

A função de probabilidade a posteriori pode ser reescrita como,

$$f(\mathbf{x},\mathbf{z}|\theta) = \frac{n!}{z_1!z_2!x_2!x_3!x_4!}\left(\frac{1}{2}\right)^{z_1}\left(\frac{1-\theta}{4}\right)^{x_2+x_3}\left(\frac{\theta}{4}\right)^{z_2+x_4} \propto (1-\theta)^{x_2+x_3}\theta^{z_2+x_4}.$$

Exemplo - Rao (1973)

Considere agora outro vetor aleatório $\mathbf{Y} = (Z_1, Z_2, X_2, X_3, X_4)$, tal que $Z_1 + Z_2 = X_1$, com probabilidades,

$$\pi_1 = \frac{1}{2}, \pi_2 = \frac{\theta}{4}, \pi_3 = p_2, \pi_4 = p_3, \pi_5 = p_4.$$

A função de probabilidade a posteriori pode ser reescrita como,

$$f(\mathbf{x},\mathbf{z}|\theta) = \frac{n!}{z_1!z_2!x_2!x_3!x_4!}\left(\frac{1}{2}\right)^{z_1}\left(\frac{1-\theta}{4}\right)^{x_2+x_3}\left(\frac{\theta}{4}\right)^{z_2+x_4} \propto (1-\theta)^{x_2+x_3}\theta^{z_2+x_4}.$$

Exemplo - Rao (1973)

A distribuição condicional dos dados faltantes dado a amostra observada e θ é,

$$k(\mathbf{z}|\mathbf{x},\theta) = \frac{f(\mathbf{x},\mathbf{z}|\theta)}{g(\mathbf{x}|\theta)} \propto \left(\frac{\theta}{2+\theta}\right)^{z_2} \left(\frac{2}{2+\theta}\right)^{x_1-z_2},$$

então,

$$Z_2 \sim Binomial\left(X_1, \frac{\theta}{2+\theta}\right).$$

Exemplo - Rao (1973)

A distribuição condicional dos dados faltantes dado a amostra observada e θ é,

$$k(\mathbf{z}|\mathbf{x},\theta) = \frac{f(\mathbf{x},\mathbf{z}|\theta)}{g(\mathbf{x}|\theta)} \propto \left(\frac{\theta}{2+\theta}\right)^{z_2} \left(\frac{2}{2+\theta}\right)^{x_1-z_2},$$

então,

$$Z_2 \sim Binomial\left(X_1, \frac{\theta}{2+\theta}\right).$$

Exemplo - Rao (1973)

Pode-se mostrar que a função Q é expressa por:

$$Q(\theta, \theta_i) = \log(\theta_i) E[Z_2|\mathbf{x}, \theta_i] + x_4 \log(\theta_i) + (x_2 + x_3) \log(1 - \theta),$$

e,

$$\frac{\partial Q(\theta, \theta_i)}{\partial \theta} = 0 \Rightarrow \theta_i = \frac{E[Z_2|\mathbf{x}, \theta_i] + x_4}{E[Z_2|\mathbf{x}, \theta_i] + x_2 + x_3 + x_4}$$

Exemplo - Rao (1973)

Considerando a seguinte amostra observada $\mathbf{X} = (125,18,20,34)$, o algoritmo EM então pode ser implementado da seguinte maneira:

Passo E: Calcule $E[Z_2|\mathbf{x}, \theta_i]$

$$E[Z_2|\mathbf{x}, \theta_i] = 125\left(\frac{\theta}{2+\theta}\right)$$

Passo M: Calcular θ_{i+1} que maximiza $Q(\theta, \theta_i)$

$$\theta_{i+1} = \frac{E[Z_2|\mathbf{x}, \theta_i] + 34}{E[Z_2|\mathbf{x}, \theta_i] + 18 + 20 + 34}$$

Comandos R Core Team (2014)

```
oa = 0.1
repeat{

  # Passo E
    e = 125*(oa/(2 + oa))

  # Passo M
    on = (e+34)/(e + 18 + 20 + 34)

# Critério de parada
  if(abs(oa-on)<10^(-7)){
    break
  }else{
    oa = on
  }
}
```

Exemplo - Rao (1973)

Tabela: Atualização da estimativa EM.

Iteração	θ_i	$\theta_* - \theta^i$
1	0,1	0,5268215
2	0,5125229	0,1142986
3	0,6102501	0,0165714
4	0,624594	0,0022275
5	0,6265252	0,0002963
6	0,6267822	0,0000393
7	0,6268163	0,0000052
8	0,6268208	0,0000007
9	0,6268214	0,0000001

em que θ^* é a estimativa de máxima verossimilhança de θ.

Referências

- *CASELLA, George.* ***Monte Carlo Statistical Methods.*** *University of Florida:. 2008.*
- *EHLERS, Ricardo S.* ***Métodos Computacionalmente Intensivos em Estatística****.2.ed.Paraná: UFPR. 2004.*
- *R Core Team.* ***R: A language and environment for statistical computing.*** *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.2014*
- *TANNER, Martin A. Tools for statistical inference. New York: Springer, 1996.*