Técnicas Computacionais em Probabilidade e Estatística I

Aula IX

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

Slide 1

c1 chang; 22/4/2008

Método de otimização Numérica

1. Introdução:

Objetivo: apresentar alguns algoritmos utilizados para maximizar uma função $f(\theta)$, onde θ é um parâmetro d-dimensional.

$$\theta = (\theta_1, ..., \theta_d)'$$

Algoritmos: Newton-Raphson

Método Scoring

Gauss- Newton

Situação de interesse: $f(\theta)$ é a função de verossimilhança ou a função densidade a posteriori.

Todos esses algoritmos são procedimentos iterativos:

$\Theta^{(i)}$: valor computado no i-ésimo estágio.

$\Theta^{(i+1)}$: valor atualizado no estágio (i+1).

O procedimento é repetido até obter convergência.

O contexto da aplicação desses algoritmos é o de estimação de parâmetros. No caso da função de verossimilhança (fv), busca-se o estimador de máxima verossimilhança (EMV) de θ; no caso de função densidade a posteriori (fdp), a moda dessa função.

Modelos lineares com erros independentes e normais $\rightarrow$ fv quadrática e podemos obter o máximo resolvendo-se um sistema de equações lineares nos parâmetros (forma fechada)

Modelos não lineares $\rightarrow$fv não é quadrática nos parâmetros e temos que resolver equações não lineares. Ex: modelos de regressão não linear, etc,

$X=(X_1, ..., X_n)'$ dados observados, obtidos da densidade $f(x/\theta)$.

$L(\theta/X)$: função de verossimilhança (fv)

$l(\theta/X)$: logaritmo da fv

Se $X_1, ..., X_n$ são iid com densidade $f(x/\theta)$, então:

$L(\theta/X) = \sum l_i(\theta/X_i)$

Além disso, se as v.a. X_i são gaussianas, a log-verossimilhança será quadrática.

No enfoque Bayesiano, suponha que tenha densidade a priori p(θ). Então o teorema de bayes dá a densidade a posteriori (condicionada sobre as observações X)

$$p(\theta / X) = \frac{f(x/\theta)\, p(\theta)}{f(x)}$$

$$f(x) = \int f(x,\theta) d\theta \qquad f(x) = \sum_{\theta} f(x,\theta)$$

$$p(\theta / x) \propto p(\theta) L(\theta / X)$$

O EMV de θ é o valor do parâmetro que maximiza $L(\theta/X)$ ou $l(\theta/X)$. Se a fv for derivável, unimodal e limitada superiormente, então a moda (EMV) $\hat{\theta}$

é obtida derivando-se L ou l, com respeito às componentes de θ, considerando-se $\partial l/\partial\theta=0$ resolvendo as d equações resultantes.

De um modo geral, uma solução analítica em forma fechada das equações de verossimilhança não pode ser encontrada. Recorre-se, então, a algum procedimento de otimização numérica para encontrar $\hat{\theta}$

Informações de Fisher (I(θ))

a) Caso unidimensional(d=1)

Função escore ⟶ equação de verossimilhança(EV)

$$\dot{l}(\theta / X) = \frac{\partial l(\theta / X)}{\partial \theta} \Rightarrow \dot{l}(\theta / X) = 0$$

Uma solução da EV é um EMV se

$$\ddot{l}(\theta / X) = \frac{\partial^2 l(\theta / X)}{\partial \theta^2} < 0$$

$$I(\theta) = -E_\theta\left[\ddot{l}(\theta / X)\right] = E_\theta\left[\dot{l}(\theta / X)\right]^2$$

$I(\theta)$: informação de Fisher sobre θ, contida em X.

$E_\theta(.)$: esperança relativa à distribuição de X para o valor do parâmetro igual a θ.

Resultado: sob condições de regularidade bastante gerais sobre a forma da fv, quando o verdadeiro valor do parâmetro é θ_0, a variância assintótica do EMV é

$$A\,\mathrm{var}_{\theta_0}(\hat{\theta}) = I^{-1}(\theta_0)$$

Como θ_0 é desconhecido, a precisão do EMV é medida de duas maneiras:

i) Informação de Fisher estimada: $\left|I(\hat{\theta})\right|$

ii) Informação observada: $\left[-\ddot{l}(\hat{\theta})\right]$

b) Caso vetorial, $\theta=(\theta_1, ..., \theta_d)$

$$I(\theta) = -E_\theta\left[\ddot{l}(\theta / X)\right] = E_\theta\left[\dot{l}(\theta / X)\right]\left[\dot{l}(\theta / X)\right]'$$

daqui em diante, utilizaremos a seguinte notação:

$\partial f(\theta)/\partial\theta = g(\theta)$, dx1, gradiente

$\partial^2 f(\theta)/\partial\theta\partial\theta = G(\theta)$, dxd, hessiano

No caso bidimensional, temos

$$\frac{\partial f(\theta)}{\partial \theta} = \left(\frac{\partial f}{\partial \theta_1} \quad \frac{\partial f}{\partial \theta_2}\right)$$

$$\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta} = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 f}{\partial \theta_2^2} \end{bmatrix}$$

Observação: problema de minimização são reduzidos a de maximização, pois

$$\max_{\theta} f(\theta) = \min_{\theta}(-f(\theta))$$

Resultado: Se $g(\theta)$ e $G(\theta)$ existirem e forem continuas na vizinhança de $\hat{\theta}$

$$g(\hat{\theta}) = 0$$

$$G(\hat{\theta}) < 0$$

São condições suficientes para que $\hat{\theta}$ seja um máximo local de $f(\theta)$.

Essas condições não garante que $\hat{\theta}$

seja um maximizador global de f(θ).

A maioria dos procedimentos iterativos são métodos gradientes, ou seja, baseados no cálculo de derivadas de f(θ) e são da forma :

$$\theta^{(i+1)} = \theta^{(i)} + \lambda s(\theta)$$

$\theta^{(i)}$: aproximação atual do máximo

$\theta^{(i+1)}$: estimador revisado

s(θ): vetor de direção, dx1, dependendo do gradiente g(θ)

λ: tamanho do passo para a mudança de θ.

Em geral,

$$s(\theta) = V(\theta)g(\theta),$$

onde $V(\theta)$ depende do método utilizado.

Dizemos que o procedimento convergiu se

a) $f(\theta^{(i+1)})$ estiver próxima de $f(\theta^{(i)})$;

b) $\theta^{(i+1)}$ estiver próximo de $\theta^{(i)}$

c) $g(\theta^{(i+1)})$ estiver próxima de $g(\theta^{(i)})$.

Se ε for um escalar pequeno e positivo, então (a) estará satisfeita se

$$\left| f(\boldsymbol{\theta}^{(i+1)}) - f(\boldsymbol{\theta}^{(i)}) \right| < \varepsilon$$

No caso de (a) e (b) temos que usar algum tipo de norma para medir a proximidade de dois vetores.

Caso univariado:

Se a solução estiver próxima de zero

$$\left|\theta^{(i+1)} - \theta^{(i)}\right| < \delta$$

Se a solução for um valor grande

$$\left|\frac{\theta^{(i+1)} - \theta^{(i)}}{\theta^{(i)}}\right| < \delta$$

Problemas de Convergência:

Usualmente relacionados com a escolha de um valor inicial, θ_0. Dependendo da forma de $f(\theta)$ e de θ_0, o algoritmo pode convergir para um máximo local e não para um global.

Os procedimentos iterativos podem depender de primeira e segunda derivadas, que devem ser calculadas no valor atual, $\theta^{(i)}$. Essas derivadas podem ser calculadas analiticamente, quando possível, ou então, numericamente.

Por exemplo,

$$\frac{\partial f}{\partial \theta_j} \cong \frac{f(\boldsymbol{\theta}^{(i)} + \delta_j \boldsymbol{e}_j) - f(\boldsymbol{\theta}^{(i)})}{\delta_j}, \; j = 1, \cdots, d$$

e_j: vetor(dx1) com coordenadas nulas, com exceção da j-ésima que é igual a 1;

δ_j: é um passo de comprimento suficientemente pequeno;

Derivadas segundas podem ser calculadas numericamente de modo análogo.