Aula 9: Otimização numérica

Prof. Dr. Eder Angelo Milani

05/06/2023

Introdução

Há, basicamente, duas classes de algoritmos de otimização: algoritmos baseados no gradiente de uma função objetivo, possivelmente definida no espaço R^n, e algoritmos não gradientes. Os primeiros são indicados para maximizar funções objetivo suaves, ou seja, deriváveis, em que há informação confiável sobre o seu gradiente. Em caso contrário, devemos recorrer a métodos não gradientes. Neste texto vamos nos concentrar na primeira classe.

Essencialmente, nosso objetivo é apresentar alguns algoritmos utilizados para maximizar ou minimizar uma função $f(\theta)$, em que $\theta = (\theta_1, \ldots, \theta_d)^T$ é um parâmetro d-dimensional.

Dentre os algoritmos mais utilizados, destacamos os algoritmos de Newton-Raphson, scoring, Gauss-Newton e Quase-Newton. Em Estatística, esses algoritmos são geralmente utilizados para maximizar a função de verossimilhança ou, no caso de inferência bayesiana, a função densidade a posteriori. Uma característica desses algoritmos é que eles são procedimentos iterativos em que em determinado estágio computa-se o valor $\theta^{(i)}$ que é utilizado para obter um valor atualizado $\theta^{(i+1)}$ no estágio seguinte. Esse processo é repetido até que haja convergência, ou seja, até que a diferença entre os resultados de dois passos consecutivos seja arbitrariamente pequena, por exemplo, em que $||\theta^{(i+1)} - \theta^{(i)}|| < \epsilon$ com $\epsilon > 0$, escolhido convenientemente.

O contexto mais comum de aplicação desses algoritmos é o de estimação de parâmetros. No caso de funções de verossimilhança, busca-se o estimador de máxima verossimilhança do parâmetro θ. No caso da função densidade a posteriori, procura-se sua moda (ou modas). Esse tipo de algoritmo também é usado em redes neurais, em que se minimiza uma função de perda, que pode ser uma soma de quadrados como em regressão, ou a entropia, no caso de classificação.

Em geral, para a maximização de uma função f em relação a um parâmetro d-dimensional θ, consideramos:

$$\frac{\partial f(\theta)}{\partial \theta} = g(\theta),$$

dimensão $d \times 1$.

$$\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T} = H(\theta),$$

com dimensão $d \times d$.

As funções $g(\theta)$ e $H(\theta)$ são conhecidas, respectivamente, por gradiente e hessiano da função f. Por exemplo, no caso bidimensional, em que $\theta = (\theta_1, \theta_2)^T$, temos

$$g(\theta) = \Big(\frac{\partial f(\theta)}{\partial \theta_1}, \frac{\partial f(\theta)}{\partial \theta_2}\Big),$$

e

$$H(\theta) = \Big(\frac{\partial^2 f(\theta)}{\partial \theta_1^2}, \frac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_2}, \frac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_2}, \frac{\partial^2 f(\theta)}{\partial \theta_2^2}\Big)$$

Problemas de minimização muitas vezes podem ser reduzidos a problemas de maximização, pois maximizar $f(\theta)$, com respeito a (θ), é equivalente a minimizar $-f(\theta)$ com respeito a θ.

Se $g(\theta)$ e $H(\theta)$ existirem e forem contínuas na vizinhança de $\hat{\theta}$, então $g(\hat{\theta}) = 0$ e $H(\hat{\theta})$ negativa definida são condições suficientes para que $\hat{\theta}$ seja um máximo local de $f(\theta)$. Essas condições não garantem que $\hat{\theta}$ seja um maximizador global de $f(\theta)$. Uma raiz nula da equação de estimação pode não ser um ponto de máximo ou mínimo, mas um ponto de sela, que é um máximo local com respeito a uma direção e um mínimo local com respeito a outra direção. Nesse caso, a matriz hessiana não é negativa definida.

Problemas de convergência dos algoritmos utilizados na maximização estão usualmente relacionados com a escolha de um valor inicial, $\theta^{(0)}$ para o processo iterativo.

A maioria dos procedimentos iterativos são métodos gradientes, ou seja, baseados no cálculo de derivadas de $f(\theta)$, e no caso uniparamétrico ($d = 1$), são da forma

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

em que $\theta^{(i)}$ é a aproximação atual do máximo, $\theta^{(i+1)}$ é o estimador revisado, $s(\theta)$ é o gradiente $g(\theta^{(i)})$ calculado no ponto $\theta^{(i)}$ e $\lambda > 0$ é o "tamanho do passo" para a mudança de $\theta^{(i)}$. Em geral, $s(\theta) = V(\theta)g(\theta)$ com $V(\theta)$ dependente do algoritmo usado para a maximização. Diferentes algoritmos baseados no método do gradiente descendente (steepest descent, ou gradient descent), quadratic hill climbing, método de Newton-Raphson etc, são tradicionalmente usados.

Dizemos que o procedimento iterativo convergiu se uma das seguintes condições for satisfeita:

i. $f(\theta^{(i+1)})$ estiver próxima de $f(\theta^{(i)})$;
ii. $\theta^{(i+1)}$ estiver próximo de $\theta^{(i)}$;
iii. $g(\theta^{(i+1)})$ estiver próxima de $g(\theta^{(i)})$.

Dado ϵ, um escalar pequeno positivo, então i) estará satisfeita se

$$|f(\theta^{(i+1)}) - f(\theta^{(i)})| < \epsilon.$$

No caso ii), para definir a convergência do algoritmo, podemos usar $|\theta^{(i+1)} - \theta^{(i)}| < \epsilon$, se a solução envolver um valor pequeno ou $|(\theta^{(i+1)} - \theta^{(i)})/\theta^{(i)}| < \epsilon$, se a solução for um valor grande. No caso multiparamétrico, ii) e iii) dependem de algum tipo de norma para medir a proximidade de dois vetores.

Os procedimentos iterativos podem depender de primeiras e segundas derivadas de $f(\theta)$, que, em cada passo, devem ser calculadas analítica ou numericamente no valor atual, $\theta^{(i)}$. Por exemplo, $\partial f(\theta)/\partial \theta_i$ pode ser calculada por

$$\frac{f(\theta^{(i)} + \delta) - f(\theta^{(i)})}{\delta},$$

em que δ é um passo de comprimento suficientemente pequeno. Derivadas segundas também podem ser calculadas numericamente de modo análogo.

Exemplo 1

Consideremos a função dada por

$$f(\theta) = \theta^2 - 4\theta + 3.$$

Então $g(\theta) = df(\theta)/df(\theta) = 2\theta - 4$ e $H(\theta) = d^2f(\theta)/d\theta^2 = 2 > 0$. Logo $\theta = 2$ é ponto de mínimo e o valor mínimo é -1.

Tomemos $\theta_1 = 0,5$ e $\theta_2 = \theta_2 + \delta$, com $\delta = 0,01$. Os verdadeiros valores da derivada em θ_1 e θ_2 são -3 e -2,98, respectivamente. Uma aproximação numérica da derivada no ponto $\theta_1 = 0,5$ é $(f(0,51) - f(0,5))/0,01 = (1,2201 - 1,25)/(0,01) = -2,99$, que está entre os dois valores acima.

Dada uma densidade $f(x|\theta)$, a função de verossimilhança, denotada por $L(\theta|x)$ é qualquer função de θ proporcional a $f(x|\theta)$. O logaritmo da função de verossimilhança (simplesmente, log-verossimilhança) será representado por $l(\theta|x)$. Se as variáveis $X_1, X_2, \ldots, X_n$ forem independentes e identicamente distribuídas, com densidade $f(x|\theta)$, então

$$l(\theta|x) = \sum_{i=1}^{n} l_i(\theta, x_i).$$

Um estimador de máxima verossimilhança de θ é um valor do parâmetro que maximiza $L(\theta)$ ou $l(\theta)$. Se a função de verossimilhança for derivável, unimodal e limitada superiormente, então o estimador de máxima verossimilhança (que é a moda, nesse caso) $\hat{\theta}$ é obtido derivando-se L ou l, com respeito aos componentes de θ, igualando essa derivada a zero e resolvendo as d equações resultantes. Em geral, uma solução analítica em forma fechada dessas d equações de estimação não pode ser encontrada e precisamos recorrer a algum procedimento de otimização numérica para obter $\hat{\theta}$.

Um instrumento importante na análise da verossimilhança é o conceito de informação de Fisher. Consideremos inicialmente, o caso unidimensional. A equação de estimação obtida por meio da maximização da log-verossimilhança é

$$g(\theta|x) = \frac{dl(\theta|x)}{d\theta} = 0,$$

em que nesse contexto, o gradiente $g(\theta|x)$ é conhecido como função escore. Uma solução dessa equação é um estimador de máxima verossimilhança se

$$h(\theta|x) = \frac{d^2l(\theta|x)}{d\theta^2} < 0.$$

A informação de Fisher sobre θ contida em x é definida por

$$I(\theta) = E_\theta([g(\theta|x)])^2 = E_\theta[h(\theta|x)],$$

em que E_θ denota a esperança relativa à distribuição de x, calculada com o valor do parâmetro igual a θ. Quando o verdadeiro valor do parâmetro é θ_0, pode-se demonstrar sob condições de regularidade bastante gerais sobre a forma da função de verossimilhança, que a variâmcia assintótica do estimador de máxima verossimilhança é

$$Var_{\theta_0}(\hat{\theta}) = I(\theta_0)^{-1}.$$

Como θ_0 não é conhecido, a precisão do estimador de máxima verossimilhança pode ser avaliada de duas maneiras, nomeadamente

i) informação de Fisher estimada

$$[I(\hat{\theta})] = \left\{ n^{-1} \sum_{i=1}^{n} E_\theta[d^2l(\theta|x_i)/d\theta]_{\theta=\hat{\theta}} \right\}^{-1}$$

ii) informação observada

$$-[H(\hat{\theta})]^{-1} = -\Big\{n^{-1}\sum_{i=1}^{n} d^2 l(\theta|x_i)/d\theta|_{\theta=\hat{\theta}}\Big\}^{-1}$$

No caso vetorial, em que $\theta = (\theta_1, \ldots, \theta_d)^T$, a matriz de informação de Fisher é definida por

$$I(\theta) = E_\theta\{[g(\theta|x)][g(\theta|x)]^T\} = -E_\theta[H(\theta|x)].$$

O método de Newton-Raphson

O procedimento de Newton-Raphson baseia-se na aproximação da função que se deseja maximizar por uma função quadrática. Para maximizar a log-verossimilhança, $l(\theta|x)$, consideremos a expansão de Taylor de segunda ordem ao redor do máximo $\hat{\theta}$

$$l(\theta|x) \approx l(\hat{\theta}|x) + (\theta - \hat{\theta})^T \frac{\partial l(\theta|x)}{\partial \theta}\Big|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})^T \frac{\partial^2 l(\theta|x)}{\partial\theta\partial\theta^T}\Big|_{\theta=\hat{\theta}}(\theta - \hat{\theta})$$

Então, para θ numa vizinhança de $\hat{\theta}$,

$$\frac{\partial l(\theta|x)}{\partial \theta} \approx \frac{\partial l(\theta|x)}{\partial \theta}\Big|_{\theta=\hat{\theta}} + \frac{\partial^2 l(\theta|x)}{\partial\theta\partial\theta^T}\Big|_{\theta=\hat{\theta}}(\theta - \hat{\theta}) = 0$$

e como o primeiro termo do segundo membro é igual a zero, obtemos

$$\hat{\theta} \approx \theta - \Big[\frac{\partial^2 l(\theta|x)}{\partial\theta\partial\theta^T}\Big|_{\theta=\hat{\theta}}\Big]^{-1} \frac{\partial l(\theta|x)}{\partial \theta}\Big|_{\theta=\hat{\theta}}$$

De modo geral podemos escrever

$$\theta^{(i+1)} \approx \theta^{(i)} - [H(\theta^{(i)})]^{-1} g(\theta^{(i)}),$$

em que $\theta^{(i)}$ é a aproximação do máximo na i-ésima iteração.

A sequência de iterações convirgirá para um ponto de máximo se $H(\hat{\theta}) < 0$, que acontecerá se a função a maximizar for convexa, o que pode não valer em geral. O procedimento não convirgirá se o hessiano calculado no ponto de máximo for singular.

Exemplo 2

Retomemos a função do Exemplo 1 e iniciemos as iterações com $\theta^{(0)} = 1,5$, o que implica $g(\theta^{(0)}) = -1$ e $H(\theta^{(0)}) = 2$. Logo usando a equação do método obtemos

$$\theta^{(1)} \approx 1,5 + \frac{1}{2} = 2,$$

e é fácil verificar que nas próximas iterações o valor de $\theta^{(i)}$ é igual a 2, indicando a convergência em uma iteração.

Exemplo 3

Consideremos, agora, a função

$$f(\theta) = \theta^3 - 3\theta^2 + 1,$$

que tem um ponto de máximo na origem e um ponto de mínimo em $\theta = 2$. Nesse caso,

i) $g(\theta) = 3\theta(\theta - 2)$

ii) $H(\theta) = 6(\theta - 1)$

O valor máximo é 1 e o valor mínimo é -3. Inicializemos o algoritmo com $\theta^{(0)} = 1,5$, para determinar o ponto de mínimo. Então, $g(1,5) = -2,25$ e $H(1,5) = 3$, de modo que na primeira iteração,

$$\theta^1 \approx 1,5 + \frac{2,25}{3} = 2,25,$$

continuando as iterações, obtemos $\theta^{(2)} = 2,025$, $\theta^{(3)} = 2,0003$, indicando a convergência para 2.

Se começarmos com $\theta^0 = 0,5$, na primeira iteração obtemos $\theta^{(1)} = -0,25$, mostrando que, como $H(0,5) < 0$, a primeira iteração direciona o estimador para o ponto de máximo.

Exercício

Desenvolva um programa que calcule as iterações do Exemplo 3.