Importance Sampling

Aishameriane Schmidt

23 de abril de 2017

- Capítulo 7 de 1 e seção 6.2.2 de 2

Métodos de Monte Carlo

Ideias básicas

- Métodos de Monte Carlo são uma alternativa para resolução de integrais (especialmente em casos multivariados onde a dimensão do problema torna os algoritmos não estocásticos muito lentos);
- Uma vez que a abordagem bayesiana requer o cálculo de distribuições a posteriori que muitas vezes envolve a resolução de integrais, os algorítmos de MC acabam sendo muito úteis neste contexto;
- MC é baseado na ideia de reamostrar valores de uma distribuição de probabilidade (simulação estocástica). Utilizando um gerador de números pseudo-aleatórios podemos obter valores de qualquer distribuição (através da $F^{-1}(\cdot)$)

Método

Considere a seguinte integral:

$$\int g(\theta)h(\theta|x)d\theta = \mathbb{E}[g(\theta)|x] \tag{01}$$

Podemos ainda nos utilizar da probabilidade condicional $f(x|\theta) = \frac{f_{X,\theta}(x,\theta)}{\pi(\theta)} \Rightarrow f_{X,\theta}(x,\theta) = f(x|\theta)\pi(theta)$ para reescrever $h(\theta|x) = \frac{f(x|\theta)\pi(theta)}{f_X(x)}$. Como o denominador é uma constante, podemos simplesmente definir o problema da integral acima da seguinte maneira:

$$\int_\Theta g(\theta)f(x|\theta)\pi(\theta)d\theta \tag{02}$$

A primeira forma é como está definido o problema em 1 (página 286) e a segunda é como está em 2 (página 294).

(Murteira) Se pudermos simular uma amostra $\theta_1, \ldots, \theta_n$ da densidade *a posteriori* $h(\theta|x)$, o método de MC irá aproximar **??** por uma média amostral:

$$\hat{\mathbb{E}}[g(\theta)|x] = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \tag{03}$$

Utilizando a lei dos grandes números, pode-se demonstrar que (03) converge quase certamente para a média $\mathbb{E}[g(\theta)|x]$ dada em (01). O método nos diz que se conseguirmos amostras da distribuição *a posteriori* $h(\theta|x)$, podemos resolver as integrais da forma descrita em (01).

(Robert) Se for possível obter valores $\theta_1, \ldots, \theta_n$ da distribuição $\pi(\theta)$, então a média amostral

$$\frac{1}{n}\sum_{i=1}^{m} g(\theta_i) f(x|\theta_i) \tag{04}$$

converge quase certamente para a média dada em (02) quando $m \to \infty$, pela lei dos grandes números. De maneira similar, se uma amostra aleatória de θ_i's da distribuição $\pi(\theta|x)$ pode ser obtida, então

$$\frac{1}{n}\sum_{i=1}^{m} g(\theta_i) \tag{05}$$

converge para

$$\frac{\int_{\Theta} g(\theta) f(x|\theta)\pi(\theta) d\theta}{\int_{\Theta} f(x|\theta)\pi(\theta) d\theta} \tag{06}$$

Amostragem por importância

Muitas vezes não é possível obter uma amostra aleatória de $h(\theta|x)$. O método de MC é flexível o suficiente para ser aplicado de formas alternativas, como por exemplo, simular de uma distribuição similar à posteriori.

(1) Considere $p(x)$ uma função de densidade que seja fácil de simular valores e que aproxima $h(\theta|x) = cf(x|\theta)h(\theta)$. Então:

$$\int g(\theta)h(\theta|x)d\theta = \frac{\int g(\theta)f(x|\theta)h(\theta)d\theta}{\int f(x|\theta)h(\theta)d\theta} \tag{1}$$

$$= \frac{\int g(\theta)\frac{f(x|\theta)h(\theta)}{p(\theta)}p(\theta)d\theta}{\int \frac{f(x|\theta)h(\theta)}{p(\theta)}p(\theta)d\theta} \tag{2}$$

$$= \frac{\int g(\theta)\omega(\theta)p(\theta)d\theta}{\int \omega(\theta)p(\theta)d\theta} \tag{3}$$

Caso tenhamos uma a.a. $\theta_1, \ldots, \theta_n$ de $p(\theta)$, podemos usar MC como em (03), de forma a obter uma aproximação para (01):

$$\hat{\mathbb{E}}[g(\theta)|x] = \frac{1}{\sum_{i=1}^{n}\omega_i}\sum_{i=1}^{n}\omega_i g(\theta_i) \tag{07}$$

onde $\omega_i = \frac{f(x|\theta_i)h(\theta_i)}{p(\theta_i)}$ e é chamado de *importance weights*.

Observe que o método atribui maior "peso" a regiões onde $p(\theta) < h(\theta|x)$ e menos peso onde $p(\theta) > h(\theta|x)$. É possível mostrar que (07) converge quase certamente para (01).

(2) Os métodos de MC tem aplicação muito mais geral que a descrita inicialmente, de forma que não é necessário amostrar da distribuição $\pi(\theta|x)$ ou de $\pi(\theta)$ para ter uma boa aproximação de (02). Se m é uma densidade de probabilidade com suporte $supp(m)$ que tem áreas em comum com o suporte de $g(\theta)f(x|\theta)\pi(\theta)$, então a integral em (02) pode ser escrita como uma esperança em termos de m:

$$\int \frac{g(\theta)f(x|\theta)\pi(\theta)}{m(\theta)}m(\theta)d\theta \tag{08}$$

Que nos leva à representação do método de Monte Carlo *com amostragem por importância*: geramos uma a.a. $\theta_1, \ldots, \theta_n$ de valores de m e aproximamos (02) por:

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)\omega(\theta_i) \tag{09}$$

onde $\omega(\theta_i) = \frac{f(x|\theta_i)\pi(\theta_i)}{m(\theta_i)}$. Pela L.G.N. essa quantidade converge quase certamente para (02). Uma aproximação para $\mathbb{E}^{\pi}[g(\theta)|x]$ é dada por:

$$\frac{\sum_{i=1}^{n} g(\theta_i)\omega(\theta_i)}{\sum_{i=1}^{n} \omega(\theta_i)} \tag{10}$$

Uma vez que o numerador de (10) converge para $\int_{\Theta} g(\theta)f(x|\theta)\pi(\theta)$ e o denominador para $\int_{\Theta} f(x|\theta)\pi(\theta)d\theta$, se $supp(m)$ tem interseção com $supp(f(x|\cdot)\pi)$. Note que (10) não depende de constantes normalizadoras em nenhum dos termos, o que indica que podemos utilizar o método mesmo quando temos apenas o núcleo das distribuições.

Embora (10) deva convergir para $\mathbb{E}^{\pi}[g(\theta)|x]$ para todas as funções que satisfazem a condição do suporte comum, a escolha a função de importância é crucial pelas seguintes razões:

- Primeiro, é necessário que a simulação dos valores de m seja de fácil implementação;
- $m(\theta)$ precisa ser próximo o suficiente de $g(\theta)\pi(\theta|x)$ de maneira a reduzir a variabilidade de (10) tanto quanto possível. Se isso não ocorre, os pesos $\omega(\theta_i)$ serão demasiadamente pequenos e poucas observações serão de fato relevantes. Além disso, corremos o risco de $\mathbb{E}^{m}[g^2(\theta)\omega^2(\theta)]$ não ser finito e a variância do estimador em (10) não estaria definida.

(3) Amostragem por importância é um método de monte carlo que visa reduzir a variância (das estimativas) ao amostrar de uma densidade mais apropriada do que a f.d.p. original.

Suponha que você deseje calcular a esperança de uma $g(\theta)$ cuja densidade é dada por $p(\theta|y)$, porém esta expressão é desconhecida ou é difícil obter amostras de seus valores. Podemos então utilizar um truque matemático combinado com a ideia de MC para obter uma aproximação para esta esperança:

$$I = \int_{\Theta} g(\theta)p(\theta|y)d\theta = \int_{\Theta} \frac{g(\theta)p(\theta|y)}{m(\theta)}m(\theta)d\theta = \mathbb{E}_m\left[\frac{g(\theta)p(\theta|y)}{m(\theta)}\right] \tag{11}$$

E então aproximamos a expressão em (11) por:

$$I \approx \hat{I}_s(\theta) = \frac{1}{S}\sum_{i=1}^{S} \frac{g(\theta^i)f(\theta^i)}{m(\theta^i)} \tag{12}$$

A variância do estimador em (12) será dada por:

$$Var[\hat{I}_s(\theta)] = \mathbb{E}_m[\hat{I}_s{}^2] - \mathbb{E}_m[\hat{I}_s]^2 = \int g^2(\theta)\frac{f^2(\theta)}{m(\theta)}d\theta - I^2 \tag{13}$$

Então, para que a variância de $\hat{I}_S$ seja finita, precisamos que $\int g^2(\theta)\frac{f^2(\theta)}{m(\theta)}d\theta < \infty$, isto é, a densidade $m(\cdot)$ deve ter caudas mais pesadas que $f(\cdot)$. Um estimador por importância considerado bom será aquele que minimiza a quantidade em (12). Isto ocorre quando $m(\cdot)$ se assemelha ao comportamento do produto $g(\cdot)f(\cdot)$.

(4) Como existem diversos estimadores de Monte Carlo, se torna um problema saber decidir qual das estimativas é a melhor. O critério para esta decisão será com base na variância do estimador.

De acordo com 4, a redução da variância pode ser vista como uma forma de utilizar conhecimento prévio sobre o problema. Em um extremo, quando não se sabe nada a respeito das densidades envolvidas, não é possível reduzir a variabilidade. Por outro lado, se temos total conhecimento do problema, a variância é zero e métodos de MC não seriam necessários. Em suas palavras: *"Variance reduction cannot be obtained from nothing; it is merely a way of not wasting information".

(5) Exemplo motivador:

Queremos estimar a probabilidade de que uma variável aleatória X, com distribuição de Cauchy de parâmetros (0,1), seja maior do que 2. Isto é, para $X \sim \mathcal{C}(0,1)$, queremos calcular $\mathbb{P}(X \geq 2)$:

$$p = \mathbb{P}(X \geq 2) = \int_2^\infty \frac{1}{\pi(1+x^2)} dx \tag{14}$$

Imagine que os valores em (14) não sejam de fácil obtenção. Podemos utilizar as ideias de cadeias de markov e, para uma amostra aleatória $X_1, \cdots, X_m$ da distribuição de X, aproximar p por:

$$p \approx \hat{p}_1 = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{X_j > 2} \tag{15}$$

Observe que a quantidade acima é uma v.a. com distribuição de Bernoulli.

(6)

Exemplo

(Exemlo adaptado de 5.2 e 7.1 de 1)

Segundo um modelo genético, pokémons de uma determinada região estão distribuídos em 4 categorias, de acordo com as seguintes probabilidades:

$$p_1 = \frac{2+\theta}{4} \quad p_2 = \frac{1-\theta}{4} \quad p_3 = \frac{1-\theta}{4} \quad p_4 = \frac{\theta}{4} \tag{4}$$

onde $0 \leq \theta \leq 1$ é um parâmetro desconhecido que desejamos fazer inferências a respeito. Suponha que sua priori é $\theta \sim Beta(a,b)$ e que para uma amostra de tamanho N se observaram y_i pokémons do i-ésimo tipo ($i \in \{1,2,3,4\}$ e $\sum_i y_i = N$). Nessas condições, a distribuição a posteriori de θ é:

$$h(\theta|y) \propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3+b-1}\theta^{y_4+1-1}, \quad 0 \leq \theta \leq 1 \tag{12}$$

E

$$L(\theta|y) = logh(\theta|y) \propto (y_1)log(2+\theta) + (y_2+y_3+b-1)log(1-\theta) + (y_4+1-1)log(\theta) \tag{5}$$

$$L'(\theta) = \frac{y_1}{2+\theta} - \frac{y_2+y_3+b-1}{1-\theta} + \frac{y_4+1-1}{\theta} \tag{6}$$

$$-L''(\theta) = \frac{y_1}{(2+\theta)^2} + \frac{y_2+y_3+b-1}{(1-\theta)^2} + \frac{y_4+1-1}{(\theta)^2} \tag{7}$$

Uma função de importância bastante utilizada é a densidade da Normal, já que o que se pretende simular deve ser similar à distribuição a posteriori, porém como nem sempre isso é adequado, vamos tentar achar uma

função de importância. A representação gráfica da verossimilhança pode ajudar na seleção da função, neste caso, uma vez que $\theta \in [0,1]$, podemos buscar uma função $Beta$ como candidata à função de importância.

Vamos comparar a função de importância com distribuição normal e com distribuição beta para duas amostras de tamanho N. Usaremos $p_N(\theta)$ para a função de importância normal e $p_B(\theta)$ para a função de importância beta.

Seja $\hat{\theta}$ o valor de θ para o qual $L'(\theta) = 0$ e $\hat{\sigma}^2 = \{-L''(\hat{\theta})\}^{-1}$. Vamos considerar esses valores como aproximações para a média e variância *a posteriori*, eles serão necessários para obter os parâmetros das distribuições a serem simuladas. O algorítmo então terá os seguintes passos:

1. Simulamos $\theta_1, \ldots, \theta_m \overbrace{\sim}^{iid} p(\theta)$;
2. Calculamos $\omega_i = \frac{h(\theta_i|y)}{p(\theta_i)}$
3. Calculamos $\frac{1}{\sum_{i=1}^m \omega_i} \sum_{i=1}^m \omega_i g(\theta_i)$ com

- $g(\theta) = \theta$ para o cálculo aproximado a média a posteriori
- $g(\theta) = \theta^2$ para a aproximação da variância a posteriori

Com o procedimento acima, basta conhecer o núcleo da distribuição a posteriori, isto é, basta conhecer $h(\theta|Y)$ a menos da constante de proporcionalidade. Também podemos obter uma aproximação boa para a densidade a posteriori atribuindo pesos $\omega_i / \sum_{j=1}^m \omega_j$ aos valores simulados θ_i.

Referências