Joao Felipe Guedes da Silva

guedes.joaofelipe@poli.ufrj.br

CPS767 - Algoritmo de Monte Carlo e Cadeias de Markov
Lista 2

2020/06/16

1 *Cauda do dado em ação*: Considere um icosaedro (um sólido Platônico de 20 faces) honesto, tal que a probabilidade associada a cada face é 1/20. Considere que o dado será lançado até que um número primo seja obtido, e seja Z a variável aleatória que denota este número. Responda às perguntas abaixo:

1.1 Determine a distribuição de Z, ou seja $P[Z = k],\ k = 1, 2, \ldots$. Que distribuição é esta?

Considere Y uma variável aleatória indicadora da primalidade da face. Ou seja, $Y = 1$ se $x \in \{2, 3, 5, 7, 11, 13, 17, 19\}$ e $Y = 0$ caso contrário.
Como $P[X = i] = 1/20$, temos que

$$P[Y = 1] = \frac{8}{20} = \frac{2}{5}$$

Esta variável pode ser modelada como uma Bernoulli com probabilidade $p = 2/5$. Neste contexto, a variável aleatória Z, que denota a quantidade de vezes que precisamos realizar a Bernoulli até que a saída seja 1, possui uma distribuição geométrica com igual probabilidade. Portanto:

$$Z \approx Geom\Big(p = \frac{2}{5}\Big)$$

1.2 Utilize a desigualdade de Markov para calcular um limitante para $P[Z \geq 10]$.

A desigualdade de Markov é utilizada para calcular um limitante superior da probabilidade de uma variável aleatória. Esta desigualdade é dada por:

$$P[Z \geq a] \leq \frac{E[Z]}{a}$$

Ou seja, precisamos apenas do valor esperado $E[Z]$ para calcular a desigualdade de Markov.
Como Z segue uma distribuição geométrica, seu valor esperado é dado por:

$$E[Z] = \frac{1}{p} = \frac{5}{2}$$

Assim,

$$P[Z \geq 10] \leq \frac{5/2}{10} = \frac{1}{4}$$

1.3 Utilize a desigualdade de Chebyshev para calcular um limitante para $P[Z \geq 10]$.

Diferente da desigualdade de Markov, a desigualdade de Chebyshev utiliza a variância da variável aleatória para calcular um limitante superior para a probabilidade aplicada em cima da variável aleatória. Sua formulação é dada por:

$$P[|Z - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Assim, queremos saber a probabilidade de uma realização da variável aleatória estar além de k desvios padrões da média.
Como vimos anteriormente,

$$\mu = E[Z] = \frac{5}{2} = 2.5$$

Além disso, a variância de uma variável aleatória que segue uma distribuição geométrica é dada por:

$$\sigma^2 = Var[Z] = \frac{1-p}{p^2} = \frac{1-2/5}{\left(2/5\right)^2} \to \sigma = 3.75$$

Para $Z - \mu \geq 0$:

$$\begin{gathered} Z - \mu \geq k\sigma \\ Z \geq \mu + k\sigma \end{gathered}$$

Assim, a probabilidade $P[Z \geq 10]$ é calculada para:

$$\begin{gathered} \mu + k\sigma = 10 \\ 2.5 + k \cdot 3.75 = 10 \\ k = 2 \end{gathered}$$

Portanto:

$$P[Z \geq 10] \leq \frac{1}{k^2} = \frac{1}{4}$$

1.4 Calcule o valor exato de $P[Z \geq 10]$ (dica: use probabilidade complementar). Compare os valores obtidos.

Utilizando a probabilidade complementar, temos:

$$P[Z \geq 10] = 1 - P[Z < 10]$$

Portanto, queremos saber a probabilidade de um número primo sair em até 1 jogadas OU até 2 jogadas OU ... até 9 jogadas. Isto pode ser dado por:

$$\begin{aligned} P[Z < 10] &= P[Z = 1 \cup Z = 2 \cup Z = 9] \\ &= \sum_{i=1}^{9} P[Z = i] \\ &= \sum_{i=1}^{9} (1-p)^{i-1} p \\ &= \sum_{i=1}^{9} (1-0.4)^{i-1} 0.4 \\ &= 0.98992 \end{aligned}$$

Portanto,

$$P[Z \geq 10] = 1 - P[Z < 10] = 0.010$$

Tanto a desigualdade de Markov quanto a de Chebyschev se mantiveram válidas após o cálculo do valor exato da probabilidade. No entanto, o valor exato foi consideravelmente menor do que o limite superior estabelecido pelas desigualdades.

2 *Cauda do dado em ação 2:* Considere um icosaedro (um sólido Platônico de 20 faces) honesto, tal que a probabilidade associada a cada face é 1/20. Considere que o dado será lançado n vezes, e seja Z a variável aleatória que denota o número de vezes que o resultado do dado foi um múltiplo de seis. Responda às perguntas abaixo:

2.1 Determine a distribuição de Z, ou seja $P[Z = k], k = 0, 1, l \ldots, n$. Que distribuição é esta?.

Considere Y uma variável aleatória indicadora da multiplicidade por 6 no número do dado. Ou seja, $Y = 1$ se $x \in \{6, 12, 18\}$ e $Y = 0$ caso contrário.

Como $P[X = i] = 1/20$, temos que:

$$\begin{aligned}P[Y = 1] &= Y[x = 6]P[X = 6] + Y[x = 6]P[X = 6] + Y[x = 6]P[X = 6] \\ &= 1 \cdot \frac{1}{20} + 1 \cdot \frac{1}{20} + 1 \cdot \frac{1}{20} \\ &= \frac{3}{20}\end{aligned}$$

Esta variável pode ser modelada como uma Bernoulli com probabilidade $p = 3/20$. Neste contexto, a variável Z que denota o número de vezes com que $Y = 1$ em n realizações possui uma distribuição binomial:

$$Z \approx Binom\Big(n, p = \frac{3}{20}\Big)$$

2.2 Seja $n = 1000$, utilize a desigualdade de Chebyshev para calcular um limitante para $P[Z > 300]$.

A desigualdade de Chebyshev é dada por:

$$P[|Z - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Ao ignorarmos o módulo, a desigualdade se mantém, mas fica mais ”frouxa”.
Como Z segue uma distribuição binomial:

$$E[Z] = np \to E[Z] = 1000 \cdot \frac{3}{20} = 150$$

$$\sigma^2 = Var[Z] = np(1 - p) = 1000 \cdot \frac{3}{20}\Big(1 - \frac{3}{20}\Big) = 127.5 \to \sigma = 11.29$$

$$P[Z > 300] = P[Z - 150 > 150]$$

Desta forma,

$$k\sigma = 150 \to k = 13.28$$

Finalmente,

$$P[Z > 300] = P[Z - \mu \geq k\sigma] \leq 0.005$$

2.3 Seja n = 1000, utilize a desigualdade de Chernoff para calcular um limitante para $P[Z > 300]$.

A desigualdade de Chernoff é dada por:

$$P[Z \geq (1 + \delta)\mu] \leq \left(\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right)^{\mu}$$

E é utilizada quando a variável Z é composta por uma soma de variáveis aleatórias independentes. Como Z possui distribuição binomial, esta condição é satisfeita.
Conforme visto anteriormente,

$$\mu = E[Z] = 150$$

Assim, para calcularmos $P[Z \geq 300]$:

$$(1 + \delta)\mu = 300$$

$$(1 + \delta) \cdot 150 = 300$$

$$\delta = 1$$

Portanto

$$P[Z \geq (1 + \delta)\mu] \leq \left(\frac{e^{1}}{(1 + 1)^{1+1}}\right)^{150} = 5.921 \times 10^{-23}$$

2.4 Seja $n = 1000$, calcule o valor exato para $P[Z > 300]$. Compare os valores obtidos.

Como Z está definido o conjunto dos naturais de 0 a 1000:

$$
\begin{aligned}
P[Z \leq 300] &= \sum_{i=301}^{1000} P[Z = i] \\
&= \sum_{i=301}^{1000} \binom{n}{i} p^i (1-p)^{n-i} \\
&= 9.90 \times 10^{-34}
\end{aligned}
$$

Como o valor exato é menor do que os limitantes calculados, as desigualdades se mantém para o problema.

2.5 Determine o valor z em função de n tal que $Z \leq z$ whp (with high probability).

Dado um modelo com parâmetro n, um evento $A(n)$ no espaço-amostral deste modelo ocorre *with high probability* quando

$$P[A(n)] \geq 1 - \frac{1}{n^\alpha}$$

para algum $\alpha \geq 1$ constante. Isso é análogo a dizer que

$$P[!A(n)] = 1 - P[A(n)] < \frac{1}{n^\alpha}$$

Considerando z uma constante que faz com que os eventos $Z \leq z$ ocorram $w.h.p.$, então:

$$A(n) : Z \leq z \rightarrow !A(n) : Z > z$$

$$P[Z > z] < \frac{1}{n^\alpha}$$

Combinando com a desigualdade de Chernoff

$$P[Z \geq (1+\delta)\mu] \leq e^{-\delta^2 \mu/3}$$

e aplicando $\alpha = 1$ temos que:

$$(1+\delta)\mu = z \rightarrow \delta = \frac{z}{\mu} - 1$$

$$\frac{1}{n} = e^{-(\frac{z}{\mu}-1)^2 \mu/3}$$

$$-\ln n = -\Big(\frac{z}{\mu} - 1\Big)^2 \mu/3$$

$$\mu \cdot \left(\sqrt{\frac{3\ln n}{\mu}} + 1\right) = z$$

$$z = 150\left(\sqrt{\frac{\ln n}{50}} + 1\right)$$

3 *Pesquisa*: Você leu no jornal que uma pesquisa com 1500 pessoas indicou que 40% dos entrevistados preferem o candidato A enquanto 60% preferem o candidato B. Estime a margem de erro desta pesquisa usando uma confiança de 90%. O que você precisa assumir? (dica: use a lei dos grandes números).

Considere o mapeamento $A \rightarrow 1$ e $B \rightarrow 0$. A resposta de uma pessoa à pesquisa possa ser modelada como uma Bernoulli de parâmetro $p = 40\%$. Isto é, dada uma pessoa Y aleatória, seu voto será computado como a

realização de uma Bernoulli com $p = 40\%$. Desta forma, temos 600 ocorrências de $Y = 1$ e 900 ocorrências de $Y = 0$.
A média amostral M_n da variável aleatória

$$M_n = \frac{600 \cdot 1 + 900 \cdot 0}{1500} = 0.4$$

Considerando que a lei dos grandes números tenha sido atendida,

$$E[Y] = M_n = p \to p = 0.4$$

Como Y é modelada como uma Bernoulli, sua variância é dada por:

$$\sigma^2 = p(1-p) = 0.24$$

De acordo com a desigualdade de Chebyschev,

$$P[|M_n - \mu| < \epsilon] > 1 - \frac{\sigma^2}{\epsilon^2 n} = \beta$$

$$\epsilon = \sqrt{\frac{\sigma^2}{n(1-\beta)}}$$

Assim, para uma confiança $\beta = 90\%$, temos:

$$\epsilon = 0.04$$

4 *Moedas:* Você tem duas moedas: uma honesta e outra enviesada que produz cara com probabilidade 3/4. Uma das duas moedas é escolhida aleatoriamente e lançada n vezes. Seja S_n o número de caras que foram observadas nas n jogadas. Responda às perguntas abaixo:

4.1 A lei dos grandes números pode ser aplicada para prever a fração de caras que será observada?

Considere que a moeda possa ser modelada como uma Bernoulli de parâmetro p. Para o caso da moeda honesta (caso 1), $p_1 = 0.5$. No segundo caso, com a moeda enviesada, $p_2 = 3/4$.
Considerando que os eventos são i.i.d, no limite, quando $n\infty$, a probabilidade da média amostral da fração de vezes que apareceu cara ser igual ao valor esperado μ da variável aleatória é 1. Em outras palavras:

$$\lim_{n\to\infty} P[M_n = \mu] = 1$$

Como $\mu = p$ para uma Bernoulli, então a lei dos grandes números pode ser aplicada para estimar o parâmetro p. Assim,

$$p = \lim_{n\to\infty} \frac{S_n}{n}$$

4.2 Podemos determinar qual moeda foi escolhida, depois da mesma ser lançada um número n grande?

Tendo-se estimado p pelo passo anterior, então a moeda honesta terá sido jogada se $p = 0.5$. Caso contrário, se $p = 3/4$, a moeda enviesada terá sido escolhida.

4.3 Determine o valor de n tal que tenhamos 95% de chance de acertar qual moeda foi escolhida.

Suponha que estimemos p com uma margem de erro $\epsilon = 0.01$, e queremos uma confiança $\beta = 0.95\%$ na estimativa de p. Caso calculemos a probabilidade de $p = 0.5$ neste cenário, estaremos calculando a probabilidade da moeda 1 ter sido escolhida com uma confiança de 95%.
Portanto, pela desigualdade de Chebyschev:

$$\beta = 1 - \frac{\sigma^2}{\epsilon^2 n}$$

$$n = \frac{\sigma^2}{\epsilon^2(1-\beta)} = \frac{p(1-p)}{\epsilon^2(1-\beta)} = \frac{0.5(1-0.5)}{0.01^2(1-0.95)} = 50k$$

5 *Sanduíches*: Você convidou 64 pessoas para uma festa e agora precisa preparar sanduíches para os convidados. Você acredita que cada convidado irá comer 0, 1 ou 2 sanduíches com probabilidades 1/4, 1/2 e 1/4, respectivamente. Assuma que o número de sanduíches que cada convidado irá comer é independente de qualquer outro convidado. Quantos sanduíches você deve preparar para ter uma confiança de 95% de que não vai faltar sanduíches para os convidados?

Seja X uma variável aleatória que mapeia o número de sanduíches consumido por uma pessoa. Pela definição do valor esperado, temos:

$$\begin{aligned} E[X] &= \sum_{i=0}^{2} x \cdot P[x = i] \\ &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} \\ &= 1 \end{aligned}$$

Portanto, na média, uma pessoa come 1 sanduíche. Analogamente, podemos calcular a variância desta variável aleatória:

$$\begin{aligned} \sigma^2 &= E[(X-\mu)^2] \\ &= \sum_{i=0}^{2} (x-\mu)^2 \cdot P[x=i] \\ &= (0-1)^2 \cdot \frac{1}{4} + (1-1)^2 \cdot \frac{1}{2} + (2-1)^2 \cdot \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

De acordo com a desigualdade de Chebyschev,

$$P[|X-\mu| < \epsilon] > 1 - \frac{\sigma^2}{\epsilon^2 n} = \beta$$

$$n = \frac{\sigma^2}{\epsilon^2(1-\beta)}$$

Assim, para uma confiança $\beta = 90\%$ e considerando uma precisão 0.01, precisaríamos do consumo de:

$$n = \frac{0.5}{0.01^2(1-0.9)} = 50k$$

6 *Vértices Isolados:* Considere o modelo de grafo aleatório de Erdos-Renyi (também conhecido por $G(n,p)$), onde cada possível aresta de um grafo rotulado com n vértices ocorre com probabilidade p, independentemente. Responda às perguntas abaixo:

6.1 Determine a distribuição do grau do vértice 1 (em função de n e p)?

Dado um vértice i no grafo, para cada um dos $n-1$ vértices j restantes, a probabilidade de existir aresta é p. Desta forma, a aresta a_{ij} pode ser modelada como a realização de uma variável aleatória indicadora que segue uma distribuição de Bernoulli com probabilidade p.

$$a_{ij} \sim Bernoulli(p)$$

Desta forma, o grau do vértice i será dado pela soma das possíveis $n-1$ arestas que ele pode possuir:

$$d_i = \sum_{i=1}^{n-1} Bernoulli(p)$$

Portanto, a distribuição de graus dos vértices no modelo $G(n.p)$ segue uma distribuição binomial pois é a soma de $n-1$ variáveis aleatórias indicadoras.

$$d_i \sim Binom(n-1, p)$$

6.2 Determine a probabilidade do vértice 1 não ter arestas incidentes, ou seja, estar isolado.

Uma vez que o grau segue uma distribuição binomial, a probabilidade do grau d_i assumir um valor k é:

$$P[d_i = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Para que o vértice esteja isolado, seu grau deve ser nulo, o que ocorre com probabilidade:

$$P[d_i = 0] = \binom{n-1}{0} p^0 (1-p)^{n-1-0} = (1-p)^{n-1}$$

Outra forma de visualizar este cálculo é: para cada uma das realizações das arestas que levam a i, a variável aleatória indicadora deve dar 0, o que ocorre com probabilidade $1-p$. Desta forma, a probabilidade de um vértice estar isolado no grafo gerado pelo modelo $G(n,p)$ é:

$$P[d_i = 0] = (1-p)^{n-1}$$

6.3 Determine o valor esperado do número de vértices isolados no grafo (dica: use v.a. indicadora).

Seja I uma variável aleatória que indica o isolamento do vértice. Assim, $I = 1$ quando não há arestas incidentes no vértice e $I = 0$ caso contrário.
Como vimos no exercício anterior, a probabilidade de um vértice estar isolado no grafo gerado pelo modelo $G(n,p)$ é dada por $P[d_i = 0]$. Portanto, A variável aleatória da indicadora pode ser modelada como uma Bernoulli de parâmetro $P[d_i = 0]$.

$$I \sim Bernoulli(P[d_i = 0]) = Bernoulli((1-p)^{n-1})$$

Assim, o número de vértices isolados no grafo é dado por:

$$N = \sum_{j=1}^{n} I = \sum_{j=1}^{n} = Bernoulli((1-p)^{n-1})$$

Como este número é uma soma de variáveis aleatórias indicadoras, sua distribuição é Binomial, cujo valor esperado é dado por $nP[d_i = 0]$. Portanto,

$$E[N] = n(1-p)^{n-1}$$

6.4 Mostre que se $p = (1+\epsilon)\frac{\log n}{n}$, para qualquer $\epsilon > 0$, o modelo G(n, p) não possui vértices isolados, whp. (dica: use método do primeiro momento).

Para $p = (1+\epsilon)\frac{\log n}{n}$ temos:

$$\begin{aligned} E[N] &= n(1-p)^{n-1} \\ &= n\Big(1-(1+\epsilon)\frac{\log n}{n}\Big)^{n-1} \end{aligned}$$

Utilizando a fórmula

$$e^x = \lim_{n\to\infty}\Big(1+\frac{x}{n}\Big)^n$$

Temos que

$$\begin{aligned} \lim_{n\to\infty} E[N] &= \lim_{n\to\infty} n\Big(1-(1+\epsilon)\frac{\log n}{n}\Big)^{n-1} \\ &= \lim_{n\to\infty} ne^{-(1+\epsilon)\log n} \\ &= 0 \end{aligned}$$

Pelo método do primeiro momento,se $\lim_{n\to\infty} E[N] = 0$ então $\lim_{n\to\infty} P[N > 0] = 0$. Portanto, para o valor de p em questão, o modelo $G(n,p)$ não possui vértices isolados w.h.p.