Методы моделирования ДСВ

Руслан Назирович Мокаев

Математико-механический факультет,
Санкт-Петербургский государственный университет

Санкт-Петербург, 05.12.2023, 12.12.2023

Датчик случайных чисел (дискретный)

Определение: Многократно повторяемый эксперимент, исходы которого имеют заданный набор вероятностей, соответствующий набору вероятностей ДСВ, называют **моделью** этой ДСВ.

Пример: Моделью ξ с $Pr\{\xi = 0\} = \frac{1}{2}$ и $Pr\{\xi = 1\} = \frac{1}{2}$ можно считать бросок монетки (если считать, что стороны монетки одинаковые, монетка никогда не падает на ребро и бросают ее наугад).

Определение: Многократно повторяемый эксперимент, результатом которого является (псевдо)случайно выбранное число из некоторого конечного множества, называют **дискретным генератором (датчиком) случайных чисел**.

Пример: Пусть α дискретный генератор случайных чисел, для которого все исходы имеют одну и ту же вероятность:

$$Pr\{\alpha = k\} = \frac{1}{N} \; \forall k \in 1 : N$$

Датчик случайных чисел (непрерывный)

Определение: Многократно повторяемый эксперимент, результатом которого является (псевдо)случайно выбранное число из некоторого промежутка числовой прямой, называют **непрерывным генератором (датчиком) случайных чисел**.

Пример: Обозначим за α_0 непрерывный генератор случайных чисел, который с равной вероятностью попадает в каждую из точек отрезка $[0,1]$. Поскольку отрезок $[0,1]$ содержит несчетное количество точек, вероятность $Pr\{\alpha_0 = x\} = \frac{1}{\infty} = 0 \ \forall x \in [0,1]$. Однако, вероятность того, что точка попадет в некоторый промежуток, все же ненулевая:

$Pr\{x \in < a, b >\} = \frac{|<a,a+\delta>|}{|[0,1]|} = \frac{\delta}{1} = \delta$. В частности, $Pr\{x \in [0,1]\} = 1$.

Замечание: Поскольку для конкретной точки вероятность попадания туда α_0 равна нулю, можно смело выкидывать из множества конечное (и даже счетное) количество точек, и вероятность попадания в него останется неизменной. Следовательно, вместо $< a, b >$ можно рассматривать (a,b) или $[a,b)$. Последнее особенно удобно, поскольку полуинтервалы хорошо стыкуются.

Табличный метод моделирования ДСВ

Пусть (S, Pr) – вероятностное пространство, ξ – заданная на нем ДСВ.

При этом $S = \{0, 1, \ \dots \ , n-1\}$, $\forall i \in \overline{0:(n-1)}\ Pr(\{i\}) = p_i$, $\sum\limits_{i=0}^{n-1} p_i = 1$.

Примечание: Для ДСВ с произвольным набором значений можно их пронумеровать и свести задачу к данной. Возьмем отрезок $[0, 1)$ и поделим его на полуинтервалы вида $[x_i, x_{i+1})$, где $x_0 = 0$, $\forall i \in \overline{0:(n-1)}\ x_{i+1} = x_i + p_i$. Для такого разбиения отрезка $[0, 1]$ вероятность $Pr\{\alpha_0 \in [x_i, x_{i+1})\} = p_i$. То есть, эксперимент "в отрезок с каким номером попадет случайное число, выданное α_0?" моделирует ξ.

Замечание: При попадании в точку 1 значение α_0 оказывается вне пределов всех полуинтервалов. С одной стороны, вероятность такого события – 0, следовательно на распределение вероятностей это не повлияет. С другой стороны, если это всё же произошло, мы можем просто объявить эксперимент неудачным и провести его заново.

Основным недостатком данного метода является большое количество сравнений вещественных чисел, поскольку эта операция является не очень точной и достаточно долгой.

Неплохо бы придумать метод, который минимизирует количество таких операций!

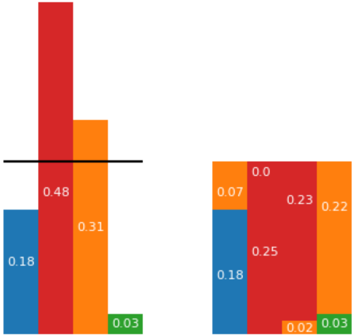
Метод Уокера (Alias-метод)

В табличном методе мы располагали полуинтервалы на отрезке $[0,1]$ в произвольном порядке.

В этот раз мы сначала разделим его на полуинтервалы $[\frac{i}{n}, \frac{i+1}{n})$, где $\forall i \in \overline{0:(n-1)}$ (назовем их **базовыми полуинтервалами**) и будем следить, чтобы в одном таком полуинтервале было **не более одной** границы полуинтервалов, соответствующих исходам.

Более того, если раньше каждому исходу соответствовал ровно **один** полуинтервал, то сейчас их может быть **несколько**, при этом сумма их длин по-прежнему должна быть равна вероятности этого исхода.

Иллюстрация метода Уокера



Специально строим "новое" распределение (приведение к Alias-формату).
Каждый элемент распределения (n штук) состоит из двух частей: $p_m + \bar{p}_{m^*} = \frac{1}{n}$.
Замечание: Если в ходе приведения к Alias-формату вероятность донорского элемента станет меньше $\frac{1}{n}$, то для него будет определен свой донор.
Замечание: Донор допускается только один.

Для корректной работы метода для каждого m надо хранить $p_m n$ (перегородка внутри элемента распределения)
и m^* (номер донора).

А. Н. Залялов, ALIAS-метод для моделирования таблично заданных распределений случайных величин, ВАНТ, сер. Математическое моделирование физических процессов. 2018. Вып. 2

Суть метода Уокера

Утверждение: $\lfloor n \cdot \alpha_0 \rfloor = i$, если $\frac{i}{n} \leq \alpha_0 < \frac{i+1}{n}$.
Доказательство: Умножим все части неравенства на n:

$$i \leq n \cdot \alpha_0 < i + 1 \Leftrightarrow \lfloor n \cdot \alpha_0 \rfloor = i$$

$\square$

Итак, мы научились понимать, в какой базовый полуинтервал попал α_0. Теперь осталось понять, что делать, если в базовом полуинтервале есть граница между полуинтервалами?
Пусть эта граница есть в точке x, $\frac{i}{n} \leq x < \frac{i+1}{n}$. Растянем наш базовый полуинтервал в n раз, тогда точка $\frac{i}{n}$ перейдет в точку $i = \lfloor n \cdot \alpha_0 \rfloor$, α_0 – в $n \cdot \alpha_0$, x – в $n \cdot x$, $\frac{i+1}{n}$ – в $i + 1 = \lfloor n \cdot \alpha_0 \rfloor + 1$.
Теперь сдвинем полуинтервал влево на i: $\lfloor n \cdot \alpha_0 \rfloor$ перейдет в 0, $n \cdot \alpha_0$ – в $\{n \cdot \alpha_0\}$ (дробная часть), $n \cdot x$ – в $n \cdot x - i$, $\lfloor n \cdot \alpha_0 \rfloor + 1$ – в 1. То есть, теперь нам достаточно сравнить $\{n \cdot \alpha_0\}$ и $n \cdot x - i$, причем $n \cdot x - i$ не зависит от α_0 и может быть посчитана заранее.

Замечание: При попадании в единицу все ещё придется повторять опыт.
Итак, пусть каждому значению случайной величины $i \in \overline{0 : (n-1)}$ соответствует множество P_i полуинтервалов, причем каждый из них полностью лежит в одном из базовых полуинтервалов и пересечение P_i и P_j пусто для любого $j \neq i$, а $\sum\limits_{[a,b) \in P_i} \Big|[a,b)\Big| = p_i$. Также, в каждом базовом полуинтервале лежит не более двух полуинтервалов.
Тогда мы научились с помощью двух сравнений понимать, в какой полуинтервал $[a,b)$ попал α_0.
А если объявить исходом эксперимента i, где $[a,b) \in P_i$, то этот эксперимент будет моделировать ξ, так как вероятность попасть в полуинтервал, лежащий в P_i, равна
$\sum\limits_{[a,b) \in P_i} Pr\{\alpha_0 \in [a,b)\} = \sum\limits_{[a,b) \in P_i} \Big|[a,b)\Big| = p_i$.
Теперь осталось построить такое разбиение. Для этого докажем две леммы.

Метод Уокера. Продолжение доказательства

Лемма 1: Если $n > 1$, то $\exists\ l \in \overline{0:(n-1)}$ такое, что $p_l \leq \frac{1}{n}$.

Доказательство: Если предположить обратное, то $\sum\limits_{i=0}^{n-1} p_i > 1$.

Лемма 2: Если $n > 1$, то $\forall l \in \overline{0:(n-1)}\ \exists\ m \in 0:(n-1), m \neq l$ такое, что $p_l + p_m > \frac{1}{n}$.

Доказательство: Если предположить обратное, то $\exists\ l_0 \in \overline{0:(n-1)}$ такое, что: $\sum\limits_{m \neq l_0} (p_{l_0} + p_m) \leq \frac{n-1}{n} < 1$.

С другой стороны,

$$\sum_{m \neq l_0} (p_{l_0} + p_m) = \sum_{m \neq l_0} p_{l_0} + \sum_{m \neq l_0} p_m = (n-2) \cdot p_{l_0} + \sum_{m \in 0:(n-1)} p_m \geq 1$$

Противоречие. $\square$

Метод Уокера. Продолжение доказательства

Построим разбиение:
Определим $\xi^{(0)} : p_i^{(0)} = p_i$, выберем $l_0 :\ p_{l_0}^{(0)} < \frac{1}{n}$, $m_0 :\ p_{l_0}^{(0)} + p_{m_0}^{(0)} > \frac{1}{n}$.
Определим ДСВ $\psi^{(0)}$:
$Pr\{\psi^{(0)} = l_0\} = n \cdot p_{l_0}^{(0)}, Pr\{\psi^{(0)} = m_0\} = 1 - n \cdot p_{l_0}^{(0)}$.
Теперь отрежем от отрезка $[0, 1]$ первый базовый полуинтервал:
$\xi^{(1)} : p_{l_0}^{(1)} = 0, p_{m_0}^{(1)} = \left(p_{m_0}^{(0)} - (\frac{1}{n} - p_{l_0}^{(0)})\right) \cdot \frac{n}{n-1}, p_i^{(1)} = p_i \cdot \frac{n}{n-1}$ и перейдем к следующему шагу алгоритма.
$\forall k \in \overline{0 : (n-1)}\ p_k^{(0)}$ через $\xi^{(1)}$: $p_k^{(0)} = \frac{1}{n} Pr\{\psi^{(0)} = k\} + \frac{n-1}{n} p_k^{(1)}$.
Давайте это аккуратно проверим!

Метод Уокера. Продолжение доказательства

Общий случай: пусть построено $\xi^{(i)}$, $i \geq 1$, у этой ДСВ $n-i$ возможных исходов; $\exists l_i, m_i: \ m_i \neq l_i, p_{l_i}^{(i)} \leq \frac{1}{n-i}, p_{m_i}^{(i)} + p_{l_i}^{(i)} > \frac{1}{n-i}$.

$\psi^{(i)}: \ Pr\{\psi^{(i)} = l_i\} = (n-i) \cdot p_{l_0}^{(i)}, Pr\{\psi^{(i)} = m_i\} = 1 - (n-i) \cdot p_{l_0}^{(i)}$.

Строим $\xi^{(i+1)}$: $p_{l_i}^{(i+1)} = 0, p_{m_i}^{(i+1)} = p_{m_i}^{(i)} - \frac{n-i}{n-i-1}(\frac{1}{n-i} - p_{l_i}^{(i)})$,

$p_s^{(i+1)} = \frac{n-i}{n-i-1} \cdot p_s^{(i)}$.

Тогда $p_k^{(i)} = \frac{1}{n-i} Pr\{\psi^{(i)} = k\} + \frac{n-i-1}{n-i} \cdot p_k^{(i+1)}$.

Так до $n-1$ шага. На $n-1$ шаге получаем, что $\xi^{(n-1)}$ имеет всего один возможный исход: $Pr\{\xi^{(n-1)} = l_{n-1}\} = 1$.

Тогда $p_k^{(n-1)} = \frac{1}{1} Pr\{\psi^{(n-1)} = k\} + 0 = Pr\{\psi^{(n-1)} = k\}$.

Т. о., мы получили рекурсивную формулу для $p_k^{(i)}$. Раскроем ее:

$$p_k = p_k^{(0)} = \frac{1}{n} Pr\{\psi^{(0)} = k\} + \frac{n-1}{n} p_k^{(1)} = \frac{1}{n} Pr\{\psi^{(0)} = k\} +$$

$$+ \frac{n-1}{n} \left(\frac{1}{n-1} Pr\{\psi^{(1)} = k\} + \frac{n-2}{n-1} p_k^{(2)} \right) = \frac{1}{n} Pr\{\psi^{(0)} = k\} +$$

$$+ \frac{1}{n} Pr\{\psi^{(1)} = k\} + \frac{n-2}{n} p_k^{(2)} = \ldots = \frac{1}{n} \sum_{j \in 0:(n-1)} Pr\{\psi^{(j)} = k\}$$

Пример

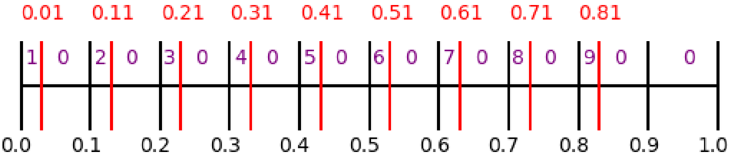
$p_0 = 0.91, p_i = 0.01,\ \forall i \in \overline{1:9}$

Первый шаг: $l_0 = 1, m_0 = 0,\ p_0^{(0)} = p_0 = 0.91,\ p_i^{(0)} = p_i = 0.01$ для $i \in \overline{1:9}$.

Тогда $p_1^{(1)} = 0, p_0^{(1)} = \frac{10}{9}\left(p_0^{(0)} - (\frac{1}{10} - p_1^{(0)})\right) = \frac{10}{9}(p_0 - \frac{1}{10} + p_1) =$

$= \frac{10}{9}(0.91 - 0.1 + 0.01) = \frac{10}{9} \cdot 0.82,\ p_i^{(1)} = \frac{10}{9} \cdot p_i^{(0)} = \frac{0.1}{9}$

И так далее, на последнем шаге получим $p_0^{(9)} = 1, p_i^{(9)} = 0.$



Вычислительная схема метода Уокера

Замечание: Постоянное домножение на коэффициенты в методе Уокера нужно только чтобы на каждом шаге $\xi^{(i)}$ оставалась ДСВ, что необходимо для формального доказательства, но не нужно для практических вычислений.

$\xi=0$	$\xi=1$	$\xi=2$	$\xi=3$	$\xi=4$	l_i	m_i	$n\cdot p^i_{l_i}$	$p^{i+1}_{m_i}$	$\psi=0$	$\psi=1$	$\psi=2$	$\psi=3$	$\psi=4$	i
0.02	0.41	0.21	0.17	0.19	0	1	0.1	0.23	0.1	0.9	0	0	0	0
0	0.23	0.21	0.17	0.19	3	1	0.85	0.2	0	0.15	0	0.85	0	1
0	0.2	0.21	0	0.19	4	1	0.95	0.19	0	0.05	0	0	0.95	2
0	0.19	0.21	0	0	1	2	0.95	0.2	0	0.95	0.05	0	0	3
0	0	0.2	0	0	2	–	–	–	0	0	1	0	0	4

Набор вероятностей для $\xi^{(i)}$; l_i, m_i – база и донор; перегородка $n\cdot p^{(i)}_{l_i}$; $p^{(i+1)}_{m_i} = p^{(i)}_{m_i} - \frac{1}{n} + p^{(i)}_{l_i}$ – "остаток" донорской вероятности (без домножения на коэффициент), необходимо для пересчёта $\xi^{(i+1)}$; набор вероятностей $\psi_{(i)}$; i – номер текущего шага (и номер базового полуинтервала).

Пример: $\alpha_0 = 0.73$. $n\cdot\alpha_0 = 5\cdot 0.73 = 3.65 \Rightarrow \lfloor n\cdot\alpha_0\rfloor = 3$. Граница в 3-м базовом полуинтервале в $n\cdot p^{(3)}_{l_3} = 0.95$.

$\{n\cdot\alpha_0\} = 0.65 < 0.95 \Rightarrow$ результат эксперимента это $l_i = l_3 = 1$.

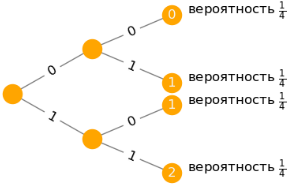
Моделирование ДСВ с помощью последовательности (псевдо)случайных бит

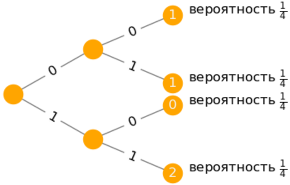
Определение: Псевдослучайный бит – 0 или 1 равновероятно (бросок монетки). Хотим научиться моделировать ДСВ, вероятности которой – рациональные двоичные числа, с помощью последовательности случайных бит.

Пример:

ξ: $Pr\{\xi = 0\} = \frac{1}{4} = 0.01_2, Pr\{\xi = 1\} = \frac{1}{2} = 0.1_2, Pr\{\xi = 2\} = \frac{1}{4} = 0.01_2$

Интуиция: построить полное двоичное дерево нужной глубины $h =$ максимальному количеству значащих двоичных цифр после запятой среди p_i и **каким-то** образом распределить листья по исходам так, чтобы количество листьев, соответствующих i-му исходу, было равно $2^h \cdot p_i$.





1-й случай – бросаем монетку дважды, 2-й – можно обойтись одним разом.

Метод формально

Пусть есть ДСВ ξ, $\forall i \in \overline{0:(n-1)}$ $Pr\{\xi = i\} = p_i$, $\sum_{i \in \overline{0:(n-1)}} p_i = 1$

p_i в двоичной записи имеет представление $0.\pi_1^i \pi_2^i \ldots \pi_m^i$, т.е.

$p_i = \frac{\pi_1^i}{2} + \frac{\pi_2^i}{4} + \ldots + \frac{\pi_m^i}{2^m}$, где $\pi_k^i \in \{0, 1\} \Rightarrow 2^m p_i = 2^{m-1}\pi_1^i + \ldots + \pi_m^i$.

Рассмотрим множество A исходов m бросков монетки; $|A| = 2^m$.

Разбиение $A = A_0 \cup A_1 \cup \ldots \cup A_{n-1}$, где $\forall i \in \overline{0:(n-1)}$ $|A_i| = 2^m p_i$

Тогда, если исход эксперимента "m бросков монетки" лежит в A_i, то регистрируем исход i. $Pr\Big\{$зарегистрирован исход $i\Big\} = \frac{|A_i|}{2^m} = p_i$.

Как построить оптимальное разбиение?

Заведем набор множеств $I_k = \left\{i \in 0:(n-1) \mid \pi_k^i = 1\right\}$ $\forall k \in \overline{1:m}$;

Пусть мы бросили монету 1 раз. Множество исходов $B_1 : |B_1| = 2$.

Выберем какое-то $M_1 \subseteq B_1 :\ |M_1| = |I_1| = \sum_{j \in \overline{0:(n-1)}} \pi_1^j$. Установим биекцию между M_1 и I_1. Во всех этих случаях исход уже определен.

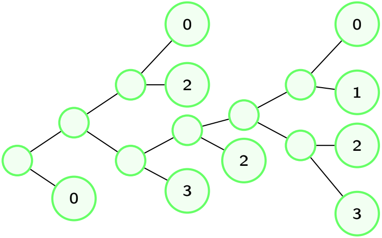
В остальных $B_1 \setminus M_1$ случаях нам нужно продолжать бросать.

Теперь у нас есть B_2, $|B_2| = 2 \cdot |B_1 \setminus M_1|$, используем тот же алгоритм.

Общий случай: B_k – множество исходов после k бросков; выбираем из них множество $M_k \subseteq B_k$ исходов: $|M_k| = |I_k|$, строим биекцию между M_k и I_k, определяя исход эксперимента для этих случаев, а в остальных $B_k \setminus M_k$ случаях продолжаем бросать.

Пример

$p_0 = 0.101001, p_1 = 0.000001, p_2 = 0.001101, p_3 = 0.001001.$
$I_1 = \{0\}, I_2 = \varnothing, I_3 = \{0, 2, 3\}, I_4 = \{2\}, I_5 = \varnothing, I_6 = \{0, 1, 2, 3\},$



$B_1 = \{0, 1\}$, $M_1 \subseteq B_1, |M_1| = |I_1| = 1$, $M_1 = \{1\}$,
биекция рез-ту $\{0\} \leftrightarrow$ бросок $\{1\}$;
$B_2 = \{00, 01\}$, $M_2 \subseteq B_2, |M_2| = |I_2| = 0$
броски не соответствуют никаким результатам;
$B_3 = \{000, 001, 010, 011\}$, $M_3 \subseteq B_3$,
$|M_3| = |I_3| = 3$
биекция $\{0, 2, 3\} \leftrightarrow$ броскам $\{000, 001, 011\}$;
$B_4 = \{0010, 0101\}$ и т.д.

$$E\{\text{число бросков}\} = \sum_{k \in 1:m} k \cdot Pr\{\text{монета брошена } k \text{ раз}\} =$$

$$\sum_{k \in 1:m} k|I_k|\frac{1}{2^k} = \sum_{k \in 1:m} \frac{k}{2^k} \sum_{j \in 0:(n-1)} \pi_k^j$$