Метод Монте-Карло в задачах вычисления малых вероятностей

Логинов Андрей Сергеевич, группа 18.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Отчет по производственной практике

Санкт-Петербург
2021г.

Введение

X — вещественная случайная величина.
Хотим оценивать следущую вероятность:

$$p = \mathbb{P}(X \geq a).$$

При достаточно больших a вероятность $p \to 0$.
Событие $\{X \geq a\}$ будем называть **редким**, а вероятность p — **малой**.
Примеры:

- Задачи молекулярной динамики
- Цифровые водяные знаки
- Задачи, связанные с выравниванием последовательностей в биоинформатике

Постановка задачи

$\mathbb{S}^{(n)}$ — множество строк длины n над алфавитом $\mathfrak{A}$ мощности l.
Рассмотрим отображение $d : \mathbb{S}^{(n)} \times \mathbb{S}^{(n)} \to \mathbb{Z}^+$
$s_0 \in \mathbb{S}^{(n)}$ — фиксированная строка
$s \in \mathbb{S}^{(n)}$ — случайная строка
Представляет интерес событие

$$A = \{d(s_0, s) \geq a\}$$

и его вероятность

$$p = \mathbb{P}(A).$$

Важным частным случаем является расстояние Хэмминга

$$\mathcal{H}(s_0, s) = \sum_{k=1}^{n} [s_0^k = s^k].$$

Можно легко вычислять p и проверять построенные оценки, которые затем можно будет использовать для более сложных отображений d.

Метод Монте-Карло

$(X_i)_{i\in\mathbb{N}}$ — последовательность независимых одинаково распределенных случайных величин.

- Для любого $N \in \mathbb{N}$ определена несмещенная оценка p:

$$\hat{p}_{MC} = \frac{1}{N}\sum_{k=1}^{N}[X_i \geq a].$$

- Дисперсия оценки:

$$\mathbb{D}\hat{p}_{MC} = \frac{p(1-p)}{N}.$$

- Относительная ошибка:

$$\mathrm{RE}(\hat{p}_{MC}) = \frac{\sqrt{\mathbb{D}\hat{p}_N}}{p} = \sqrt{\frac{1-p}{Np}}.$$

Метод Монте-Карло

- Для относительной ошибки ε получаем требуемый порядок объема выборки:
$$N \sim \frac{1-p}{\varepsilon^2 p}.$$
- Например, хотим оценить вероятность
$$p_1 = \mathbb{P}(\mathcal{H}(s_0, s) \geq a) = \sum_{k=a}^{n} C_n^k \left(\frac{1}{l}\right)^k \left(\frac{l-1}{l}\right)^{n-k}.$$
В случае, когда $s_0, s_1 \in \mathbb{S}^{20}$, $l = 4$, получается $p_1 = 3.81 \cdot 10^{-6}$. Чтобы оценить такую вероятность с относительной оишбкой $\varepsilon = 0.01$, потребуется $N \sim 10^9$.
- Актуальна задача уменьшения дисперсии при фиксированном объеме выборки.

Метод существенной выборки

Пусть $\mathcal{G}$, $\mathcal{Q}$ — распределения на $\mathbb{S}^{(n)}$:

- g, q — их плотности,
- $\mathcal{G} \prec \mathcal{Q}$.

Рассмотрим выборку $s_1, \ldots, s_N \sim \mathcal{Q}$ и фиксированную строку $s_0 \in \mathbb{S}^{(n)}$.
Оценка вероятности $p = \mathbb{P}(d(s_0, s) \geq a) = \mathbb{P}(A)$ по методу существенной выборки:

$$\hat{p}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(s_i)}{q(s_i)} \mathbb{1}_A(s_i).$$

Существенная выборка: моделирующее распределение

Свойства оценки по методу существенной выборки зависят от выбора моделирующего распределения $\mathcal{Q}$.

- Для расстояния Хэмминга можно использовать $\mathcal{Q}$ такое, что $\forall s \sim \mathcal{Q}$

$$s^j = \begin{cases} s_0^j \text{ с вероятностью } p^* \\ x \in \mathfrak{A} \setminus \{s_0^j\} \text{ с вероятностью } (1-p^*)/l, \end{cases}$$

где p^* — параметр.
Однако, такую плотность не получится обобщить для для других отображений d.

Существенная выборка: моделирующее распределение

- В общем случае будем использовать распределение с плотностью вида

$$q(s) = \frac{1}{Z}\tilde{q}(s) = \frac{1}{Z}g(s)w(s),$$

где Z — нормализующая константа, $w(s)$ — весовая функция. Будем рассматривать весовые функции вида

$$w(s) = w(d(s_0, s)) = \exp\{\gamma \cdot d(s_0, s)\},$$

где γ — положительный параметр.

Алгоритм Метрополиса-Гастингса

- Для построения $\hat{p}_{IS}$ нужно уметь моделировать некоторое распределение $\mathcal{Q}$, что может быть весьма затруднительно.
- Алгоритм Метрополиса-Гастингса (**Hastings_1970**) позволяет строить цепь Маркова, стационарное распределение которой есть нужное $\mathcal{Q}$.
- Чтобы построить такую марковскую цепь, достаточно знать плотность распределения $\mathcal{Q}$ с точностью до нормализующей константы.

Алгоритм Метрополиса-Гастингса: практические соображения

- Важно только стационарное распределение получаемой цепи Маркова.
  Определять нужную часть траектории можно по стабилизации накопленных средних.
- Выборку получаем с помощью цепи Маркова, поэтому ее элементы зависимы.
  Прореживанием траектории можно добиться снижения автокорреляции.

Нормализующая константа

- С учетом того, что моделирующая плотность имеет вид $q(s) = \frac{1}{Z} g(s) w(s)$, оценка $\hat{p}_{IS}$ принимает вид

$$\hat{p}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_A(s_i) \frac{Z}{w(s_i)}.$$

- Нормализующую константу Z можно оценить:

$$\hat{Z} = \left(\frac{1}{N} \sum_{i=1}^{N} (w(s_i))^{-1} \right)^{-1}.$$

- Тогда можно переписать оценку

$$\hat{p}_{IS} = \frac{\sum_{i=1}^{N} \mathbb{1}_A(s_i)(w(s_i))^{-1}}{\sum_{i=1}^{N} (w(s_i))^{-1}}.$$

Оценка дисперсии

- В случае независимой выборки дисперсию $\hat{p}_{IS}$ можно найти по формуле (**Owen2020**):

$$\mathrm{Var}(\hat{p_{IS}}) = \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{\mathbb{1}_A(s_i) g(s_i)}{q(s_i)} - \hat{p}_{IS} \right)^2.$$

- Дисперсию оценки вдоль траектории марковской цепи можно найти с помощью метода batch means (**Jones2006**):

$$\mathrm{Var}(\hat{p}_{IS}) = \frac{v}{(u-1)N} \sum_{j=1}^{u} (Y_j - \hat{p}_{IS})^2,$$

где $u \cdot v = N$, а Y_j находится по формуле:

$$Y_j = \frac{1}{v} \sum_{i=(j-1)v}^{jv-1} \mathbb{1}_A(x_i) \frac{g(s_i)}{q(s_i)},$$

Построенные оценки

Были построены:

- оценки по методу Монте-Карло $\hat{p}_{MC}$,
- оценки по методу существенной выборки для повторных независимых выборок $\hat{p}_{IS}$,
- оценки по методу существенной выборки, использующие алгоритм Метрополиса-Гастингса, с известной и неизвестной нормализующей константой $\hat{p}_{MHK}$ и $\hat{p}_{MHU}$,
- оценки по методу существенной выборки, использующие алгоритм Метрополиса-Гастингса с предложенными модификациями, с известной и неизвестной нормализующей константой $\hat{p}_{MHMK}$ и $\hat{p}_{MHMU}$,

для событий $\{\mathcal{H}(s_0, s \geq k)\}$, $s_0, s \in \mathbb{S}^{(20)}$, $l = 4$, $k \in \{5, 10, 15, 20\}$, вычислены относительные ошибки.

Замечание

Для оценок $\hat{p}_{MHMK}$ и $\hat{p}_{MHMU}$ под объемом выборки понимается длина цепи до прореживания.

Численные результаты: фиксированный объем выборки

Для оценок, построенных по одинаковому объему выборки $N = 10^5$:

p	0.585158	0.013864	$3.81 \cdot 10^{-6}$	$9.10 \cdot 10^{-13}$
RE_{MC}	0.002663	0.026538	1.337253	-
RE_{IS}	0.004608	0.005153	0.005981	0.004231
RE_{MHK}	0.012064	0.028712	0.031248	0.092181
RE_{MHU}	0.0120	0.038373	1.043263	18.864911
RE_{MHMK}	0.038793	0.086880	0.086781	0.182050
RE_{MHMU}	0.039509	0.103369	2.825464	97.941459

- Относительные ошибки оценок $\hat{p}_{MHK}$ и $\hat{p}_{MHMK}$ при уменьшении p растут намного медленнее оценок $\hat{p}_{MC}$, а относительная ошибка $\hat{p}_{IS}$ остается практически неизменной.
- Быстрый рост RE_{MHU} и RE_{MHMU} при уменьшении p связан с неточностью оценки Z.

Доверительные интервалы

Доверительные интревалы для $\hat{p}_{MC}$ и $\hat{p}_{IS}$.

p	CI_{MC}	CI_{IS}
0.585158	$(0.581380; 0.589419)$	$(0.577726; 0.591639)$
0.013864	$(0.012945; 0.014856)$	$(0.013679; 0.014047)$
$3.81 \cdot 10^{-6}$	$(-1.09 \cdot 10^{-5}; 1.65 \cdot 10{-}5)$	$(3.75 \cdot^{-6}; 3.87 \cdot 10^{-6})$
$9.10 \cdot 10^{-13}$	$(0; 0)$	$(9.00 \cdot 10^{-13}; 9.20 \cdot 10^{-13})$

- Доверительные интервалы оценок по методу существенной выборки содержат p для всех k.
- При увеличении k доверительные интрвалы оценок по методу Монте-Карло перестают накрывать истинное значение p.

Убывание относительной ошибки

Сравнение скорости убывания относительной ошибки оценок $\hat{p}_{MC}$, $\hat{p}_{IS}$, $\hat{p}_{MHK}$, $\hat{p}_{MHU}$ при увеличении объема выборки для вероятности события $\mathcal{H}(s_0, s) \geq 12$, $s_0, s \in \mathbb{S}^{(20)}$, $l = 4$.

Заключение

Что сделано:

- Были построены оценки по методу существенной выборки с предложенным простым для моделирования распределения и с применением алгоритма Метрополиса-Гастингса.
- Проведено сравнение относительных ошибок при фиксированном объеме выборки и разных вероятностях с оценками по методу Монте-Карло.
- Проведено сравнение скоростей убывания относительной ошибки с ростом объема выборки.

Что нужно сделать:

- Построить доверительные интервалы для полученных оценок.
- Научиться более точно оценивать нормализующую константу Z вдоль траектории марковской цепи.