Применение цепей Маркова в анализе данных

Саттаров Никита Дмитриевич, группа 19.Б04

Санкт-Петербургский Государственный Университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — к.ф.-м.н. Н.Э. Голяндина

Санкт-Петербург
2021г.

Введение

Имеем таблицу данных:

Номер наблюдения	Начальное состояние	Конечное состояние
1	s_{start_1}	s_{finish_1}
2	s_{start_2}	s_{finish_2}
3	s_{start_3}	s_{finish_3}
...	...	...

$\forall i: \ s_{start_i}, s_{finish_i} \in S, \ |S| < +\infty$

Хотим извлечь из таблицы информацию: вероятность попасть из s_i в s_j, среднее количество шагов, чтобы попасть из s_i в s_j.

Постановка задачи

Последовательность дискретных случайных величин $\{X_n\}_{n\geq 0}$ называется цепью Маркова, если

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n)$$

Хотим проверить корректность построения цепи Маркова по таблице данных.

- Построим цепь Маркова, которая будет соответствовать вероятностям попасть из состояния s_i в состояние s_j для нашей таблицы данных
- Сгенерируем новые наблюдения на основе вероятностей из цепи Маркова
- Построим новую цепь Маркова по сгенерированным наблюдениям
- Сверим две получившиеся цепи

Обыкновенный метод Монте-Карло

$(S_i)_{i \in \mathbb{N}}$ — последовательность случайных строк из $\mathbb{S}^{(n)}$

- Для любого N имеем несмещённую оценку p:

$$\hat{p}_N = \frac{\#\{S_i : d(S_0, S_i) > a\}}{N}$$

- Дисперсия оценки:

$$\mathbb{D}\hat{p}_N = \frac{p(1-p)}{N}$$

- Относительная ошибка:

$$\frac{\sqrt{\mathbb{D}\hat{p}_N}}{p} = \sqrt{\frac{1-p}{Np}}$$

Обыкновенный метод Монте-Карло

- Для относительной ошибки ε получаем требуемый порядок объема выборки:
$$N \sim \frac{1-p}{\varepsilon^2 p}$$
- Тогда, напрмер, для оценивания некоторой вероятности $p_0 \sim 10^{-10}$ с относительной ошибкой $\varepsilon = 0.01$ имеем необходимый объем выборки $N \sim 10^{17}$.
- Актуальна задача уменьшения дисперсии при фиксированном объеме выборки.

Адаптивное многоуровневое расщепление (Brehier2014)

- Введем J промежуточных уровней: $a_0 = 0 < a_1 < \ldots < a_J = a$ и соответствующие им вероятности:

$$p_j = \mathbb{P}(X > a_j | X > a_{j-1})$$

- Малая вероятность p должна удовлетворять равенству:

$$p = \prod_{i=1}^{J} p_i$$

- Дисперсия будет минимизирована, если $p_1 = \ldots = p_J = p^{1/J}$, поэтому промежуточные уровни определяются так, чтобы выполнялось равенство множителей p_j.

AMS

При попытке применить алгоритм возникли следующие трудности:

- Необходимо получать условные распределения $\mathcal{L}(X|X > a_i)$.
- Алгоритм определен и доказан для с.в. $X \in \mathbb{R}$, но рассматриваем $d(S_0, S_1) \in \mathbb{Z}^+$.
  Поэтому в текущем виде применить его для оценивания малых вероятнстей, связанных со строками, нельзя. Требуется адаптация алгоритма.