

Задача

Предлагается следующий алгоритм reservoir sampling:

```
class ReservoirSampler:
    def __init__(self, nsamples):
        self.nsamples = nsamples
        self.nseen = 0
        self.result = []

    def sample(self, item):
        self.nseen += 1
        if self.nseen <= self.nsamples:
            self.result += [item]
        else:
            if random() < self.nsamples/self.nseen:
                i = int(random()*self.nsamples)
                self.result[i] = item
```

Необходимо доказать, что любой элемент входной последовательности окажется в конечной выборке с равной вероятностью.

Решение

Рассмотрим алгоритм поэтапно и вычислим вероятности включения каждого элемента.

Обозначения

- n — общее количество элементов во входной последовательности (известно только в конце).
- k — размер резервуара (`self.nsamples`).
- Элементы последовательности нумеруются от 1 до n .

Алгоритм работает так

1. Первые k элементов всегда попадают в резервуар (вероятность = 1).
2. Для элемента номер m (где $m > k$):
 - Он выбирается с вероятностью k/m .
 - Если выбран, то случайно заменяет один из текущих k элементов (с вероятностью $1/k$ для каждого).

Доказательство равномерности

Нужно показать, что для любого элемента i ($k < i \leq n$), вероятность того, что он окажется в итоговом резервуаре, равна k/n .

Случай 1: $i \leq k$ (элемент попал в резервуар на начальном этапе)

Изначально элемент i находится в резервуаре.

Чтобы он остался в резервуаре, на каждом шаге $m > k$:

- Либо элемент m не выбран (вероятность $1 - k/m$);
- Либо элемент m выбран (вероятность k/m), но заменяет другой элемент (вероятность $(k-1)/k$).

Таким образом, вероятность того, что элемент i останется после шага m , равна:

$$\left(1 - \frac{k}{m}\right) + \frac{k}{m} \cdot \frac{k-1}{k} = \frac{m-k+k-1}{m} = \frac{m-1}{m}$$

Общая вероятность того, что элемент i останется после всех шагов:

$$P_i = \prod_{m=k+1}^n \left(\frac{m-1}{m}\right)$$

Раскроем произведение:

$$P_i = \frac{k}{k+1} \cdot \frac{k+1}{k+2} \cdots \frac{n-1}{n} = \frac{k}{n}$$

Здесь числители и знаменатели сокращаются телескопически.

Случай 2: $i > k$ (элемент пришел позже)

Элемент i должен быть выбран на шаге $m = i$ с вероятностью k/i .

Далее, для всех последующих шагов $m > i$ он должен не быть заменен:

- Аналогично случаю 1, вероятность не быть замененным на шаге m равна $(m-1)/m$.

Общая вероятность:

$$P_i = \frac{k}{i} \cdot \prod_{m=i+1}^n \left(\frac{m-1}{m}\right)$$

Раскроем произведение:

$$P_i = \frac{k}{i} \cdot \frac{i}{i+1} \cdot \frac{i+1}{i+2} \cdots \frac{n-1}{n} = \frac{k}{n}$$

Вывод

В обоих случаях вероятность того, что элемент i окажется в итоговой выборке, равна k/n .