Число различных элементов

Всеволод Опарин

CS Клуб, Осень 2014

19 Октября

Базовые понятия

- Случайная величина (С.В.) X.
- $\mathbf{E}[\,X\,] = \sum_{i \in Dom(X)} i \cdot \mathbf{Pr}[\,X = i\,]$.
- Для набора С.В. $X_1, X_2, \cdots, X_n$ и чисел $\alpha_1, \cdots, \alpha_n$
  $\mathbf{E}[\,\sum_i \alpha_i \cdot X_i\,] = \sum_i \alpha_i \cdot \mathbf{E}[\,X_i\,]$
- $\mathbf{D}[\,X\,] = \mathbf{E}\left[\,(X - \mathbf{E}[\,X\,])^2\,\right] = \mathbf{E}\left[\,X^2\,\right] - \mathbf{E}[\,X\,]^2$.
- $\mathbf{D}[\,\alpha X\,] = \alpha^2 \mathbf{D}[\,X\,]$.

Базовые понятия

- С.В. X и Y независимы, если $\forall a, b \in \mathbb{R}$

$$\mathbf{Pr}[\, X = a \wedge Y = b \,] = \mathbf{Pr}[\, X = a \,] \cdot \mathbf{Pr}[\, Y = b \,].$$

- Для независмых С.В. X и Y
  $\mathbf{E}[\, XY \,] = \mathbf{E}[\, X \,]\, \mathbf{E}[\, Y \,]$,
  $\mathbf{D}[\, X + Y \,] = \mathbf{D}[\, X \,] + \mathbf{D}[\, Y \,]$.
- Набор С.В. $X_1, \cdots, X_n$ 2-независим, если для любых $1 \leq i \neq j \leq n$ X_i и X_j независимы.
- $X_1, X_2, X_3 \in \{0, 1\}$, $X_1 + X_2 + X_3 = 0 \bmod 2$.
- Для 2-независимого набора С.В.
  $\mathbf{D}[\, \sum_i X_i \,] = \sum_i \mathbf{D}[\, X_i \,]$.

Основные неравенства

- $\mathbf{Pr}[\, A \vee B \,] \leq \mathbf{Pr}[\, A \,] + \mathbf{Pr}[\, B \,]$.
- $\mathbf{Pr}[\, A \wedge B \,] \leq \mathbf{Pr}[\, A \,]$.

Неравенство Маркова

Для С.В. $X \geq 0$ и числа $a > 0$

$$\mathbf{Pr}[\, X \geq a \,] \leq \frac{\mathbf{E}[\, X \,]}{a}.$$

Основные неравенства

Неравенство Чебышёва

Для С.В. X при $\mathbf{E}[\,X\,] = \mu$

$$\mathbf{Pr}[\,|X - \mu| \geq \varepsilon\,] \leq \frac{\mathbf{D}[\,X\,]}{\varepsilon^2}.$$

Неравенство Чернова

Для набора независимых С.В. $X_1, X_2, \cdots, X_n$ таких, что $X_i \in \{0, 1\}$ и $\mathbf{E}[\,X_i\,] = \mu$, и константы $\varepsilon > 0$

$$\mathbf{Pr}\left[\,\left|\frac{1}{n} \cdot \sum_i X_i - \mu\right| \geq \varepsilon\,\right] \leq exp(-c \cdot n).$$

Хэш-функции

- Пусть $\mathcal{H} = \{h_i\}_{i=1}^{t}$, где $h_i : K \to V$. $\mathcal{H}$ – семейство 2-независимых хэш-функций, если для любых $k_1 \neq k_2$ и любых v_1, v_2

$$\Pr_{h \leftarrow U(\mathcal{H})}[\, h(k_1) = v_1 \wedge h(k_2) = v_2 \,] = \frac{1}{|V|^2}.$$

- Пример $h_i : \mathbb{F}_p \to \mathbb{F}_p$. $h_{a,b}(x) = a \cdot x + b$.

$$\Pr_{h \leftarrow U(\mathcal{H})}[\, a \cdot k_1 + b = v_1 \wedge a \cdot k_2 + b = v_2 \,] = \frac{1}{p^2}.$$

$$\begin{cases} a \cdot k_1 + b = v_1, \\ a \cdot k_2 + b = v_2. \end{cases}$$

Число различных элементов

Задача

- $\sigma = \langle a_1, \cdots, a_m \rangle$, $a_i \in [n] = [2^l]$.
- $f[x] = \#\{i \mid a_i = x\}$ – частота.
- $d = \#\{x \mid f[x] > 0\}$ – число различных элементов.
- Разрешен один проход. Найти d.

- FM'83 AMS'99: $\mathbf{Pr}\left[\ ans \in [\frac{d}{3}, 3d]\ \right] \geq 1 - \delta$.

  Память $O(\log \frac{1}{\delta} \log n)$.
- BJKST'04: $\mathbf{Pr}\left[\ |ans - d| \leq \varepsilon d\ \right] \geq 1 - \delta$

  Память $O(\log \frac{1}{\delta}(\log n + \frac{1}{\varepsilon^2}(\log \frac{1}{\varepsilon} + \log \log n)))$.

AMS

- zero(p) $= \max\{i \mid p$ делится на $2^i\}$.
- Алгоритм:
  1. init(): z $= 0$, взять $h : [n] \to [n]$ из 2-независимого семейства.
  2. process(y) : z $= \max($z, zero(h(y))$)$.
  3. answer(): вернуть $2^{z+\frac{1}{2}}$.
- Пусть d – ответ. $\bar{d}$ – результат алгоритма.
  $\mathbf{Pr}\left[\ \bar{d} \geq 3 \cdot d\ \right] \leq 0.47.$

  $\mathbf{Pr}\left[\ \bar{d} \leq \frac{1}{3} d\ \right] \leq 0.47.$

AMS. Анализ

- $X_{j,r} = [zero(h(j)) \geq r]$.

$$\mathbf{E}[\, X_{j,r} \,] = \frac{1}{2^r},$$
$$\mathbf{D}[\, X_{j,r} \,] \leq \frac{1}{2^r}.$$

- $Y_r = \sum_{j|f[j]>0} X_{j,r}$.

$$\mathbf{E}[\, Y_r \,] = \frac{d}{2^r},$$
$$\mathbf{D}[\, Y_r \,] \leq \frac{d}{2^r}.$$

- $\mathbf{Pr}[\, Y_r > 0 \,] = \mathbf{Pr}[\, Y_r \geq 1 \,] \leq \mathbf{E}[\, Y_r \,]/1 = \frac{d}{2^r}$.
- $\mathbf{Pr}[\, Y_r = 0 \,] \leq \mathbf{Pr}\left[\, |Y_r - \frac{d}{2^r}| \geq \frac{d}{2^r} \,\right] \leq \frac{\mathbf{D}[\, Y_r \,]}{(d/2^r)^2} \leq \frac{2^r}{d}$.

AMS. Анализ

- Пусть a – минимальное такое, что $3 \cdot d \leq 2^{a+\frac{1}{2}}$. Тогда

$$\mathbf{Pr}[\ Y_a > 0\ ] \leq \frac{d}{2^a} \leq \frac{\sqrt{2}}{3}$$

.

- Пусть b – максимальное такое, что $2^{b+\frac{1}{2}} \leq \frac{1}{3} \cdot d$. Тогда

$$\mathbf{Pr}[\ Y_{b+1} = 0\ ] \leq \frac{2^{b+1}}{d} \leq \frac{\sqrt{2}}{3}$$

.

Медиана

- Алгоритм возвращает значение, большее разрешенного, с вероятностью $p < \frac{1}{2}$.
- Запустим алгоритм независимо k раз и возьмем медиану.
- $T_i = [\text{ошибка при } i\text{-ом запуске}]$.
- $\mathbf{E}[\ T_i\ ] \leq p$.
- Если медина плохая, то алгоритм ошибся хотя бы половину раз.
- $\mathbf{Pr}\left[\ \left|\frac{1}{k}\sum_i T_i - \mathbf{E}[\ T_i\ ]\right| \geq \frac{1}{2} - p\ \right] \leq exp(-c \cdot k)$.
- Запустим $\log(\frac{2}{\delta})$ и получим ошибку $\frac{\delta}{2}$ с одной стороны.
- $\mathbf{Pr}[$ оценка мала или велика $] \leq$
  $\mathbf{Pr}[$ мала $] + \mathbf{Pr}[$ велика $] \leq \delta$

BJKST. Алгоритм

```
init():
  B = []
  z = 0
  pick 2-ind h : [n] -> [n]

process(y):
  zy = zero(h(y))
  if zy >= z:
    B <- (y, zy)
    while (B > c / (eps * eps)):
      z++
      remove all (a, b) from B s.t. b < z

ans = |B| * pow(2, z)
```

BJKST. Анализ

- $|B| = Y_z$.
- $\mathbf{E}[\,|B|\,] = \mathbf{E}[\,Y_z\,] = \frac{d}{2^z}$.
- $\mathbf{D}[\,|B|\,] = \mathbf{D}[\,Y_z\,] \leq \frac{d}{2^z}$.
- При $z = 0$ всё точно.
- $||B| \cdot 2^z - d| \geq \varepsilon \cdot d \Leftrightarrow \left||B| - \frac{d}{2^z}\right| \geq \varepsilon \cdot \frac{d}{2^z}$

BJKST. Анализ

- Пусть s такое, что $\frac{12}{\varepsilon^2} \leq \frac{d}{2^s} \leq \frac{24}{\varepsilon^2}$.
- $\mathbf{Pr}\left[\ \left||B| - \frac{d}{2^z}\right| \geq \varepsilon \cdot \frac{d}{2^z}\ \right] =$

  $\sum_{i=1}^{\log n} \mathbf{Pr}\left[\ \left||B| - \frac{d}{2^z}\right| \geq \varepsilon \cdot \frac{d}{2^z} \wedge z = i\ \right] \leq$

  $\sum_{i=1}^{s-1} \mathbf{Pr}\left[\ \left||B| - \frac{d}{2^z}\right| \geq \varepsilon \cdot \frac{d}{2^z}\ \right] + \sum_{i=s}^{\log n} \mathbf{Pr}\left[\ z = i\ \right] \leq \frac{1}{6}$
- $\mathbf{Pr}\left[\ \left||B| - \frac{d}{2^i}\right| \geq \varepsilon \cdot \frac{d}{2^i}\ \right] \leq \frac{\mathbf{D}[\ |B|\ ]}{(\varepsilon d/2^i)^2} \leq \frac{d}{2^i}\frac{1}{\varepsilon^2}\left(\frac{2^i}{d}\right)^2 = \frac{1}{\varepsilon^2}\frac{2^i}{d}$.
- $\sum_{i=1}^{s-1} \mathbf{Pr}\left[\ \left||B| - \frac{d}{2^z}\right| \geq \varepsilon \cdot \frac{d}{2^z}\ \right] \leq \frac{1}{\varepsilon^2}\frac{2^s}{d} \leq \frac{1}{12}$.
- $\sum_{i=s}^{\log n} \mathbf{Pr}\left[\ z = i\ \right] = \mathbf{Pr}\left[\ Y_{s-1} \geq \frac{c}{\varepsilon^2}\ \right] \leq \frac{d}{2^{s-1}} \cdot \frac{\varepsilon^2}{c} \leq \frac{24 \cdot 2}{\varepsilon^2} \cdot \frac{\varepsilon^2}{c} \leq \frac{48}{c}$.

BJKST. Память

- $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \log n)$.
- Было:
  `B <- (y, zero(h(y)))`.
- Стало:
  1. В `init`: вытаскиваем еще одну 2-независимую функцию $g : [n] \to [b \cdot \frac{1}{\varepsilon^4} \log^2 n]$.
  2. В `process`: $B \leftarrow (g(y), zero(h(y)))$.
  3. $\mathbf{Pr}[\text{ коллизия }] \leq \frac{1}{6}$.
- Память: $O(\log \frac{1}{\delta}(\log n + \frac{1}{\varepsilon^2}(\log \frac{1}{\varepsilon} + \log \log n)))$

Спасибо!

Вопросы?