Моменты и частые элементы

Всеволод Опарин

CS Клуб, Осень 2014

9 ноября

Задачи

- $\sigma = \langle (a_1, c_1), \cdots, (a_m, c_m) \rangle$.
- $f_x = \sum_{i:a_i=x} c_i$ – частота.
- Кассовая модель: $c_i > 0$.
- Турникетная модель: c_i – произвольные.

Частота элемента

- По заданному x найти f_x.

k-ый момент

- Найти момент $F_k = \sum_{x:f_x>0} f_x^k$.

Эскиз (англ. sketch)

Определение

$\texttt{DS}(\sigma)$ – эскиз, если для двух потоков σ_1 и σ_2

$$\texttt{DS}(\sigma_1 \circ \sigma_2) = \texttt{COMB}(\texttt{DS}(\sigma_1), \texttt{DS}(\sigma_2)).$$

$\texttt{COMB}$ – эффективен по памяти.

Определение

Пусть $A \in \mathbb{R}^{k \times n}$, $\mathbf{f} \in \mathbb{R}^n$. $A\mathbf{f}$ – линейный эскиз.

Эскиз-счетчик

Идея

Берем случайную матрицу A, у которой в каждом столбце один ненулевой элемент: 1 или -1.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}$$

- Ответ для x: $\mathbf{v}^T \cdot A \cdot \mathbf{e}_x$.

Эскиз-счетчик

Алгоритм

- Пусть $k = \frac{1}{\varepsilon^2}$.
- Выбираем две функции
  - $h : [n] \to [k]$,
  - $g : [n] \to \{-1, +1\}$

  из 2-независимых семейств.
- Заводим массив `v[1..k]`.
- Обработка пары (x, c): `v[h(x)] += g(x)` $\cdot$ `c`.
- Ответ для x: $g(x) \cdot \mathtt{v}[h(x)]$.

Эскиз-счетчик

Анализ

- Спросили про элемент a.
- $Y_j = [h(a) = h(j)]$.
- $X = g(a) \sum_j f_j \cdot g(j) \cdot Y_j$.
- $\mathbf{E}[\, X \,] = f_a$.
- $\mathbf{D}[\, X \,] = \frac{||\mathbf{f}||_2^2 - f_a^2}{k}$.
- $\mathbf{Pr}[\, |X - f_a| \geq \varepsilon \cdot ||f_{-a}||_2 \,] \leq \frac{1}{3}$.

Повторить $O(\log \frac{1}{\delta})$ раз, взять медиану.

- Гарантия: $\mathbf{Pr}[\, |X - f_a| \geq \varepsilon \cdot ||f_{-a}||_2 \,] \leq \delta$.
- Память: $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \cdot (\log m + logn))$.

Эскиз-мин. счетчик

Идея

Берем случайную матрицу A, у которой в каждом столбце ровно одна единица.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}$$

- Ответ для x: $\mathbf{v}^T \cdot A \cdot \mathbf{e}_x$.

Эскиз-мин. счетчик

Алгоритм

- Пусть $k = \frac{1}{2 \cdot \varepsilon}$.
- Выбираем одну функцию
  - $h : [n] \to [k]$,

  из 2-независимого семейства.
- Заводим массив `v[1..k]`.
- Обработка пары (x, c): `v[h(x)] += c`.
- Ответ для x: $\mathtt{v}[h(x)]$.

Эскиз-мин. счетчик

Анализ

- Спросили про элемент a.
- $Y_j = [h(a) = h(j)]$.
- $\mathtt{ans} = f_a + X$.
- $X = \sum_{j \neq a} f_j \cdot Y_j$.
- $\mathbf{E}[\, X \,] = \frac{||\mathbf{f}_{-a}||_1}{k}$.
- $\mathbf{Pr}[\, X \geq \varepsilon \cdot ||f_{-a}||_1 \,] \leq \frac{1}{2}$.

Повторить $O(\log \frac{1}{\delta})$ раз, взять минимум.

- Гарантия: $\mathbf{Pr}[\, |\mathtt{ans} - f_a| \geq \varepsilon \cdot ||f_{-a}||_1 \,] \leq \delta$.
- Память: $O(\frac{1}{\varepsilon} \log \frac{1}{\delta} \cdot (\log m + logn))$.

Момент F_2

Задача

- Дан поток $\sigma = \langle (a_1, c_1), \cdots, (a_m, c_m) \rangle$. $f_x = \sum_{i:a_i=x} c_i$.
- Посчитать $F_2 = \sum f_x^2$.

Идея

Берем случайную матрицу $A \in \mathbb{R}^{t \times n}$, у которой все элементы из $\{-1, +1\}$.

$$\frac{1}{\sqrt{t}} \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_t \end{pmatrix}$$

- Ответ: $||\mathbf{v}||_2^2$.

Момент F_2

Алгоритм

- Выбираем одну функцию
  - $g : [n] \to \{-1, +1\}$

  из 4-независимого семейства.
- Заводим переменную $\mathtt{x} = 0$.
- Обработка пары (x, c): $\mathtt{x} \mathrel{+}= c \cdot \mathtt{g}(\mathtt{x})$.
- Ответ: $\mathtt{x}^2$.

Медиана средних

- $\mathbf{E}[\,X\,] = Q$. Известна $\mathbf{D}[\,X\,]$
- Цель – Z: $\mathbf{Pr}[\,|Z - Q| \geq \varepsilon \cdot Q\,] \leq \delta$.
- $Y = \frac{1}{k}\sum_{j=1}^{k} X_j$.
- $\mathbf{D}[\,Y\,] = \frac{1}{k}\mathbf{D}[\,X\,]$.
- $\mathbf{Pr}[\,|Y - Q| \geq \varepsilon \cdot Q\,] \leq \frac{\mathbf{D}[\,X\,]}{k \cdot \varepsilon^2 Q^2}$.
- $k = \frac{3 \cdot \mathbf{D}[\,X\,]}{Q^2\varepsilon^2}$, чтобы получить $\frac{1}{3}$.
- Повторить $O(\log \frac{1}{\delta})$ раз, взять медиану.
- Ухудшение по памяти: $O(\varepsilon^{-2} \log \frac{1}{\delta} \frac{\mathbf{D}[\,X\,]}{Q^2})$.

Момент F_2

Анализ

- $Y_j = h(j)$
- $X = \sum_j f_j \cdot Y_j$.
- $\mathbf{E}\left[X^2 \right] = F_2$.
- $\mathbf{E}\left[X^4 \right] = F_4 + 6 \cdot \sum_{i<j} f_i^2 \cdot f_j^2$.
- $\mathbf{D}\left[X^2 \right] = 2 \cdot F_2^2$.

Применяем медиану средних.

- Гарантия: $\mathbf{Pr}\left[\left| ||\mathbf{v}||_2^2 - F_2 \right| \geq \varepsilon \cdot F_2 \right] \leq \delta$.
- Память: $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} (\log m + \log n))$.

Регрессия

Задача

- Y – зависимая переменная.
- A_j – предикторы.
- Хотим построить линейную зависимость $Y = x_1 \cdot A_1 + \cdots + x_d \cdot A_d$.
- Наблюдение за предикторами дается матрицей A. Наблюдения за Y – вектором y.
- Найти $x* = \arg\min ||Ax - y||_2$.

Нужно решить $A^T Ax = A^T y$, а наблюдений много...

Регрессия

Идея

- Пусть $m = O(d\varepsilon^{-1} \log \frac{1}{\delta})$.
- Возьмем матрицу $S \in \{-1, +1\}^{n \times m}$.
- Матрица должна быть $\Omega(d + \log \frac{1}{\delta})$-независимой.
- Найдем $\hat{x} = \arg\min ||S^T(Ax - y)||_2$.
- Гарантия: $\mathbf{Pr}\left[\, ||A\hat{x} - y||_2 \leq (1+\varepsilon)||Ax^* - y||_2 \,\right] \geq 1 - \delta$.
- Память: $O(d^2\varepsilon^{-1} \log \frac{1}{\delta} \log N)$.

Момент F_k

Задача

- Дан поток $\sigma = \langle a_1, \cdots, a_m \rangle$, $a_i \in [n]$.
- Найти $F_k = \sum_{x:f_x>0} f_x^k$.

Алгоритм

- Взять случайный элемент a в потоке.
- Посчитать, сколько раз он встретился. Счетчик в r.
- Вернуть $m \cdot (r^k - (r-1)^k)$.

Момент F_k

Анализ

- A – выбранный элемент. R – значение r в конце.
- Выбираем элемент, выбираем появление.
- $\mathbf{E}[\,X\,] = F_k$.
- $\mathbf{D}[\,X\,] \le \mathbf{E}\left[\,X^2\,\right] = k \cdot F_1 \cdot F_{2k-1}$.

Лемма

Для $n > 0$, $x_i \ge 0$, $k \ge 1$

$$(\sum x_i)(\sum x_i^{2k-1}) \le n^{1-\frac{1}{k}}(\sum x_i^k)^2.$$

Суммы берутся по $i \in [n]$.

- $\mathbf{D}[\,X\,] \le k n^{1-\frac{1}{k}} F_k^2$.
- Память: $O(\varepsilon^{-2} \cdot \log \frac{1}{\delta} \cdot k \cdot n^{1-\frac{1}{k}}(\log m + \log n))$

Спасибо!

Вопросы?