DS v sublineárním prostoru

- vstup se nevejde do paměti → streaming model (1 nebo více sekvenčních průchodů)
- typicky toliko aproximace, často randomizovaná

Bloomův filtr – reprezentace množiny s 1strannou chybou

- pole k bitů $B[0] \dots B[k-1]$ + hešovací fce $h: \mathcal{U} \to [k]$
- Insert(x): $B[h(x)] \leftarrow 1$
- Find(x) testuje $B[h(x)]$
- Delete neumíme

} $x \in M \Rightarrow$ odpovíme ANO; $x \notin M \Rightarrow$ někdy také ANO

Rozbor: $\Pr_{(x \notin M)}[\text{chyba}] = 1 - \prod_{a \in M} (1 - \underbrace{\Pr[h(a) = h(x)]}_{1/k}) = 1 - (1 - 1/k)^n \approx 1 - e^{-\frac{n}{k}}$ … stlačíme pod 1/k volbou $k = cn$

↑ pozor, silné požadavky na nezávislost h!

→ pro $\delta = 1/2$ [Pr chyby] potřebujeme prostor $\Theta(n)$ bitů a čas $O(1)$ na op.

Zesílení: Pořídíme si t filtrů paralelně s různými h, $\text{Find}(x) = \bigwedge_{i=1}^{t} \text{Find}_i(x)$.

$\Rightarrow \delta \leq 2^{-t}$, čas $O(t)$, prostor $\Theta(tn)$.

[též lze nechat všechny h_i hešovat do téhož pole apod.]

Majorita – úkol: najít prvek, který se ve streamu vyskytuje více jak $n/2$-krát.

- udržujeme kandidáta k a jeho počítadlo c

Init: $k \leftarrow \emptyset$, $c \leftarrow 0$
Process(x): Pokud $k = \emptyset$: $k \leftarrow x$
Pokud $x = k$: c++
Jinak: c––
Pokud $c = 0$: $k \leftarrow \emptyset$

Věta: Pokud prvek m je majoritní, pak na konci $k = c$. (dalším průchodem možno ověřit)

Dk: Rozdělíme vstup na bloky, blok končí resetem k:

V těchto blocích má kandidát přesně polovinu zastoupení

↑ zde musí mít nadpoloviční ⇒ je to kandidát

$\Rightarrow O(n)$ času, $O(\log n + \log \mathcal{U})$ prostoru

Odhad frekvencí (zejm. pro hledání prvků vyskytujících se více jak (n/k)-krát) – Misra, Gries 1982

- udržujeme slovník počítadel $A[x]$, třeba v BVS

Init: $A \leftarrow \emptyset$
Process(x): Pokud $x \in \text{keys}(A)$: $A[x]$++
Jinak je-li $|\text{keys}(A)| < k-1$: $A[x] \leftarrow 1$
Jinak: $\forall a \in \text{keys}(A)$: $A[a]$––
Pokud klesl na 0, smažeme $A[a]$.

pro analýzu: představme si, že $A[x]$ vždy zvýšíme a pak možná zase hned snížíme

Output: Pro $x \in \text{keys}(A)$ hlásíme četnost $A[x]$, jinak hlásíme 0.

obecně dodef. $A[a] \leftarrow 0$, kdykoli $a \notin \text{keys}(A)$

Paměť: $O(k \cdot (\log n + \log \mathcal{U}))$
Čas: $O(n \cdot \log k)$

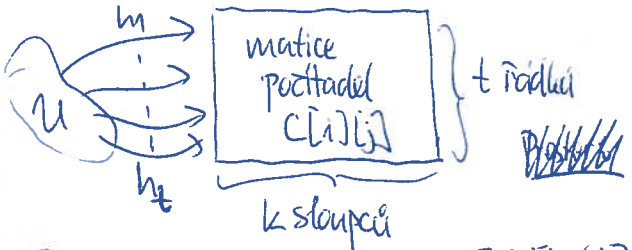
Značení: f_a = skutečná frekvence a
$\hat{f}_a$ = náš odhad, tedy $A[a]$ na konci výpočtu

Věta: $\forall a$ $f_a - n/k \leq \hat{f}_a \leq f_a$

Dk: kdykoli snižuji $A[a]$, snížím i $k-1$ dalších počítadel ⇓ stane se to max. n/k-krát (potenciál: Σ počítadel); $\hat{f}_a \leq f_a$ zjevně

Důsledek: Umíme najít (na 2 průchody) $\{a \mid f_a > n/k\}$

Pravděpodobnostní odhad frekvencí - Count-Min



$h_1 \dots h_t : U \to [k]$ náhodně vybereme z 2-nezávislého systému funkcí

(tady ještě stavila obyč. univerzalita)

$\forall x \neq y \in U \; \forall i,j \in [k] \; Pr[h(x)=i \wedge h(y)=j] = \frac{1}{k^2}$

Process(x): Pro $i=1 \dots t$: $C[i][h_i(x)]++$

Result(x): $\min_i C[i][h_i(x)]$

Chceme (ε,δ)-estimátor: $Pr\left[\left|\frac{x-OPT}{OPT}\right| > \varepsilon\right] < \delta$ (něco jako FPTAS)

Nastavíme $k := \lceil 2/\varepsilon \rceil$, $t := \lceil \log 1/\delta \rceil$

Analýza: f_a opt. frekvence (skutečné), $\hat{f}_a$ výstup algoritmu ... evidentně $\hat{f}_a \geq f_a$ ... o kolik víc?

Fix a: $X_i := C[i][h_i(a)] - f_a$

Pro $j \in U \setminus \{a\}$: $Y_{i,j} := [h_i(j) = h_i(a)]$ ... $X_i = \sum_{j \neq a} f_j \cdot Y_{i,j}$

$\mathbb{E}[Y_{i,j}] = 1/k \Rightarrow \mathbb{E}[X_i] = \frac{\|f\|_1 - f_a}{k} = \frac{\|f-a\|_1}{k}$

Podle Markova: $Pr[X_i > \varepsilon \cdot \underbrace{\|f-a\|_1}_{\leq n}] < \frac{1}{k\varepsilon} \leq \frac{1}{2}$

Pak: $Pr[\hat{f}_a - f_a > \varepsilon n] = Pr[\forall i \; X_i > \varepsilon n] < 2^{-t} \leq \delta$, ✓ je to (ε,δ)-estimátor.

prostor: $O(\log \varepsilon \log 1/\delta \log n) + \log 1/\delta \log U$

Odhad # různých prvků - Alon, Matias, Szegedy 1999 [AMS]

Def: $tz(x) := \max\{i \mid 2^i | x\}$ (# trailing zeroes) "pravitková funkce"

Init: $z \leftarrow 0$, $h : U \to U$ z 2-nezávislého syst. náhodně vybraná

Process(x): $z \leftarrow \max(z, tz(h(x)))$

Output: $2^{z+1/2}$

Intuice: každý 2^i-tý prvek má $tz(x) \geq i$

Analýza: $\forall j \in U \; \forall r \geq 0$ $X_{r,j} := [tz(h(j)) \geq r]$ ... tyto jsou 2-nezávislé jevy

$Y_r := \sum_{j : f_j > 0} X_{r,j}$

t := hodnota z na konci výpočtu

d := # různých prvků

$\hat{d}$:= náš odhad, tedy $2^{t+1/2}$

$t \geq r \Leftrightarrow Y_r > 0$ (vidíme)

$\mathbb{E}[X_{r,j}] = 1/2^r$, $\mathbb{E}[Y_j] = d/2^r$

$var[X_{r,j}] \leq \mathbb{E}[X_{r,j}^2] = \mathbb{E}[X_{r,j}] = 1/2^r$

$var[Y_r] = \sum_{j : f_j > 0} var[X_{r,j}] \leq d/2^r$ (díky 2-nezávislosti)

Nyní zvolme a nejmenší celé t.ž. $2^{a+1/2} \geq 3d$. Pak:

$$\Pr[\hat{d} \geq 3d] = \Pr[t \geq a] = \Pr[Y_a > 0] = \Pr[Y_a \geq 1] \leq \mathbb{E}[Y_a] = \frac{d}{2^a} \leq \frac{\sqrt{2}}{3} \doteq 0{,}47$$

($t \geq a$: z def. Y_a; $Y_a > 0$: z celočíselnosti; $Y_a \geq 1$: Markov)

Pro opačný směr zvolme b největší celé, pro které $2^{b+1/2} \leq d/3$. Pak:

$$\Pr[\hat{d} < d/3] = \Pr[t \leq b] = \Pr[Y_{b+1} = 0] \leq \Pr\left[|Y_{b'} - \mathbb{E}[Y_{b'}]| > \frac{d}{2^{b'}}\right] \leq \frac{2^{b+1}}{d} \leq \frac{\sqrt{2}}{3}$$

($t \leq b$: z def. Y_{b+1}; b'; $\frac{d}{2^{b'}} = \sqrt{\frac{d}{2^{b'}}} \cdot \sqrt{\mathrm{var}\, Y_{b'}}$; Čebyšev)

Získali jsme tedy $(3, \frac{\sqrt{2}}{3})$-estimátor.

Pr chyby lze snadno zlepšit: spustíme paralelně t estimátorů, vrátíme medián ...
-- podle Černova $\delta < 2^{-\Theta(t)}$.
→ $(3,\delta)$-estimátor pro libovolné δ, prostor $O(\log \frac{1}{\delta} \cdot \log U)$, čas $O(\log \frac{1}{\delta} \cdot n)$

Lepší odhad # různých prvků, ... tentokrát pro libovolné $\varepsilon, \delta > 0$. Bar-Yossef et al., 2004 (BJKST)

- Opět $h: U \to U$ a počítání $tz(h(x))$
- Tentokrát si pamatujeme B – množinu hodnot s daným tz, ale omezíme její velikost.

Init: Zvolíme h náhodně, $z \leftarrow 0$, $B \leftarrow \emptyset$

Process(x): Je-li $tz(h(x)) \geq z$:
$B \leftarrow B \cup \{(x, tz(h(x)))\}$
Dokud $|B| \geq c/\varepsilon^2$: (nastavíme později)
$z \leftarrow z+1$
odstraníme z B všechny dvojice s $tz < z$

Intuice: cca $n/2^z$ prvků se objeví v B ↓ $|B| \cdot 2^z$ je dobrý odhad

Output: $|B| \cdot 2^z$

Prostor: $O(\frac{1}{\varepsilon^2} \cdot \log U)$ [lze zlepšit: místo x si v B pamatujeme nějaké lusce do dost velkého prostoru]

Analýza: $X_{r,j}$ a Y_r jako předtím.
$\hat{d} = 2^t \cdot Y_t$ (t je poslední hodnota proměnné z)
Pokud $t=0$, alg. počítá přesně.

Chyba $\Leftrightarrow$ $|Y_t \cdot 2^t - d| > \varepsilon \cdot d \Leftrightarrow |Y_t - \frac{d}{2^t}| > \frac{\varepsilon d}{2^t}$ ($Y_t \cdot 2^t = \hat{d}$)

pro malá a velká t odhadneme jinde, prahy zvolíme s tak, aby $\frac{12}{\varepsilon^2} \leq \frac{d}{2^s} < \frac{24}{\varepsilon^2}$

$$\text{Nyní: } \Pr[\text{chyba}] = \sum_{r=1}^{\log n} \Pr\left[|Y_r - \tfrac{d}{2^r}| > \tfrac{\varepsilon d}{2^r} \;\&\; t=r\right] \leq \sum_{r=1}^{s-1} \Pr\left[|Y_r - \tfrac{d}{2^r}| > \tfrac{\varepsilon d}{2^r}\right] + \sum_{r=s}^{\log n} \Pr[t=r]$$

(zanedbáme pro malá r)

$\Pr[|Y_r - \mathbb{E}[Y_r]| > \frac{\varepsilon d}{2^r}]$ — Čebyšev — $\leq \frac{2^r}{\varepsilon^2 d}$

$\Pr[t \geq s] = \Pr[Y_{s-1} \geq c/\varepsilon^2]$

Viz minul. cvičení

$$\star \le \sum_{r=1}^{s-1} \Pr\left[\left|Y_r - \frac{d}{2^r}\right| > \frac{\varepsilon d}{2^r}\right] + \sum_{r=s}^{\log n} \Pr[t=r]$$

$$\Pr\left[\left|Y_r - \mathbb{E}[Y_r]\right| > \frac{\varepsilon d}{2^r}\right]$$

↓ Čebyšev

$$< \frac{d}{2^r} \Big/ \left(\frac{\varepsilon d}{2^r}\right)^2 = \frac{d \cdot 2^r \cdot 2^r}{d^2 \varepsilon^2 2^r} = \frac{2^r}{\varepsilon^2 d}$$

$$\sum_{r=1}^{s-1} \frac{2^r}{\varepsilon^2 d} \le \frac{2^s}{\varepsilon^2 d} \le \frac{\varepsilon^2}{\varepsilon^2 \cdot 12} = \frac{1}{12}$$

↑ geom. řada ↑ volba s

$$= \Pr[t \ge s] = \Pr[Y_{s-1} \ge c/\varepsilon^2]$$

↓ Markov

$$< \frac{d}{2^{s-1}} \Big/ \frac{c}{\varepsilon^2} = \frac{2d\varepsilon^2}{2^s \cdot c} \le \frac{2 \cdot 24\,\varepsilon^2}{\varepsilon^2 \cdot c} = \frac{48}{c}$$

⇒ celé určeme stálnost pod $\frac{1}{12}$ volbou vhodné c

$\Pr[\text{chyba}] \le 1/6$ … pod 5 stranné medicínským trikem.