Eksamensdisposition - Distinct Elements

Søren Mulvad, rbn601

17. juni 2019

- **Problem**
- **AMS algoritmen**
- **Kvaliteten af estimatet**
  - For højt estimat $\hat{d}$
  - For lavt estimat $\hat{d}$
- **Median tricket**

Eksamensdisposition - Distinct Elements

Problem

Vi lader en strøm $j_1, \ldots, j_s \in [n]$ passere. Da lader vi f_j betegne antal forekomster af j i strømmen, svarende til antal elementer $\{i | j_i = j\}$.

Nu ønsker vi at give et estimat på d, som beskriver hvor mange forskellige elementer der var i strømmen, hvor

$$S = \{j | f_j > 0\}, \qquad\qquad d = |S|$$

AMS algoritmen

Vi definerer nu for et heltal $y > 0$ funktionen zeros(y) til at være antal trailing 0'er i bit-formen af y.

F.eks. vil $y = 288_{10}$ give zeros$(y) = 1001\underbrace{00000}_{zeros} = 5$ og vi ser $y \bmod 2^5 = 0$ mens $y \bmod 2^6 \neq 0$.

Algorithm 1: AMS

Init
1 h = new random strong-universel hash function s.t. $h : [n] \to [n]$
2 $z = 0$
Process j
3 **if** $\texttt{zeros}(h(j)) > z$
4 $\quad z = \texttt{zeros}(h(j))$
5 **return** $\hat{d} = 2^{z+1/2}$

Hvor Process j svarer til vi løbende kører alle j'er i strømmen igennem i linje 3-4, hvor vi potentielt kunne stoppe på et vilkårligt tidspunkt.

Intuitionen for algoritmen er, at vi forventer 1 ud af de d unikke elementer hashes så $\texttt{zeros}(h(j)) \geq \log d$, og vi forventer ikke nogle elementer at ramme $\texttt{zeros}(h(j)) \gg \log d$. Derfor vil maksværdien af $\texttt{zeros}(h(j))$ være en god approksimation af $\log d$.

Kvaliteten af estimatet

For ethvert $j \in [n]$ og ethvert heltal $r \geq 0$, definer da indikatorvariablen:

$$X_{r,j} = [\texttt{zeros}(h(j)) \geq r]$$

Lad Y_r være en stokastisk variabel for et givent r som beskriver antal elementer j der både indgår i strømmen og opfylder at $\texttt{zeros}(h(j)) \geq r$:

$$Y_r = \sum_{j : f_j > 0} X_{r,j} \tag{1}$$

Lad t beskrive værdien af z idet algoritmen terminerer. Så har vi følgende:

$$t \geq r \iff Y_r > 0 \tag{2}$$

Idet der minimum har været et element der opfyldte at $\texttt{zeros}(h(j)) \geq r$.

Vi kan også omskrive det til, at såfremt intet element opfyldte det har vi:

$$t \leq r - 1 \iff Y_r = 0 \tag{3}$$

Siden $h(j)$ er uniformt distribueret over $(\lg n)$-bitstrengene har vi:

$$\mathbb{E}[X_{r,j}] = \mathbb{P}[\texttt{zeros}(h(j)) \geq r] = \frac{1}{2^r}$$

Vi kan nu bestemme forventning Y_r som følger, idet vi bruger linearity of expectation på (1) og vi har defineret d til at være antal unikke elementer:

$$\mathbb{E}\left[Y_r\right] = \sum_{j:f_j>0} \mathbb{E}\left[X_{r,j}\right] = \frac{d}{2^r} \tag{4}$$

Variansen fås til:

$$\begin{aligned}
\mathrm{Var}\left[Y_r\right] &= \sum_{j:f_j>0} \mathrm{Var}\left[X_{r,j}\right] && (5)\\
&= \sum_{j:f_j>0} \mathbb{E}\left[X_{r,j}^2\right] - \mathbb{E}\left[X_{r,j}\right]^2 && (6)\\
&\leq \sum_{j:f_j>0} \mathbb{E}\left[X_{r,j}^2\right]\\
&= \sum_{j:f_j>0} \mathbb{E}\left[X_{r,j}\right] && (7)\\
&= \frac{d}{2^r}
\end{aligned}$$

Hvor vi i (5) bruger den parvise uafhængighed der følger af at vi valgte en stærk universel hashing-funktion, i (6) bruger definitionen på varians og i (7) bruger at de er indikator-variabler.

Lad os nu definere $\hat{d}$ til at være estimatet af d som algoritmen returnerer, hvorved vi har $\hat{d} = 2^{t+\frac{1}{2}}$.

For højt estimat $\hat{d}$

Vi lader nu a være det mindste heltal så $2^{a+\frac{1}{2}} \geq 6d$. Så får vi:

$$\begin{aligned}
\mathbb{P}\left[\hat{d} \geq 6d\right] &= \mathbb{P}\left[t \geq a\right] && (8)\\
&= \mathbb{P}\left[Y_a > 0\right] && (9)\\
&= \mathbb{P}\left[Y_a \geq 1\right]\\
&= \frac{\mathbb{E}\left[Y_r\right]}{1} && (10)\\
&\leq \frac{d}{2^a} < \frac{1}{4} && (11)
\end{aligned}$$

I (8) benytter vi værdien for $\hat{d}$ og vores definition på a.
I (9) benytter vi $t \geq r \iff Y_r > 0$ fra (2).
I (10) bruger vi Markovs ulighed.
I (11) benytter vi vores værdi for $\mathbb{E}\left[Y_r\right]$. Man får herefter en talværdi mindre end 1/4 såfremt man indsatte a.

For lavt estimat $\hat{d}$

Lad os tilsvarende kigge på sandsynligheden for at vi får noget for småt. Lad b være det største heltal så $2^{b+\frac{1}{2}} \leq d/6$. Da får vi:

$$\begin{aligned}
\mathbb{P}\left[\hat{d} \leq d/6\right] &= \mathbb{P}\left[t \leq b\right] && (12)\\
&= \mathbb{P}\left[Y_{b+1} = 0\right] = \mathbb{P}\left[\left|Y_{b+1} - \mathbb{E}\left[Y_{b+1}\right]\right| \geq \mathbb{E}\left[Y_{b+1}\right]\right] && (13)\\
&\leq \frac{\mathrm{Var}\left[Y_{b+1}\right]}{\mathbb{E}\left[Y_{b+1}\right]^2} = \frac{1}{\mathbb{E}\left[Y_{b+1}\right]} \leq \frac{2^{b+1}}{d} < \frac{1}{4} && (14)
\end{aligned}$$

I (12) benytter vi vores værdier for $\hat{d}$ og b.
I (13) benytter vi $t \leq r \iff Y_b = 0$ fra (3).

I (14) benytter vi Chebyshevs ulighed og at $\mathrm{Var}[Y_r] = \mathbb{E}[Y_r]$.

Vi ser at garantierne er relativt små. Det ses både ved at $\hat{d}$ kan afvige med en del og at vores sandsynligheder kun er begrænset af omkring 25 %.

Median tricket

Lav nu k uafhængige $X_0, \ldots, X_{k-1}$ estimater i parallel (ved at bruge forskellige hashfunktioner) og returner medianen $X_{(\lceil k/2 \rceil)}$ af de k svar.

Lad $B_i = [\hat{X}_i \geq 6d]$ og definer $B = \sum_{i \in [k]} B_i$. Vi ser at median tricket fejler når $B \geq k/2$.

Den forventede værdi af B er:

$$\mathbb{E}[B] = \mu = \sum_{i \in [k]} \mathbb{P}[B_i] \leq k/4$$

Da får vi

$$\mathbb{P}[B \geq 2\mu] = \mathbb{P}[B \geq (1+\delta)\mu] \leq e^{-\delta^2 \mu/3} = e^{-k/12}$$

Hvor vi bruger Chernoff Bounds til at begrænse sandsynligheden, og herudover i eksponentialet benytter $\delta = 1$.

Herved har vi altså væsentligt begrænset sandsynligheden for at få noget rimelig forkert.