Eksamensdisposition - Heavy Hitters

Søren Mulvad, rbn601

17. juni 2019

- **Problem**
- **Basic Count Sketch Algoritme**
- **Kvalitet af Basic Count Sketch estimatet**
  - Forventet værdi
  - Varians
  - Afvigelse
- **Full Count Sketch og Median tricket**

Eksamensdisposition - Heavy Hitters

Problem

Vi har en strøm af par $(j_0, \Delta_0), \ldots, (j_{s-1}, \Delta_{s-1}) \in [n] \times \mathbb{Z}$ (sættet af alle heltal).
Vi definerer frekvensen f_j for hver eneste $j \in [n]$ som

$$f_j = \sum_{i \in [s], j_i = j} \Delta_i$$

Nu ønsker vi givet et a at beregne et estimat $\hat{f}_a$ for f_a.

Basic Count Sketch Algoritme

Algorithm 1: Basic Count Sketch

Init
1 $k = \lceil \frac{4}{\epsilon^2} \rceil$
2 $C[0, \ldots, k-1] = 0$
3 $h =$ strong-universel $h : [n] \to [k]$
4 $s =$ strong-universel $s : [n] \to \{-1, +1\}$
Process (j, Δ)
5 $C[h(j)] \mathrel{+}= s(j) \cdot \Delta$
Output
6 **return** $\hat{f}_a = s(a) \cdot C[h(a)]$

Hvor Process (j, Δ) svarer til vi løbende kører alle par i strømmen igennem i linje 5, hvor vi potentielt kunne stoppe på et vilkårligt tidspunkt.

Kvalitet af Basic Count Sketch estimatet

Lad os fiksere en nøgle a og betragte outputtet $X = \hat{f}_a$ for en query a.
For enhver nøgle $j \in [n]$, definer da indikatorvariablen $Y_j = [h(j) = h(a)]$.
Vi ser, at en nøgle j bidrager til tælleren $C[h(a)]$ hvis og kun hvis $h(j) = h(a)$.
Mængden den bidrager med er den frekvens f_j ganget med den tilfældige fortegn $s(j)$. Derfor:

$$\begin{aligned} X = \hat{f}_a &= s(a) \cdot C[h(a)] \\ &= s(a) \sum_{j \in [n]} f_j \, s(j) Y_j && (1) \\ &= f_a \, s(a) s(a) Y_a + s(a) \sum_{\substack{j \in [n] \\ j \neq a}} f_j \, s(j) Y_j && (2) \\ &= f_a + s(a) \sum_{\substack{j \in [n] \\ j \neq a}} f_j \, s(j) Y_j && (3) \end{aligned}$$

I (1) benytter vi at et element kun tælles med når $Y_j = 1$.
I (2) splitter vi vores sum i to, så vi tager højde for den unikke case når $j = a$ og alle andre cases.
I (3) benytter vi $s(a)s(a) = 1 \cdot 1$ eller $(-1)(-1) = 1$ og $Y_a = [h(a) = h(a)] = 1$.

Vi kan da regne på den forventede værdi af udtrykket i sumtegnet i (3):

$$\mathbb{E}[f_j \, s(j) Y_j] = f_j \underbrace{\mathbb{E}[s(j)]}_{0} \mathbb{E}[Y_j] = 0 \tag{4}$$

Hvor forventningen af produktet er lig produktet af de forskellige forventninger da s er 2-uafhængig, og s og h er uafhængige af hinanden.

Da kan vi bruge (4) til at regne videre på (3) hvorved vi får:

$$\mathbb{E}[X] = f_a + s(a) \sum_{\substack{j \in [n] \\ j \neq a}} \mathbb{E}[f_j\, s(j) Y_j] = f_a \tag{5}$$

Hermed har vi altså vist at $X = \hat{f}_a$ er en unbiased estimator for frekvensen f_a. Men vi skal stadig vise at det er usandsynligt den afviger for meget fra dens forventede værdi.

Derfor analyserer vi dens varians:

$$\begin{aligned}
\operatorname{Var}[X] &= \mathbb{E}\left[(\hat{f}_a - f_a)^2\right] \\
&= \mathbb{E}\left[\left(f_a + s(a) \sum_{\substack{j \in [n] \\ j \neq a}} f_j\, s(j) Y_j - f_a\right)^2\right] \\
&= \mathbb{E}\left[\left(s(a) \sum_{\substack{j \in [n] \\ j \neq a}} f_j\, s(j) Y_j\right)^2\right] \\
&= \underbrace{(s(a))^2}_{1} \sum_{\substack{i,j \in [n] \\ i \neq a,\, j \neq a}} \mathbb{E}[f_i f_j s(i) s(j) Y_i Y_j]
\end{aligned}$$

Vi bruger nu, at h er stærk universel, så for ethvert $j \in [n]$ hvor $j \neq a$ har vi:

$$\mathbb{E}\left[Y_j^2\right] = \mathbb{E}[Y_j] = \mathbb{P}[h(j) = h(a)] = \frac{1}{k}$$

da $Y_j = 0 \vee 1$ og vi har $0^2 = 0$ og $1^2 = 1$.

Hvis vi kigger på udtrykket i summen, så har vi:

$$\mathbb{E}[f_i f_j s(i) s(j) Y_i Y_j] = \begin{cases} f_j^2 \underbrace{(s(j))^2}_{1} \underbrace{\mathbb{E}\left[Y_j^2\right]}_{\mathbb{E}[Y_j]} = f_j^2/k & i = j \\ f_i f_j \underbrace{\mathbb{E}[s(i)]}_{0} \underbrace{\mathbb{E}[s(j)]}_{0} \mathbb{E}[Y_i Y_j] = 0 & i \neq j \end{cases}$$

Vi definerer $\sum_{j \in [n]} f_j^2 = ||\mathbf{f}||_2^2$ (udtales "to-normen i anden"), hvorved vi kan regne ud vores udtryk bliver (hvor $||\mathbf{f}_{-a}||_2^2 = ||\mathbf{f}||_2^2 - f_a^2$.):

$$\operatorname{Var}[X] = \sum_{\substack{j \in [n] \\ j \neq a}} \frac{f_j^2}{k} = \frac{||\mathbf{f}_{-a}||_2^2}{k}$$

Med vores informationer for forventning og varians kan vi nu benytte Chebyshev:

$$\mathbb{P}\left[|\hat{f}_a - f_a| \geq \epsilon ||\mathbf{f}_{-a}||_2\right] \leq \frac{\operatorname{Var}[X]}{\epsilon^2 ||\mathbf{f}_{-a}||_2^2} = \frac{1}{k\epsilon^2} \leq \frac{1}{4}$$

Idet vi satte $k = \lceil 4/\epsilon^2 \rceil$ i vores algoritme.

Median tricket

Vi laver t uafhængige estimater $X_0, \ldots, X_{t-1}$ i parallel (ved at bruge forskellige hashfunktioner) og returnerer medianen $X_{(\lceil t/2 \rceil)}$ af de t svar (Tal om hvad man skulle ændre i algoritmen. Denne algoritme kaldes ofte ”Final Count Sketch”).

Vi siger X_i fejler hvis $|X_i - \mathbb{E}[X]| \geq \epsilon ||\mathbf{f}_{-a}||_2$.
Lad $B_i = [X_i \text{ fejler}]$ og lad B være antallet der fejler:

$$B = \sum_{i \in [t]} B_i$$

Da har vi, at hvis $X_{(\lceil t/2 \rceil)}$ fejler, så betyder det at $B \geq t/2$.
Da $\mathbb{P}[B_i = 1] \leq \frac{1}{4}$ må den forventede værdi af B være

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

Da kan vi beregne:

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

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.

Count-Min Sketch

Vi har næsten samme problem som før, med den forskel at vi antager at der for hvert par (j, Δ) i strømmen gælder $\Delta > 0$.

Algorithm 2: Count-Min Sketch

Init
$k = \lceil \frac{2}{\epsilon} \rceil$
$t = \lceil \lg \frac{1}{\delta} \rceil$
for $i \in [t]$
  $C_i[0, \ldots, k-1] = 0$
  h_i = strong-universel $h : [n] \to [k]$
Process (j, Δ)
for $i \in [t]$
  $C_i[h_i(j)] \mathrel{+}= \Delta$
Output
return $\hat{f}_a = \min_{i \in [t]} C_i[h_i(a)]$

Kvalitet af Count-Min Sketch estimatet

Vi ser tydeligt at enhver tæller $C_i[h_i(a)]$ korresponderende til en nøgle a er et overestimat af f_a. Derfor vil vi altid have:

$$f_a \leq \hat{f}_a$$

For en fikseret a analyserer vi nu overskuddet i én tæller, lad os sige $C_i[h_i(a)]$. Lad den stokastiske variabel X_i beskrive dette overskud. For $j \in [n]$ hvor $j \neq a$ lader vi $Y_{i,j} = [h_i(j) = h_i(a)]$. Bemærk at j bidrager til tælleren hvis og kun hvis $Y_{i,j} = 1$ og når det gør er bidraget f_j. Således

$$X_i = \sum_{\substack{j \in [n] \\ j \neq a}} f_j Y_{i,j}$$

På grund af h_i er stærk universel er $\mathbb{E}\left[Y_{i,j} = 1/k\right]$. Således får vi ved linearity of expectation:

$$\mathbb{E}\left[X_i\right] = \sum_{\substack{j\in[n]\\ j\neq a}} \frac{f_j}{k} = \frac{||\mathbf{f}||_1 - f_a}{k} = \frac{||\mathbf{f}_{-a}||_1}{k}$$

Siden alle $f_j \geq 0$ har vi $X_i \geq 0$, og kan derfor bruge Markovs ulighed til at få:

$$\mathbb{P}\left[X_i \geq \epsilon||\mathbf{f}_{-a}||_1\right] \leq \frac{\mathbb{E}\left[X_i\right]}{\epsilon||\mathbf{f}_{-a}||_1} = \frac{1}{k\epsilon} = \frac{1}{2}$$

pga. vores valg af k i algoritmen.

Flere tællere

Ovenstående sandsynlighed er for én tæller. Vi har t tællere som er uafhængige. Overskuddet i outputtet, $\hat{f}_a - f_a$, er minimum af overskuddet for alle X_i hvor $i \in [t]$. Derfor får vi:

$$\begin{aligned}
\mathbb{P}\left[\hat{f}_a - f_a \geq \epsilon||\mathbf{f}_{-a}||_1\right] &= \mathbb{P}\left[\min\{X_1, \ldots, X_t\} \geq \epsilon||\mathbf{f}_{-a}||_1\right] && (6)\\
&= \mathbb{P}\left[\forall i \in [t] \text{ gælder } X_i \geq \epsilon||\mathbf{f}_{-a}||_1\right] && (7)\\
&= \prod_{i=1}^{t} \mathbb{P}\left[X_i \geq \epsilon||\mathbf{f}_{-a}||_1\right] && (8)\\
&\leq \frac{1}{2^t} && (9)
\end{aligned}$$

Med vores valg af t i algoritmen bliver denne sandsynlighed højest δ. Derfor har vi vist at der med høj sandsynlighed gælder:

$$f_a \leq \hat{f}_a \leq f_a + \epsilon||\mathbf{f}_{-a}||_1$$

hvor uligheden til venstre altid holder, og den højre ulighed fejler med sandsynligheden højest δ.