L12 Streaming (Count-Min Sketch)

Tuesday, February 25, 2020 6:59 AM

- we have a stream $A = \langle a_1, a_2, \ldots, a_n \rangle$ where $a_i \in [m]$
- we only get one pass over the data
- we have small relative memory space
- while n, m are very large...
  - counts can be stored with $\log(n)$ bits
  - labels can be stored with $\log(m)$ bits.

Frequency Approximation:

- $f_j = |\{a \in A \mid a = j\}|$
- $F_1 = \sum_j f_j$ = total count n
- $F_2 = \sqrt{\sum f_j^2}$ typically $F_1 >> F_2$ (thus $\varepsilon F_1 >> \varepsilon F_2$)
- $F_0 = \sum_j f_j^0$ = # of unique items

Goal:

- $\forall j \in [m] \rightarrow \hat{f}_j$ such that $|f_j - \hat{f}_j| \leq \varepsilon F_1$ size $\frac{1}{\varepsilon}$ poly$(\log n, \log m)$
- Imagine some sketch $S(A)$
  - we would want a data structure with the following capabilities:
    - insert (a_i)
    - query $(q \in [m]) \Rightarrow \hat{f}_q$
      - thus there is a trade-off between space of $S(A)$ and accuracy.
- Count-Min Sketch: $f_j \leq \hat{f}_j \leq f_j + \varepsilon n$ always overstated, but has upperbound
- Count Sketch: $f_j - \varepsilon F_2 \leq \hat{f}_j \leq f_j + \varepsilon F_2$ within some tight bound of actual f_j

} true with $P(1-\delta)$

Count-Min Sketch:

- k counters where $k = 2/\varepsilon$
- t hash functions where $t = \log(1/\delta)$, $h_j : [m] \rightarrow [k]$

h_1	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	...	$c_{1,k}$
h_2	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	...	$c_{2,k}$
⋮	⋮				⋮
h_t	$c_{t,1}$	$c_{t,2}$	$c_{t,3}$	...	$c_{t,k}$

- imagine some element comes in and it gets mapped to some column
- then there's a counter there in the column

- the data structure needs to handle two operations
  - insert $a \in A$ where $a \in [m]$ } for each row hash to counter and increment
    - for $j = 1$ to t
      - $c_{j, h_j(a)}$ ++
  - query $q \in [m]$

- $C_{j, h_j(a)}$ ++

- query $q \in [m]$
  - $\hat{f}_q = \min_{j \in [t]} C_{j, h_j(q)}$ i.e. $\hat{f}_q$ is the minimum count over each hash function of q