Streaming Algorithms

Meng-Tsung Tsai

03/20/2018

Programming Assignment #0

The OJ (oj.nctu.me) is already. You need to submit your program for assignment #0 by 23:59, Mar 30.

If you didn't use the OJ before, you may read the guidelines on new e3.

References

- "Pairwise Independence and Derandomization," Luby and Wigderson (2005)
- "Sketch Techniques for Approximate Query Processing," Cormode
- "Epsilon Nets," Welzl https://www.ti.inf.ethz.ch/ew/lehre/CG12/lecture/Chapter%2015.pdf

Heavy Hitter

Problem Definition

Input: a sequence of n elements $e_1, e_2, ..., e_n$ where each e_i in $[U] = \{1, ..., U\}$.

Output: a set $S \subseteq [U]$ so that the frequency

$$f(k) = \sum_{i \in [n]} \mathbf{1}[e_i = k] \geq \varepsilon n.$$

In words, output a set S containing frequent elements.

Goal: using $o(n \log |U|)$ bits.

Using Count-Min Sketch or Count Sketch

Though we can estimate the frequency of all elements, iterating over the entire domain U may take extremely long time.

```
foreach (k in U){
    if( f̂(k) > εn){
        output k;
    }
}
```

If U is the set of all 64-bit integers, then it takes 2^{64} iterations.

Sampling-Based Method

Upon iterating over the incoming sequence, we sample a random subsequence. Specifically, we do:

```
foreach (incoming element e_i){
   S ← ∅;
   flip a coin that it heads up with probabiliy p;
   if(the coin heads up){
      S ← S ∪ {e_i};
   }
}
```

By Chernoff bound, $|S|$ has size in $[(1-\boldsymbol{\delta})\,pn, (1+\boldsymbol{\delta})\,pn]$ for some constant $\boldsymbol{\delta} > 0$ with probability at least $1\text{-}1/e^{\Omega(pn)}$.

Sampling-Based Method

For each subset A that has size at least εn.

$$\begin{aligned}\Pr[S \cap A = \varnothing] &\leq (1\text{-}p)^{|A|} \\ &\leq (1\text{-}p)^{\varepsilon n} \\ &\leq e^{-p\varepsilon n} \qquad \text{(note that } 1+x \leq e^x \text{ for all real x)}\end{aligned}$$

By the union bound, the probability that all frequent elements are contained in S is $1\text{-}1/(\varepsilon e^{p\varepsilon n})$. (Why?)

Sampling-Based Method

Thus, heavy hitter has an efficient implementation like:

foreach (k in S){
  if($\hat{f}(k)$> εn){
    output k;
  }
}

Check-on-Update Method (Count-Min)

foreach (incoming element e_i){
  update the data structure with e_i;
  if($\hat{f}(e_i)$ equals εn){
    // in the count-min sketch, this happens once for each e_i whose final $\hat{f}(e_i) \geq$ εn;
    output e_i;
  }
}

Issues

Both methods cannot be directly generalized to the cases when it is allowed to delete elements.

We will see in the lecture of L_p-sampler how to sample S even if deletions are allowed.

Minimum Enclosing Circle

Problem Definition

Input: a set P of n points $p_1, p_2, \ldots, p_n$ in $\mathbf{R}^2$.

Output: a circle C whose area is no more than that of the minimum enclosing circle of P so that at most εn points in P are not included in C for some constant $\varepsilon > 0$.

Problem Definition

Input: a set P of n points $p_1, p_2, \ldots, p_n$ in $\mathbf{R}^2$.

Output: a circle C whose area is no more than that of the minimum enclosing circle of P so that at most εn points in P are not included in C for some constant $\varepsilon > 0$.

Strategy. Sample a small set S of points, each point is included in S with probability p, and show that the minimum enclosing circle of S contains at least $(1-\varepsilon)n$ point with a good probability.

First Attempt

$\Pr[|P \cap MEC(S)| \geq (1-\varepsilon)\, n]$

$\geq 1 - \cup_{A \subseteq P, |A| \geq \varepsilon n} \Pr[|A \cap MEC(S)| > 0]$

$\geq 1 - 2^n \Pr[|A \cap S| > 0]$

$\geq 1 - 2^n (1/e^{\varepsilon p n}) < 0$

Second Attempt

$\Pr[|P \cap MEC(S)| \geq (1-\varepsilon)\, n]$

$\geq 1 - \cup_{A' = P \setminus MEC(A), |A'| \geq \varepsilon n} \Pr[|A' \cap MEC(S)| > 0]$

$\geq 1 - O(n^3) \Pr[|A' \cap S| > 0]$ (Why $O(n^3)$?)

$\geq 1 - O(n^3) (1/e^{\varepsilon p n}) > 0$ by setting $p = d \log n/n$ for some sufficiently large constant d

Exercise 1

Can we extend the method for the minimum enclosing circle to convex hull?

Input: a set P of n points $p_1, p_2, \ldots, p_n$ in $\mathbf{R}^2$.

Output: a convex polygon Q whose area is no more than that of the convex hull of P so that at most εn points for some constant $\varepsilon > 0$ are not included in Q.

Covering by Other Geometric Objects

If interested, check the 3rd reference. It shows the relationship between the VC dimension of geometric objects and epsilon-nets.

Let (X, R) be a range space (or set system, or a hypergraph) where

$$R \subseteq 2^X.$$

An epsilon-net is a set $A \subseteq X$ so that for every $r \in R$,

$$\text{if } (r \cap X) \geq \varepsilon\,|X|, \text{ then } (r \cap A) \neq \emptyset.$$

In the above examples, we use some kinds of epsilon-nets.