Lecture 4

Barna Saha

AT&T-Labs Research

September 19, 2013

Outline

Heavy Hitter Continued

Frequency Moment Estimation

Dimensionality Reduction

Heavy Hitter

- Heavy Hitter Problem: For $0 < \epsilon < \phi < 1$ find a set of elements S including all i such that $f_i > \phi m$ and there is no element in S with frequency $\leq (\phi - \epsilon)m$.
- Count-Min sketch guarantees: $f_i \leq \hat{f}_i \leq f_i + \epsilon m$ with probability $\geq 1 - \delta$ in space $\frac{e}{\epsilon} \log \frac{1}{(\phi - \epsilon)\delta}$.
- Insert only: Maintain a min-heap of size $k = \frac{1}{\phi - \epsilon}$, when an item arrives estimate frequency and if above ϕm include it in the heap. If heap size more than k, discard the minimum frequency element in the heap.

Heavy Hitter

- Turnstile model:
  - Maintain dyadic intervals over binary search tree and maintain $\log n$ count-min sketch with using space $\frac{e}{\epsilon} \log \frac{2 \log n}{\delta(\phi-\epsilon)}$ one for each level.
  - At every level at most $\frac{1}{\phi}$ heavy hitters.
  - Estimate frequency of children of the heavy hitter nodes until leaf-level is reached.
  - Return all the leaves with estimated frequency above ϕm.

  - Analysis
  - At most $\frac{2}{\phi-\epsilon}$ nodes at every level is examined.
  - Each true frequency $> (\phi - \epsilon)m$ with probability at least $1 - \frac{\delta(\phi-\epsilon)}{2 \log n}$.
  - By union bound all true frequencies are above $(\phi - \epsilon)m$ with probability at least $1 - \delta$.

l_2 frequency estimation

- $|f_i - \hat{f}_i| \leq \pm\epsilon\sqrt{f_1^2 + f_2^2 + ...f_n^2}$ [Count-sketch]
- $F_2 = f_1^2 + f_2^2 + ...f_n^2$
- How do we estimate F_2 in small space ?

AMS-F_2 Estimation

- $\mathcal{H} = \{h : [n] \to \{+1, -1\}\}$ four-wise independent hash functions
- Maintain $Z_j = Z_j + ah_j(i)$ on arrival of (i, a) for $j = 1, ..., t = \frac{c}{\epsilon^2}$
- Return $Y = \frac{1}{t} \sum_{j=1}^{t} Z_j^2$

Analysis

- $Z_j = \sum_{i=1}^{n} f_i h_j(i)$
- $\mathrm{E}\left[Z_j\right] = 0$, $\mathrm{E}\left[Z_j^2\right] = F_2$.
- $\mathrm{Var}\left[Z_j^2\right] = \mathrm{E}\left[Z_j^4\right] - (\mathrm{E}\left[Z_j\right])^2 \leq 4F_2^2$.
- $\mathrm{E}\left[Y\right] = F_2$. $\mathrm{Var}\left[Y\right] = \frac{1}{t^2} \sum_{j=1}^{t} Var(Z_j^2) = \frac{4\epsilon^2}{c} F_2^2$
- By Chebyshev Inequality $\Pr\left[|Y - \mathrm{E}\left[Y\right]| > \epsilon F_2\right] \leq \frac{4}{c}$

Boosting by Median

- Keep $Y_1, Y_2, \ldots Y_s, s = O(\log 1\delta)$
- Return $A = \mathit{median}(Y_1, Y_2, .., Y_s)$
- By Chernoff bound $\Pr\Big[|A - F_2| > \epsilon F_2\Big] < \delta$

Linear Sketch

- Algorithm maintains a linear sketch $[Z_1, Z_2, ..., Z_t]\mathbf{x} = R\mathbf{x}$ where R is a $t \times n$ random matrix with entries $\{+1, -1\}$.
- Use $Y = ||Rx||_2^2$ to estimate $t||x|_2^2$. $t = O(\frac{1}{\epsilon^2})$.
- Streaming algorithm operating in the sketch model can be viewed as **dimensionality reduction** technique.

Dimensionality Reduction

- Streaming algorithm operating in the sketch model can be viewed as **dimensionality reduction** technique.
  - stream S: point in n dimensional space, want to compute $l_2(S)$
  - sketch operator can be viewed as an approximate embedding of l_2^n to sketch space $\mathcal{C}$ such that
    1. Each point in $\mathcal{C}$ can be described using only small number (say m) of numbers so $C \subset \mathbb{R}^m$ and
    2. value of $l_2(S)$ is approximately equal to $F(C(S))$.
  - $F(Y_1, Y_2, ..Y_t) = \text{median}(Y_1, Y_2, .., Y_t)$

Dimensionality Reduction

- $F(Y_1, Y_2, ..Y_t) = \text{median}(Y_1, Y_2, .., Y_t)$
- Disadvantage: F is not a norm–performing any nontrivial operations in the sketch space (e.g. clustering, similarity search, regression etc.) becomes difficult.
- Can we embed from l_2^n to l_2^m, $m << n$ approximately preserving the distance ? **Johnson-Lindenstrauss Lemma**

Interlude to Normal Distribution

Normal distribution $\mathcal{N}(0,1)$:

- Range $(-\infty, \infty)$
- Density $f(x) = e^{-x^2}/\sqrt{2\pi}$
- Mean=0, Variance=1

Basic facts

- If X and Y are independent random variables with normal distribution then so is $X + Y$
- If X and Y are independent with mean 0 then $\mathrm{E}\left[[X+Y]^2\right] = \mathrm{E}\left[X^2\right] + \mathrm{E}\left[Y^2\right]$
- $\mathrm{E}\left[cX\right] = c\mathrm{E}\left[X\right], \mathrm{Var}\left[cX\right] = c^2\mathrm{Var}\left[X\right]$

A Different Linear Sketch

Instead of ± 1 let r_i be a i.i.d. random variable from $\mathcal{N}(0,1)$.

- Consider $Z = \sum_i r_i x_i$
- $\mathrm{E}\left[Z^2\right] = \mathrm{E}\left[(\sum_i r_i x_i)^2\right] = \sum_i \mathrm{E}\left[r_i^2\right] x_i^2 = \sum_i \mathrm{Var}\left[r_i\right] x_i^2 = \sum_i x_i^2 = ||x||_2^2$.
- As before we maintain $Z = [Z_1, Z_2, ..., Z_t]$ and define $Y = ||Z||_2^2$
- $\mathrm{E}\left[Y\right] = t||x||_2^2$
- We show that there exists constant $C > 0$ s.t. for small enough $\epsilon > 0$

$$\Pr\left[|Y - t||x||_2^2| > \epsilon t||x||_2^2\right] \leq e^{-C\epsilon^2 t} \text{ (JL lemma)}$$

- set $t = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$

Johnson Lindenstrauss Lemma

Lemma

For any $0 < epsilon < 1$ *and any integer* m, *let* t *be a positive integer such that*

$$t > \frac{4 \ln m}{\epsilon^2/2 + \epsilon^3/3}$$

Then for any set V *of* m *points in* R^n, *there is a map* $f : R^n \to R^t$ *such that for all* u *and* $v \in V$,

$$(1-\epsilon)||u-v||_2^2 \leq ||f(u)-f(v)||_2^2 \leq (1+\epsilon)||u-v||_2^2.$$

Furthermore this map can be found in randomized polynomial time.