Lecture 7

Barna Saha

AT&T-Labs Research

September 26, 2013

Outline

Sampling

Estimating F_k [AMS'96]

Reservoir Sampling

Priority Sampling

Estimating F_k

- Suppose, you know m, the stream length
- Sample a index p uniformly and randomly with probability $\frac{1}{m}$. Suppose $a_p = l$
- Compute $r = |\{q : q \geq p, a_q = l\}|$–the number of occurrences of l in the stream starting from a_p
- Return $X = m(r^k - (r-1)^k)$
- Show $\mathrm{E}\left[X\right] = F_k$, $\mathrm{Var}\left[X\right] \leq n^{1-\frac{1}{k}}(F_k)^2$.

Estimating F_k

- Maintain $s_1 = O(\frac{kn^{1-\frac{1}{k}}}{\epsilon^2})$ such estimates $X_1, X_2, ..., X_{s_1}$. Take the average, $Y = \frac{1}{s_1}\sum_{i=1}^{s_1} X_i$.
- Maintain $s_2 = O(\log \frac{1}{\delta})$ of these average estimates, $Y_1, Y_2, ..., Y_{s_2}$ and take the median.
- Follows $(1 \pm \epsilon)$ approximation with probability $\geq (1 - \delta)$.

Estimating F_k

Lemma

$E[X] = F_k$

$$\begin{aligned}
E[Y] &= \sum_{i=1}^{n} \sum_{j=1}^{f_i} E[X \mid i \text{ is sampled on } j\text{th occurrence}] \frac{1}{m} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{f_i} m((f_i - j + 1)^k - (f_i - j)^k) \frac{1}{m} \\
&= \sum_{i=1}^{n} \left[1^k + (2^k - 1^k) + (3^k - 2^k) + \ldots + (f_i^k - (f_i - 1)^k) \right] \\
&= F_k
\end{aligned}$$

Estimating F_k

Lemma

$\mathrm{Var}\left[X\right] \leq kn^{1-\frac{1}{k}}(F_k)^2$

$$\mathrm{E}\left[Y^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{f_i}\mathrm{E}\left[X^2 \mid i \text{ is sampled on } j\text{th occurrence}\right]\frac{1}{m}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{f_i} m^2((f_i - j + 1)^k - (f_i - j)^k)^2\frac{1}{m}$$

$$= m\sum_{i=1}^{n}\left[1^{2k} + (2^k - 1^k)^2 + (3^k - 2^k)^2 + \ldots + (f_i^k - (f_i - 1)^k)^2\right]$$

$$\leq m\sum_{i=1}^{n} k1^{2k-1} + k2^{k-1}(2^k - 1^k) + \ldots.. + f_i^{k-1}(f_i^k - (f_i - 1)^k)$$

Using $a^k - b^k = (a-b)(a^{k-1} + ba^{k-2} + .. + b^{k-1}) \leq (a-b)ka^{k-1}$

Estimating F_k

$$
\begin{aligned}
& m \sum_{i=1}^{n} k1^{2k-1} + k2^{k-1}(2^k - 1^k) + \ldots.. + f_i^{k-1}(f_i^k - (f_i - 1)^k) \\
< \quad & mk \sum_{i=1}^{n} 1^{2k-1} + 2^{2k-1} + \ldots + f_i^{2k-1} = mkF_{2k-1} \\
= \quad & kF_1 F_{2k-1} \leq kn^{1-\frac{1}{k}} \left(\sum_{i=1}^{n} f_i^k \right)^2 = kn^{1-\frac{1}{k}} (F_k)^2
\end{aligned}
$$

Reference: The space complexity of approximating the frequency moment by Alon, Matias, Szegedy.

Uniform Random Sample from Stream Without Replacement

- What happens when you do not know m ?

Check out: Algorithms Every Data Scientist Should Know: Reservoir Sampling http://blog.cloudera.com/blog/2013/04/hadoop-stratified-randosampling-algorithm/

Reservoir Sampling

- Find a uniform sample s from stream if you do not know m ?
- Initially $s = a_1$
- On seeing the t-th element set $s = a_t$ with probability $\frac{1}{t}$

$$\Pr\left[s = a_i\right] = \frac{1}{i}\left(1 - \frac{1}{i+1}\right)\left(1 - \frac{1}{i+2}\right)\ldots\left(1 - \frac{1}{t}\right) = \frac{1}{t}$$

- Can you extend AMS algorithm to a single pass now ?

Reservoir Sampling of size k

- Find a uniform sample s of size k from stream if you do not know m ?

- Initially $s = \{a_1, a_2, ..., a_k\}$
- On seeing the t-th element set, pick a number $r \in [1, t]$ uniformly and randomly
- If $r \leq k$, replace the rth element by a_t

$$\Pr\left[a_i \in s\right] = \frac{k}{i}\left(1 - \frac{1}{i+1}\right)\left(1 - \frac{1}{i+2}\right) \ldots \left(1 - \frac{1}{t}\right) = \frac{k}{t}$$

Priority Sampling

- Element i has weight w_i.
- Keep a sample of size k such that any subset sum query can be answered later.

- Uniform Sampling: Misses few heavy hitters
- Weighted Sampling with Replacements: duplicates of heavy hitters
- Weighted Sampling Without Replacement: Very complicated expression-does not work for subset sum

Priority Sampling

- For each item $i = 0, 1, .., n-1$ generate a random number $\alpha_i \in [0, 1]$ uniformly and randomly.
- Assign priority $q_i = \frac{w_i}{\alpha_i}$ to the *ith* element.
- Select the k highest priority items in the sample S.

Priority Sampling

- Let τ be the priority of the $(k+1)$th highest priority.
- Set $\hat{w}_i = \max(w_i, \tau)$ if i is in the sample and 0 otherwise.
- $\mathrm{E}\left[\hat{w}_i\right] = w_i$

Priority Sampling

- $A(\tau')$:Event τ' is the kth highest priority among all $j \neq i$.
- For any value of τ',
  $\mathrm{E}\big[\hat{w}_i \mid A(\tau')\big] = \Pr\big[i \in S \mid A(\tau')\big] \max(w_i, \tau')$
- $\Pr\big[i \in S \mid A(\tau')\big] = \Pr\big[\frac{w_i}{\alpha_i} > \tau'\big] = \Pr\big[\alpha_i < \frac{w_i}{\tau'}\big] = \min(1, \frac{w_i}{\tau'})$
- $\mathrm{E}\big[\hat{w}_i \mid A(\tau')\big] = \max(w_i, \tau') \min(1, \frac{w_i}{\tau'}) = w_i$
- Holds for all τ', hence holds unconditionally.

Priority Sampling

- Near optimality: variance of the weight estimator is minimal among all $k + 1$-sparse unbiased estimators.