Mining Data Streams-Estimating Frequency Moment

Barna Saha

February 18, 2016

Frequency Moment

- Computing "moments" involves distribution of frequencies of different elements in the stream.

Frequency Moment

- Computing "moments" involves distribution of frequencies of different elements in the stream.
- Let f_i be the number of occurrences of the ith element for any $i \in [1, n]$, then the kth frequency moment is $F_k = \sum_i f_i^k$

Frequency Moment

- The 0th moment is the sum of 1 for each $f_i > 0$. Hence it counts the number of distinct items.

Frequency Moment

- The 0th moment is the sum of 1 for each $f_i > 0$. Hence it counts the number of distinct items.
- The 1st moment is the sum of the f_is which must be the length of the stream. This is easy to calculate.

Frequency Moment

- The 0th moment is the sum of 1 for each $f_i > 0$. Hence it counts the number of distinct items.
- The 1st moment is the sum of the f_is which must be the length of the stream. This is easy to calculate.
- The 2nd moment is the sum of the squares of the f_i's. It is sometimes called the *surprise number* as it measures the unevenness of the distribution of elements.
  - Suppose we have a stream of length 100.
  - Scenario 1: There are 10 elements each with frequency 10. $F_2 = 10 * 10^2 = 1000$
  - Scenario 2: There are 10 elements, 1st item has frequency 91, and rest have each frequency 1. $F_2 = 91^2 + 9 * 1^2 = 8290$.

Computing F_2 in Small Space

- Linear Sketching
- Alon-Matias-Szegedy Sampling (read Sec 4.5 Leskovec et al.)

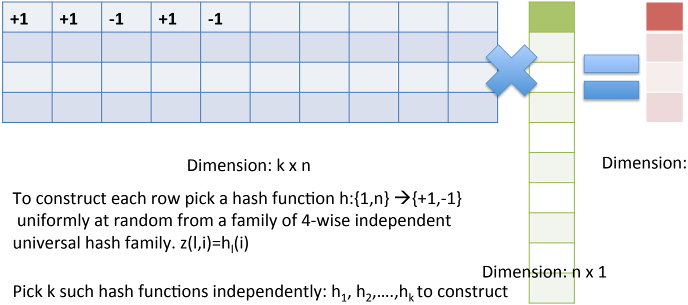
Linear Sketch for F_2

- Problem Given a stream $A_1, A_2, .., A_m$ where elements are coming from the universe $[1, n]$ estimate $F_2 = \sum_{i=1}^{n} f_i^2$ in "small space".
- Output Return an estimate $\hat{F}_2$ such that

$$\Pr\left(F_2(1-\epsilon) \leq \hat{F}_2 \leq (1+\epsilon)F_2\right) \geq (1-\delta)$$

where $\epsilon > 0$ and $\delta > 0$ are respectively the error and confidence parameters.

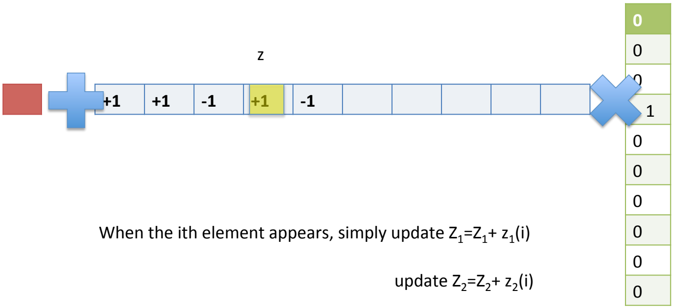
Linear Sketch for F_2



To construct each row pick a hash function $h:\{1,n\} \rightarrow \{+1,-1\}$ uniformly at random from a family of 4-wise independent universal hash family. $z(l,i)=h_l(i)$

Pick k such hash functions independently: $h_1, h_2, \ldots, h_k$ to construct The k rows.

Linear Sketch for F_2



When the ith element appears, simply update $Z_1 = Z_1 + z_1(i)$

update $Z_2 = Z_2 + z_2(i)$

update $Z_3 = Z_3 + z_3(i)$

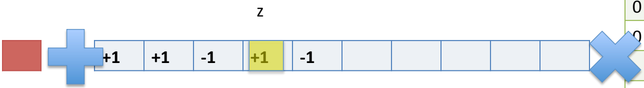
:

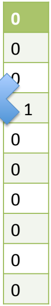
:

:

update $Z_k = Z_k + z_k(i)$

Linear Sketch for F_2





When the ith element appears, simply update $Z_1=Z_1+ z_1(i)$

update $Z_2=Z_2+ z_2(i)$

update $Z_3=Z_3+ z_3(i)$

:

:

:

update $Z_k=Z_k+ z_k(i)$

Space Requirement=klog(n)

Estimate=$(Z_1^2+Z_2^2+....+Z_k^2)/k$

Estimate: $\hat{F}_2 = \frac{1}{k} \sum_{i=1}^{k} Z_i^2$

- Why is this a good estimate?

Estimate: $\hat{F}_2 = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$

- Why is this a good estimate?
- Show $E[\hat{F}_2] = F_2$.

Estimate: $\hat{F}_2 = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$

- Why is this a good estimate?
- Show $E[\hat{F}_2] = F_2$.
- Show $Var[\hat{F}_2] \leq \frac{2F_2^2}{k}$.

Estimate: $\hat{F}_2 = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$

- Why is this a good estimate?
- Show $E[\hat{F}_2] = F_2$.
- Show $Var[\hat{F}_2] \leq \frac{2F_2^2}{k}$.
- Apply Chebyshev.

$$Prob\left(|\hat{F}_2 - F_2| > \epsilon F_2\right) \leq \frac{Var(\hat{F}_2)}{\epsilon^2 F_2^2}$$

Estimate: $\hat{F}_2 = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$

- Why is this a good estimate?
- Show $E[\hat{F}_2] = F_2$.
- Show $Var[\hat{F}_2] \leq \frac{2F_2^2}{k}$.
- Apply Chebyshev.

$$Prob\left(|\hat{F}_2 - F_2| > \epsilon F_2\right) \leq \frac{Var(\hat{F}_2)}{\epsilon^2 F_2^2}$$

- Take $k = \frac{16}{\epsilon^2}$. $Prob\left(|\hat{F}_2 - F_2| > \epsilon F_2\right) \leq \frac{1}{8}$

Estimate: $\hat{F}_2 = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$

- Why is this a good estimate?
- Show $E[\hat{F}_2] = F_2$.
- Show $Var[\hat{F}_2] \leq \frac{2F_2^2}{k}$.
- Apply Chebyshev.

$$Prob\left(|\hat{F}_2 - F_2| > \epsilon F_2\right) \leq \frac{Var(\hat{F}_2)}{\epsilon^2 F_2^2}$$

- Take $k = \frac{16}{\epsilon^2}$. $Prob\left(|\hat{F}_2 - F_2| > \epsilon F_2\right) \leq \frac{1}{8}$
- $$Prob\left(F_2(1-\epsilon) \leq \hat{F}_2 \leq (1+\epsilon)F_2\right) \geq \frac{7}{8}$$

Expectation of Z_s^2

$Z_s \sim Z, s = 1, 2, .., k$

- $Z = \sum_{i=1}^{n} f_i z(i),\ Z^2 = \sum_{i,j \in [1,n]} f_i f_j z_i z_j$

Expectation of Z_s^2

$Z_s \sim Z, s = 1, 2, .., k$

- $Z = \sum_{i=1}^{n} f_i z(i)$, $Z^2 = \sum_{i,j \in [1,n]} f_i f_j z_i z_j$
- $E[Z^2] = \sum_{i,j \in [1,n]} E[f_i f_j z_i z_j] = \sum_i E[f_i^2 z_i^2] = \sum_i f_i^2 = F_2$

since $E[z_i z_j] = 0$ if $i \neq j$ and $E[z_i^2] = 1$.

Expectation of Z_s^2

$Z_s \sim Z, s = 1, 2, .., k$

- $Z = \sum_{i=1}^{n} f_i z(i)$, $Z^2 = \sum_{i,j \in [1,n]} f_i f_j z_i z_j$
- $E[Z^2] = \sum_{i,j \in [1,n]} E[f_i f_j z_i z_j] = \sum_i E[f_i^2 z_i^2] = \sum_i f_i^2 = F_2$

  since $E[z_i z_j] = 0$ if $i \neq j$ and $E[z_i^2] = 1$.
- $$E[\hat{F}_2] = \frac{1}{k} \sum_{s=1}^{k} E[Z_s^2] = F_2$$

Variance of Z_s^2

- $Var(Z^2) = E[Z^4] - (E[Z^2])^2$

Variance of Z_s^2

- $Var(Z^2) = E[Z^4] - (E[Z^2])^2$
- $$E[Z^4] = \sum_i f_i^4 E[z_i^4] + \sum_{i,j:i<j} \binom{4}{2} f_i^2 f_j^2 E[z_i^2 z_j^2]$$
  $$= \sum_i f_i^4 + 6 \sum_{i,j:i<j} f_i^2 f_j^2$$

  since $E[z_i z_j z_k z_l] = 0$ if $i < j < k < l$ or 3 of the terms are equal.

Variance of Z_s^2

- $Var(Z^2) = E[Z^4] - (E[Z^2])^2$
- $$E[Z^4] = \sum_i f_i^4 E[z_i^4] + \sum_{i,j:i<j} \binom{4}{2} f_i^2 f_j^2 E[z_i^2 z_j^2]$$
$$= \sum_i f_i^4 + 6 \sum_{i,j:i<j} f_i^2 f_j^2$$

  since $E[z_i z_j z_k z_l] = 0$ if $i < j < k < l$ or 3 of the terms are equal.
- $$(E[Z^2])^2 = \left(\sum_i f_i^2\right)^2 = \sum_i f_i^4 + 2 \sum_{i,j:i<j} f_i^2 f_j^2$$

Variance of Z_s^2

- $Var(Z^2) = E[Z^4] - (E[Z^2])^2$
- $$E[Z^4] = \sum_i f_i^4 E[z_i^4] + \sum_{i,j:i<j} \binom{4}{2} f_i^2 f_j^2 E[z_i^2 z_j^2]$$
  $$= \sum_i f_i^4 + 6 \sum_{i,j:i<j} f_i^2 f_j^2$$
  since $E[z_i z_j z_k z_l] = 0$ if $i < j < k < l$ or 3 of the terms are equal.
- $$(E[Z^2])^2 = \left(\sum_i f_i^2\right)^2 = \sum_i f_i^4 + 2 \sum_{i,j:i<j} f_i^2 f_j^2$$
- $$Var(Z^2) = 4 \sum_{i,j:i<j} f_i^2 f_j^2 \le 2F_2^2$$

Variance of $\hat{F}_2$

$$
\begin{aligned}
Var(\hat{F}_2) &= Var(\frac{1}{k}\sum_{s=1}^{k} Z_s^2) \\
&= \frac{1}{k^2} Var(\sum_{s=1}^{k} Z_s^2)) \text{ since } Var(aX) = a^2 Var(X) \text{ for any constant } \\
&= \frac{1}{k^2}\sum_{s=1}^{k} Var(Z_s^2) \leq \frac{1}{k^2} 2kF_2^2 = \frac{2F_2^2}{k}
\end{aligned}
$$

Boosting Confidence by Median

- We have

$$Prob\left(F_2(1-\epsilon) \leq \hat{F}_2 \leq (1+\epsilon)F_2\right) \geq \frac{7}{8}$$

- We want

$$Prob\left(F_2(1-\epsilon) \leq \hat{F}_2 \leq (1+\epsilon)F_2\right) \geq 1-\delta$$

Boosting Confidence by Median

- We have

$$Prob\left(F_2(1-\epsilon) \le \hat{F_2} \le (1+\epsilon)F_2\right) \ge \frac{7}{8}$$

- We want

$$Prob\left(F_2(1-\epsilon) \le \hat{F_2} \le (1+\epsilon)F_2\right) \ge 1-\delta$$

- Take t independent estimates
  $H_1 = \hat{F_2}^1, H_2 = \hat{F_2}^2, ..., H_t = \hat{F_2}^t$

Boosting Confidence by Median

- We have

$$Prob\left(F_2(1-\epsilon) \leq \hat{F}_2 \leq (1+\epsilon)F_2\right) \geq \frac{7}{8}$$

- We want

$$Prob\left(F_2(1-\epsilon) \leq \hat{F}_2 \leq (1+\epsilon)F_2\right) \geq 1-\delta$$

- Take t independent estimates $H_1 = \hat{F}_2^{\,1}, H_2 = \hat{F}_2^{\,2}, ..., H_t = \hat{F}_2^{\,t}$
- Return the median of H_1, H_2,...,H_t.

Boosting by Median

- Suppose there is an Algorithm that returns an estimate $\hat{F}$ of a true estimate F such that $|\hat{F} - F|$ is small with probability $\frac{7}{8}$.
- How can we design an algorithm that will return an estimate G of F such that $|G - F|$ is small with probability $99/100$? (In general $1 - \delta$)

Boosting by Median

- Suppose there is an Algorithm that returns an estimate $\hat{F}$ of a true estimate F such that $|\hat{F} - F|$ is small with probability $\frac{7}{8}$.
- How can we design an algorithm that will return an estimate G of F such that $|G - F|$ is small with probability $99/100$? (In general $1 - \delta$)
- Run $s = 6 \log \frac{1}{\delta}$ independent copies of the Algorithm to obtain estimates $\hat{F}^1, \hat{F}^2, ..., \hat{F}^s$. Set $G = median_{i=1}^{s} \hat{F}^i$.

Boosting by Median

- What is the probability that the median is a bad estimate?

Boosting by Median

- What is the probability that the median is a bad estimate?
- Either all $\lfloor \frac{s}{2} \rfloor$ copies with estimate below G are bad or, $\lfloor \frac{s}{2} \rfloor$ copies with estimate above G are bad. That is there are $3 \log \frac{1}{\delta}$ copies that are at least bad for G to be a bad estimate.

Boosting by Median

- What is the probability that the median is a bad estimate?
- Either all $\lfloor \frac{s}{2} \rfloor$ copies with estimate below G are bad or, $\lfloor \frac{s}{2} \rfloor$ copies with estimate above G are bad. That is there are $3 \log \frac{1}{\delta}$ copies that are at least bad for G to be a bad estimate.
- Define an indicator random variable X_i which is 1 if the ith estimate $\hat{F}_i$ is bad. Then $E[X_i] = \frac{1}{8}$.

Boosting by Median

- What is the probability that the median is a bad estimate?
- Either all $\lfloor \frac{s}{2} \rfloor$ copies with estimate below G are bad or, $\lfloor \frac{s}{2} \rfloor$ copies with estimate above G are bad. That is there are $3 \log \frac{1}{\delta}$ copies that are at least bad for G to be a bad estimate.
- Define an indicator random variable X_i which is 1 if the ith estimate $\hat{F}_i$ is bad. Then $E[X_i] = \frac{1}{8}$.
- Then the number of bad estimates is $Y = \sum_i X_i$. and $E[Y] = \frac{6 \log \frac{1}{\delta}}{8} = \frac{3}{4} \log \frac{1}{\delta}$

Boosting by Median

- What is the probability that the median is a bad estimate?
- Either all $\lfloor \frac{s}{2} \rfloor$ copies with estimate below G are bad or, $\lfloor \frac{s}{2} \rfloor$ copies with estimate above G are bad. That is there are $3 \log \frac{1}{\delta}$ copies that are at least bad for G to be a bad estimate.
- Define an indicator random variable X_i which is 1 if the ith estimate $\hat{F}_i$ is bad. Then $E[X_i] = \frac{1}{8}$.
- Then the number of bad estimates is $Y = \sum_i X_i$. and $E[Y] = \frac{6 \log \frac{1}{\delta}}{8} = \frac{3}{4} \log \frac{1}{\delta}$
- Bound

$$Prob(Y > 3 \log \frac{1}{\delta})$$

using Chernoff's bound.

Boosting by Median

- Upper Tail version of Chernoff Bound. For $\epsilon > 1$

$$Prob(Y > E[Y](1+\epsilon)) \leq e^{-\frac{E[Y]\epsilon^2}{2+\epsilon}}$$

.

Boosting by Median

- Upper Tail version of Chernoff Bound. For $\epsilon > 1$
$$Prob(Y > E[Y](1+\epsilon)) \leq e^{-\frac{E[Y]\epsilon^2}{2+\epsilon}}$$
.
- $$Prob\left(Y > 3\log\frac{1}{\delta}\right) = Prob\left(Y > \frac{3}{4}\log\frac{1}{\delta}(1+3)\right)$$
$$\leq e^{-\frac{3}{4}\left(\log\frac{1}{\delta}\right)9\frac{1}{5}} < \delta$$

Versions of Chernoff Bound

Reference:
https://www.cs.princeton.edu/courses/archive/fall09/cos521/Handouts/probabilityandcomputing.pdf

Frequency Moment

- For $k > 2$, the best bound known is $\tilde{O}(n^{1-\frac{2}{k}} \log \frac{1}{\delta})$ barring $poly(\frac{1}{\epsilon})$ factor. There is an almost matching lower bound of $\Omega(n^{1-\frac{2}{k}})$.
- For $k < 2$, the best bound known is $\tilde{O}(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$.
- The algorithms use clever combination of sketching and hashing

Sketching as a Versatile Tool

- Estimating entropy, quantiles, heavy hitters, fitting histograms etc.
- Applications beyond streaming: dimensionality reduction, nearest neighbors, anomaly detection, statistics over social network.
- Not only useful for small-space algorithm design, but also for fast running time, distributed processing etc.

Sketching as a Versatile Tool

A different linear sketch

- Instead of ±1, let r_i be i.i.d. random variables from N(0,1)
- Consider

$$Z=\sum_i r_i x_i$$

- We still have that $E[Z^2] = \sum_i x_i^2 = ||x||_2^2$, since:
  - $E[r_i] E[r_j] = 0$
  - $E[r_i^2]$ = variance of r_i, i.e., 1
- As before we maintain $\mathbf{Z}=[Z_1 \dots Z_k]$ and define

  $Y = ||\mathbf{Z}||_2^2 = \sum_j Z_j^2$ (so that $E[Y]=k||x||_2^2$)
- We show that there exists $C>0$ s.t. for small enough $\varepsilon>0$

$$\Pr[\, | Y - k||x||_2^2 | > \varepsilon k||x||_2^2] \leq \exp(-C \varepsilon^2 k)$$

Slide from Piotr Indyk's course on Streaming, Sketching and Compressed Sensing

Sliding Window Model

- Only the last W items matter where W is the window size.

Sliding Window Model

- Only the last W items matter where W is the window size.
- Can you extend Bloom Filter, FM sketch in this setting?

Sliding Window Model

- Only the last W items matter where W is the window size.
- Can you extend Bloom Filter, FM sketch in this setting?
- Can you extend Count-Min sketch or linear sketching techniques in this setting?

Decaying Window Model

- No fixed window size, but older items have less importance.

Decaying Window Model

- No fixed window size, but older items have less importance.
- Can you extend Bloom Filter, FM sketch in this setting?

Decaying Window Model

- No fixed window size, but older items have less importance.
- Can you extend Bloom Filter, FM sketch in this setting?
- Can you extend Count-Min sketch or linear sketching techniques in this setting?