Data Stream Algorithms

Basic definitions

- Stream: n elements from universe $[m] = \{1, 2, \dots, m\}$.

  E.g: Consider $[1000]$

  $\{x_1, x_2, \dots, x_n\} = 3, 5, 7, 100, \dots$
- Goal: Compute a function of stream,

  E.g. Median, number of distinct elements, longest increasing sequence.
- But: – limited working memory, (usually sublinear in n and m, i.e. $O(\log n)$ or $O(\log m)$).
  – Access data sequentially.
  – Process quickly.

Why do ~~we~~ care?

Faster network, cheaper data storage, ...

* Sampling: a general technique to tackling massive amounts of data.

E.g we have a large list of all queries made to a search engine, and we want to measure how many queries contain the word "~~data~~ iPhone XS". Easy! just count them ?!! But we can actually do it faster. → sampling.

Problem: Given a large set of N elements U, $(|U| = N)$, select a subset of elements $\hat{U}$ $(|\hat{U}| \approx n)$ such that from $\hat{U}$ the size of any subset $S \subset U$ can be estimated.

- Sampling approach: Pick each element from U independently into set $\hat{U}$ with probability $p = \frac{n}{N}$

Let the variable X_i be 1 if element i is picked and 0 otherwise.

The number of picked elements is $\sum_{i=1}^{N} X_i$ and its expectation is

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbb{E}[X_i] = \sum_{i=1}^{N} \frac{n}{N} = n.$$

Let $\hat{S}$ be the set of the intersection of S and $\hat{U}$.

$$\hat{S} = S \cap \hat{U}$$

Let Δ_i be 1 if $i \in S$ and 0 otherwise

indicator function.

Let $Z = \frac{N}{n}|\hat{S}|$ be our estimator of $|S|$.

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{N}{n}|\hat{S}|\right] = \frac{N}{n}\mathbb{E}\left[\sum_{i=1}^{N} X_i \Delta_i\right] = \frac{N}{n}\mathbb{E}\left[\sum_{j \in S} X_j\right] = |S|$$

$i \in \hat{S}$ if and only if $X_i \Delta_i = 1$.

$\therefore \mathbb{E}[Z] = |S|$

~~The question~~:

Question: How close is Z to $|S|$?

Chernoff's bound would help!

Lemma (Chernoff bound): Let $X_1, \dots, X_n$ be independent Bernoulli random variables $\mathbb{P}(X_i = 1) = p_i$ and $\mathbb{P}(X_i = 0) = 1 - p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}(X > (1+\varepsilon)\mu) \leq e^{-\mu\varepsilon^2/4}$$

$$\mathbb{P}(X < (1-\varepsilon)\mu) \leq e^{-\mu\varepsilon^2/2}$$

$$\mathbb{P}(|X - \mu| \geq \varepsilon\mu) \leq 2e^{-\mu\varepsilon^2/4}.$$

Recall that in our problem, we want to ~~know how~~ go over the elements of S and count how many of them were samped into $\hat{U}$, and that

$$\frac{n}{N} Z = \sum_{i=1}^{n} X_i \Delta_i = \cancel{\sum_{j=}} \sum_{j \in S} X_j \Rightarrow Z = \frac{N}{n} \sum_{j \in S} X_j.$$

and $\frac{n}{N}\mathbb{E}[Z] = |S|$

Applying Chernoff's for $\sum_{j \in S} X_j$; $\mathbb{P}\left(\left|\sum_{j \in S} X_j - \frac{n}{N}|S|\right| > \varepsilon \frac{n|S|}{N}\right) \leq 2e^{-|S|n\varepsilon^2/4N}$

$\therefore \mathbb{P}\left(\left|\frac{N}{n}\sum_{j \in S} X_j - |S|\right| > \varepsilon|S|\right) \leq 2\exp\left(-\frac{|S|n\varepsilon^2}{4N}\right)$

$$\cancel{\begin{cases} \mathbb{P}(Z > (1+\varepsilon)|S|) \leq e^{-|S|n\varepsilon^2/4N} \\ \mathbb{P}(Z < (1-\varepsilon)|S|) \leq e^{-|S|n\varepsilon^2/2N} \end{cases}}$$

union bound

$$\Rightarrow \mathbb{P}(|Z - |S|| > \varepsilon|S|) \leq 2e^{-|S|n\varepsilon^2/4N}.$$

what does it mean?

For example, if $|S|$ is of the size $10^{-5}N$ and we want to have a 10% accuracy with probability at least 0.99, we must keep a sample of roughly 10^{8} elements, regardless of N.

big number but think of N a really big number, it's still small!

* Frequency moments of Data stream:

Given a data stream $a_1, a_2, \dots, a_n$ of length n, where each $a_j \in \{1, 2, \dots, m\} =: [m]$. The frequency of $i \in [m]$ in the stream is $f_i = |\{ j \mid a_j = i \}|$.
The vector $\vec{f} = (f_1, f_2, \dots, f_m)$ is called the frequency vector.

For $p \geq 0$, The pth frequency moment of the input is defined as follows:

$$F_p = \begin{cases} |\{ i \mid f_i \neq 0 \}| & \text{if } p = 0 \\ \max_i f_i & \text{if } p = \infty \\ \|f\|_p^p = \sum_{i=1}^{m} f_i^p & \text{otherwise} \end{cases}$$

number of distinct symbols occuring in the stream

- For $p = 1$, the first frequency moment is just n, the length of the string.
- For $p = 2$, the second frequency moment is useful in computing the variance of the stream:

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

$$= \left(\frac{1}{m} \sum_{i=1}^{m} f_i^2 \right) - \frac{n^2}{m^2}.$$

- For $p = \infty$, F_∞ is the frequency of the most frequent element.

* The uniform distribution:

A r.v. X assumes values in the interval $[a, b]$ such that all subintervals of equal length have equal probability, we say that X has the uniform distribution over $[a, b]$.

The probability distribution function of X is

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x \geq b. \end{cases}$$

and its density function is

$$f(x) = \begin{cases} 0 & \text{if } x < a. \\ \dfrac{1}{b-a} & \text{if } a \leq x \leq b. \\ 0 & \text{if } x > b. \end{cases}$$

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}\, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

$$\mathbb{E}[X^2] = \overset{\text{(Exercise!)}}{\cdots} = \frac{b^2 + ab + a^2}{3}$$

$\mathrm{Var}[X] = ?$

Lemma: Let $X_1, X_2, \ldots, X_k$ be independent random variables over $[0, 1]$. Let $Y = \min(X_1, X_2, \ldots, X_k)$.

Then $E[Y] = \dfrac{1}{k+1}$.

$$\mathbb{P}(Y \geq y) = \mathbb{P}(\min(X_1, \ldots, X_k) \geq y)$$

$$= \mathbb{P}(\{X_1 \geq y\} \cap \{X_2 \geq y\} \cap \ldots \cap \{X_k \geq y\})$$

X_i's are independent

$$= \prod_{i=1}^{k} \mathbb{P}(X_i \geq y)$$

$$= (1-y)^k$$

$$\therefore \mathbb{P}(Y \leq y) = 1-(1-y)^k.$$

$$F(y) = 1-(1-y)^k$$

density function of y is $F'(y) = f(y) = k(1-y)^{k-1}$.

$$\Rightarrow E[Y] = \int_0^1 ky(1-y)^{k-1}\,dy = y(1-y)^k\Big|_{y=0}^{1} + \int_0^1 (1-y)^k\,dy.$$

Integration by parts

$u = y \qquad dv = k(1-y)^{k-1}$

$du = dy \qquad v = -(1-y)^k$

$$= 0 + \int_0^1 (1-y)^k\,dy.$$

$$= -\frac{(1-y)^{k+1}}{k+1}\Big|_{y=0}^{1}$$

$$= \frac{1}{k+1}$$

* Estimating F_0. = Counting distinct elements.

[1] Noga Alon, Yossi Matias, and Mario Szegedy.
The space complexity of approximating the frequency moments. STOC'96.

[2] Edith Cohen.
Size-estimation framework with applications to transitive closure and reachability. '97.

universe

Idea: use a (hash) function $h: [m] \to [0,1]$.

We hash each entry a_i of the data ~~at~~ we see it, and keep track of the minimum seen hash value in our memory. Suppose in $a_1, a_2, \dots, a_n$, there are k distinct elements $x_1, x_2, \dots, x_k$.

Let $Y = \min(h(x_1), h(x_2), \dots, h(x_k))$.

Suppose that the values $h(a_1), \dots, h(a_n)$ are independently distributed uniform r.v. over the interval $[0,1]$.

$\Rightarrow$ From the previous lemma:

$$\cancel{\mathbb{E}[Y] = \frac{1}{n+1}}$$

$$\mathbb{E}[Y] = \frac{1}{k+1}$$

Recall that we want to estimate k, so Y may be used to estimate it.

$\hookrightarrow$ Can use Chebyshev's inequality.

$$\mathbb{E}[Y^2] = \int_0^1 y^2 k (1-y)^{k-1} \, dy$$

$$= \dots = ? \leq \frac{2}{(k+1)^2}$$

(Exercise!)

$$\Rightarrow \mathrm{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \leq \frac{1}{(k+1)^2} = \mathbb{E}[Y]^2.$$

Chebyshev's inequality:

$$\mathbb{P}(|Y - \mathbb{E}[Y]| > \varepsilon \mathbb{E}[Y]) \leq \frac{\mathrm{Var}[Y]}{\varepsilon^2 \mathbb{E}[Y]^2} \leq \frac{1}{\varepsilon^2}$$

which is a useless bound for small ε.

To improve, we take a mean of estimators.

Consider multiple independent versions $Y_1, Y_2, \ldots, Y_t$ of Y.

Y_1 has a corresponding hash function h_1

Y_2 $\qquad h_2$

$\vdots \qquad \vdots$

Y_t $\qquad h_t$

And let $Z = \dfrac{Y_1 + Y_2 + \ldots + Y_t}{t}$ as our new estimator.

$$\mathbb{E}[Z] = \mathbb{E}[Y] = \frac{1}{k+1}.$$

Since $Y_1, \ldots, Y_t$ are independent,

$$\cancel{\mathbb{E}} \quad \mathrm{Var}[Z] = \frac{1}{t^2} \sum_{i=1}^{t} \mathrm{Var}[Y_i] = \frac{\mathrm{Var}[Y]}{t} \leq \frac{\mathbb{E}[Y]^2}{t}.$$

$$\therefore \quad \mathrm{Var}[Z] \leq \frac{\mathbb{E}[Z]^2}{t}.$$

Applying Chebyshev's ineq:

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq \varepsilon \mathbb{E}[Z] \right) \leq \frac{\mathrm{Var}[Z]}{\varepsilon^2 \mathbb{E}[Z]^2} \leq \frac{1}{\varepsilon^2 t}.$$

This means that by increasing t, we can reduce the probability of bad event $\{ |Z - \mathbb{E}[Z]| \geq \varepsilon \mathbb{E}[Z] \}$.

Setting $t = \dfrac{10}{\varepsilon^2}$, we can bound the probability of failure by $\cancel{\frac{1}{4}}$ $\dfrac{1}{10}$

* Estimating the second moment F_2.

Recall $F_2 = \sum_{i=1}^{m} f_i^2$

Goal: Estimate F_2.

~~Consider a hash function $h: [m] \to \{-1, 1\}$.~~

For each symbol i, $1 \leq i \leq m$,
independently set a random variable X_i such that

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$$

then consider

$$S = \sum_{i=1}^{m} X_i f_i \quad \text{and} \quad V = \left(\sum_{i=1}^{m} X_i f_i\right)^2.$$

Fact: $\mathbb{E}[V] = \sum_{i=1}^{m} f_i^2$.

pf: Note that

$$\underbrace{\left(\sum_{i=1}^{m} X_i f_i\right)^2}_{} = \sum_{i=1}^{m} X_i^2 f_i^2 + 2 \sum_{i \neq j} X_i X_j f_i f_j$$

$$\Rightarrow \quad \mathbb{E}[V] = \mathbb{E}\left(\sum_{i=1}^{m} X_i^2 f_i^2\right) + 2\,\mathbb{E}\left(\sum_{i \neq j} X_i X_j f_i f_j\right)$$

$$= \sum_{i=1}^{m} \mathbb{E}[X_i^2 f_i^2] + 2 \sum_{i \neq j} \mathbb{E}[X_i X_j f_i f_j].$$

$$= \sum_{i=1}^{m} f_i^2 + 0.$$

$\mathbb{E}[X_i X_j f_i f_j]$
$= \mathbb{E}[X_i]\mathbb{E}[X_j] f_i f_j$ since X_i, X_j independent for $i \neq j$.
$= 0$.

$\Rightarrow$ V is an estimator of F_2.

We can show that (see [BHK] p. 190).

$$\mathbb{E}[V^2] \le 3\,\mathbb{E}[V]^2$$

$$\therefore \cancel{\le \mathbb{E}[V]}$$

$$\therefore \operatorname{Var}[V] = \mathbb{E}[V^2] - \mathbb{E}[V]^2$$

$$\le 2\,\mathbb{E}[V]^2.$$

By Chebyshev's inequality:

$$\mathbb{P}\left(|V - \mathbb{E}[V]| \geq \varepsilon \mathbb{E}[V] \right) \le \frac{\operatorname{Var}[V]}{\varepsilon^2 \mathbb{E}[V]^2} \le \frac{2}{\varepsilon^2}.$$

Not a good bound for ε small.

$\Rightarrow$ we can consider multiple independent version of V $V_1, V_2, \ldots, V_s$. and so Let $Y = \frac{1}{s}\sum_{i=1}^{s} V_i$.

Then $\mathbb{P}(|Y - F_2| \geq \varepsilon F_2) \le \delta$

if $s \geq \dfrac{2}{\varepsilon^2 \delta}$.

Alon-Matias-Szegedy was able to construct Y and V using $O(\log m)$ space. (See [BHK] p. 190).