Distinct Elements and Homework 8

December 1, 2016

Problem

We are given a sequence of elements $S = \{a_1, \ldots, a_T\}$. These elements are arriving one at a time, and as each arrives, we can do a bit of computation. Suppose that each a_i is a nonnegative 128 bit integer. We want to estimate the number n of distinct elements in the sequence in **constant time and space**. (Think of T as being astronomically large and the number of distinct elements may also be extremely huge.)

For example if $S = \{32, 5, 17, 32, 14, 5, 17, 5, 32, 17\}$, then n, the number of distinct elements, is 4.

1 An idea

Suppose that we had a hash function h that maps each 128 bit integer to a uniformly random real number in $[0, 1]$. Define

$$Y := \min_{1 \le i \le T} h(a_i).$$

Notice that as each a_i "arrives", Y can be updated in constant time and the total space used is constant.

For example, if $h(32) = 0.43$, $h(5) = 0.19$, $h(17) = 0.85$ and $h(14) = 0.61$, then Y would be equal to 0.19 at the end of the process.

The ingenious idea is the following: If there are n distinct elements in the sequence (and remember that the algorithm does *not* know n), then[1]

$$E(Y) = \frac{1}{n+1}. \tag{1.1}$$

Problem 1: Prove displayed equation (1.1).

Thus, if we computed Y, a natural way to estimate n would be to output

$$n := \frac{1}{Y} - 1.$$

[1] See the last section below for a start on how you would prove this.

(In our example above, that would mean that our estimate for n would be $1/0.19 - 1 = 4.26$, which is not bad..)

This would work superbly if we could guarantee that Y was close to its expectation. Unfortunately, this is not the case. For example, if Y was likely to be less than say 0.03, then the above estimate of n would be roughly 32, which is pretty far off!

In particular, the variance of Y is approximately

$$Var(Y) \approx \frac{1}{(n+1)^2}, \tag{1.2}$$

so the standard deviation is basically equal to the expectation.

Problem 2: Compute the variance of Y exactly. (You may need to use integration by parts.)

In general it can be very likely for a random variable to be more than one standard deviation from its mean.

How can we improve the quality of our estimate?

Repetition!!!!

2 Streaming Algorithm for estimating the number of distinct elements

Construct k hash functions $h_1, \ldots, h_k$, where each one independently maps each 128 bit integer to a uniformly random real number in $[0, 1]$. Define

$$Y_j := \min_{1 \le i \le T} h_j(a_i).$$

As above

$$E(Y_j) = \frac{1}{n+1} \quad \text{and} \quad Var(Y_j) \approx \frac{1}{(n+1)^2}.$$

Henceforth, we will pretend that this estimate for the variance is exact. (Do the same on all remaining homework problems.) Define

$$X := \frac{1}{k} \sum_{j=1}^{k} Y_j.$$

Then

$$E(X) = \frac{1}{n+1} \quad \text{and} \quad Var(X) = \frac{1}{k(n+1)^2}. \tag{2.1}$$

In other words, we've reduced the variance by a factor of k.

Problem 3: Suppose that $W_1, \ldots, W_k$ are independent random variables that each have mean μ and variance σ^2. Let

$$W := \frac{1}{k} \sum_{i=1}^{k} W_i.$$

Compute the expectation and variance of W in terms of μ and σ. Then apply your result to get displayed equation (2.1). (The answer to this question should be about 2 sentences long.)

We will use as our estimate for n

$$\hat{n} := \frac{1}{X} - 1.$$

2.1 Analysis

We will use the following tail bound known as **Chebychev's inequality**.

Theorem 2.1. *Let W be a random variable with mean μ and variance σ^2. Then*

$$Pr(|W - \mu| \geq c\,\sigma) \leq \frac{1}{c^2}.$$

For a proof, see http://inst.eecs.berkeley.edu/~cs70/sp16/notes/n18.pdf. Hopefully, I will also do this in class next week.

Let's apply Chevychev's inequality to our random variable X. Recall that

$$X := \frac{1}{k} \sum_{j=1}^{k} Y_j.$$

We have

$$E(X) = \frac{1}{n+1} \quad \text{and} \quad \sigma(X) = \frac{1}{\sqrt{k}(n+1)}.$$

so if we want to know

$$Pr(\left|X - \frac{1}{n+1}\right| \geq \frac{\epsilon}{n+1})$$

we should set

$$c\,\sigma(X) = \frac{c}{\sqrt{k}(n+1)} = \frac{\epsilon}{n+1},$$

or equivalently

$$c = \epsilon\sqrt{k}.$$

Thus,

$$Pr\left(\left|X - \frac{1}{n+1}\right| \geq \frac{\epsilon}{n+1}\right) \leq \frac{1}{c^2} = \frac{1}{\epsilon^2 k}.$$

We can play with these parameters to decide how good an approximation we want with what probability.

2.1.1 Example

If we want $\epsilon = 0.1$ and 90% chance of success, we take $\epsilon^2 k = 10$, i.e.

$$k = 1000.$$

This gives us that with 1000 hash functions

$$Pr\left(\left|X - \frac{1}{n+1}\right| \geq \frac{0.1}{n+1}\right) \leq \frac{1}{10}.$$

I.e.,

$$X \in \left[\frac{0.9}{n+1}, \frac{1.1}{n+1}\right]$$

with probability 0.9, or equivalently

$$n+1 \in \left[\frac{0.9}{X}, \frac{1.1}{X}\right]$$

with probability 0.9. Thus, we are 90% confident that

$$\frac{1}{X}$$

is within 10% of the true value of $n+1$. This is what they call a "confidence interval" in statistics.)

Problem 4: Use Chebychev's inequality to show that if $k = 400$, then the probability that

$$X \in \frac{1}{(n+1)}\left(1 \pm \frac{1}{10}\right)$$

is at least 0.75.

Problem 5: Use the result of problem 4 to show that estimating n as

$$\hat{n} := \frac{1}{X} - 1$$

guarantees that

$$\hat{n} \in \left[\frac{9}{11}n, \frac{11}{9}n\right]$$

with probability at least 0.75. (Note that these calculations are crude.)

2.2 Using the Central Limit Theorem instead

The Central Limit Theorem is one of the most important results in probability theory. It says that if $X_1, \ldots, X_n$ are independent, identically distributed, each with mean μ and variance σ^2, then $\sum_{i=1}^n X_i$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$. More formally,

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

approaches a $N(0,1)$ random variable as $n \to \infty$. We can also say that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

For some more details see page 11 here: http://inst.eecs.berkeley.edu/~cs70/sp16/notes/n20.pdf

Applying the Central Limit Theorem to our random variable X above implies that X is approximately normal with mean $1/(n+1)$ and variance $1/[k(n+1)^2]$.

Suppose that Z is a $N(0,1)$ random variable. For what α is it true that

$$Pr(|Z| > \alpha) = 0.1?$$

Using our normal tables, the answer is approximately $\alpha = 1.64$. Thus, if X is normal with the appropriate mean and variance

$$Pr\left(\left|X - \frac{1}{n+1}\right| \geq \frac{1.64}{\sqrt{k}(n+1)}\right) = \frac{1}{10}. \tag{2.2}$$

Problem 6: Prove the displayed equation (2.2).

If we want

$$Pr\left(\left|X - \frac{1}{n+1}\right| \geq \frac{\epsilon}{n+1}\right) = \frac{1}{10}$$

for $\epsilon = 0.1$, then it suffices to take

$$\frac{1.64}{\sqrt{k}} = 0.1$$

or

$$k = \left(\frac{1.64}{0.1}\right)^2 \approx 268.$$

Problem 7: As in the preceding discussion, suppose, by the Central Limit Theorem, that X is well approximated by a normal distribution with the appropriate mean and variance. For what k (as a function of ϵ) is

$$X \in \frac{1}{(n+1)}(1 \pm \epsilon) \tag{2.3}$$

with probability at least 0.95? What answer do you get for $\epsilon = 0.05$? For these choices if $X = 0.00015$, what would be the estimate of n the algorithm would output (i.e., the value of $(1/X) - 1$)? For these choices you can conclude that

$$Pr\left(n \in \left(\frac{1}{X} - 1 \pm \text{err}\right)\right) \geq 0.95,$$

where err is called the *margin of error*. What is the value of *err* you get and what is the probability of failure, i.e.

$$Pr\left(n \notin \left(\frac{1}{X} - 1 \pm \text{err}\right)\right)?$$

2.3 Don't forget

Don't forget that there are a couple more homework problems here: https://courses.cs.washington.edu/courses/cse312/16au/hw/hw8.pdf

More

For more on this and related topics, see these lecture notes:

http://www.cs.cmu.edu/afs/cs/project/pscico-guyb/realworld/www/slidesS14/stream.pdf

http://www.cs.princeton.edu/courses/archive/fall14/cos521/lecnotes/lec1.pdf

http://www.cs.princeton.edu/courses/archive/fall14/cos521/lecnotes/lec3.pdf

http://www.cs.princeton.edu/courses/archive/fall14/cos521/lecnotes/lec5.pdf

3 Calculations

Suppose that $U_1, \ldots, U_n$ are independent $U[0,1]$ random varaibles, and let

$$Y := min(U_1, \ldots, U_n).$$

Then

$$\begin{aligned} F_Y(y) &= Pr(Y \leq y) = 1 - Pr(Y > y) \\ &= 1 - \prod_{i=1}^{n} Pr(U_i > y) = 1 - (1-y)^n \qquad y \in [0,1]. \end{aligned}$$

You can now compute $f_Y(y)$ by differentiating the CDF $F_Y(y)$.