Asmt 4: Frequent Items

Gopal Menon
Turn in through Canvas by 2:45pm:
Wednesday, March 22

1 Streaming Algorithms

A: (20 points) Run the Misra-Gries Algorithm (see **L11.3.1**) with $(k-1) = 9$ counters on streams $S1$ and $S2$. Report the output of the counters at the end of the stream.

The estimated counts for streams $S1$ and $S2$ are given in tables 1 and 2 respectively.

Table 1: Misra-Gries Counter Outputs for stream $S1$

c	a	b	o	v	f	p
105,715	195,715	155,715	2	1	1	1

Table 2: Misra-Gries Counter Outputs for stream $S2$

b	c	a	h	l	j	w	r
135,715	175,715	245,715	1	1	1	1	1

In each stream, from just the counters, report how many objects might occur more than 20% of the time, and which must occur more than 20% of the time.

For any item q, the actual frequency f_q and the frequency $\hat{f}_q$ reported by the algorithm are related by the inequality

$$f_q - \frac{m}{k} \leq \hat{f}_q$$

where $m = 1{,}000{,}000$ is the size of the stream, and $k = 10$ is the number of counters plus 1.
Substituting these values into the above inequality, we get

$$\begin{aligned} f_q - \frac{1{,}000{,}000}{10} &\leq \hat{f}_q \\ f_q - 100{,}000 &\leq \hat{f}_q \\ f_q - \hat{f}_q &\leq 100{,}000 \end{aligned}$$

This means that the maximum possible undercounting is by 100,000. Given that 20% of 1,000,000 is 200,000, any label with count more than 200,000, must occur more than 20% of the time since over-counting is not possible. So label a in stream $S2$ must occur more than 20% of the time. Any label with count between 100,000 and 200,000 might occur more that 20% of the time. So labels a, b and c in stream $S1$ and labels b and c in stream $S2$ might occur more than 20% of the time.

B: (20 points) Build a Count-Min Sketch (see **L12.1.1**) with $k = 10$ counters using $t = 5$ hash functions. Run it on streams $S1$ and $S2$.
For both streams, report the estimated counts for objects a, b, and c. Just from the output of the sketch, which of these objects, with probably $1 - \delta = \frac{31}{32}$, might occur more than 20% of the time?

The estimated counts for streams $S1$ and $S2$ are given in table 3.

Table 3: Count-Min Sketch Counter Outputs

Stream	a	b	c
$S1$	266,758	243,029	176,294
$S2$	303,485	206,983	233,533

For any item q, the actual frequency f_q and the frequency $\hat{f}_q$ reported by the algorithm are related by the PAC bound,

$$f_q \leq \hat{f}_q \leq f_q + \varepsilon m$$

where the second inequality holds with a probability of $1 - \delta$. The number of hash functions is related to δ by $t = \log_2\left(\frac{1}{\delta}\right)$ the number of counters per hash function is related to ε by $k = \frac{2}{\varepsilon}$, and m is the number of items in the stream.

Given that

$$\begin{aligned}
1 - \delta &= \frac{31}{32} \\
\delta &= 1 - \frac{31}{32} \\
&= \frac{1}{32} \\
t &= \log_2\left(\frac{1}{\delta}\right) \\
&= \log_2\left(\frac{1}{\frac{1}{32}}\right) \\
&= \log_2(32) \\
&= 5
\end{aligned}$$

This value of t matches with the number of hash functions in the experiment.

$$\begin{aligned}
k &= \frac{2}{\varepsilon} \\
\varepsilon &= \frac{2}{k} \\
&= \frac{2}{10}
\end{aligned}$$

From the above inequality, we can see that

$$\begin{aligned}
\hat{f}_q &\leq f_q + \varepsilon m \\
&\leq f_q + \frac{2}{10} \times 1{,}000{,}000 \\
&\leq f_q + 200{,}000 \\
\hat{f}_q - f_q &\leq 200{,}000
\end{aligned}$$

This means that the over-counting is limited by 200,000. Since under-counting is not possible, we can see that since a and b in stream $S1$ and a, b and c in stream $S2$ have an estimated count of at least 200,000, these labels might occur more that 20% of the time (since 20% of 1,000,000 is 200,000) with probability $\frac{31}{32}$.

C: (5 points) How would your implementation of these algorithms need to change (to answer the same questions) if each object of the stream was a "word" seen on Twitter, and the stream contained all tweets concatenated together?

Assuming that we knew the number of words, m, in the stream of tweets, the implementation of the algorithms could be changed to have counters for words instead of characters. If the number of words is not known, then these can be just counted using a word counter as the stream is being read. The size of the counter would be $\log_2 m$, where m is the number of words in all the tweets.

For Misra-Gries,

$$\begin{aligned} f_q - \frac{m}{k} &\leq \hat{f}_q \\ f_q - \hat{f}_q &\leq \frac{m}{k} \\ &\leq \frac{m}{10} \end{aligned}$$

For the case of Misra-Gries, the maximum undercount would be $\frac{m}{10}$. For an object to have to exist 20% of the time, the count needs to be at least $\frac{m}{5}$ (which is 20% of m) . Any object with a count between $\frac{m}{5} - \frac{m}{10}$ and $\frac{m}{5}$, which is between $\frac{m}{10}$ and $\frac{m}{5}$, may exist at least 20% of the time.

For Count-Min Sketch,

$$\begin{aligned} \hat{f}_q - f_q &\leq \varepsilon m \\ &\leq \frac{2}{k} m \\ &\leq \frac{2}{10} m \\ &\leq \frac{m}{5} \end{aligned}$$

For the case of Count-Min Sketch, any object seen between $\frac{m}{5}$ (which is 20% of m) and $\frac{m}{5} + \frac{m}{5}$, which is between $\frac{m}{5}$ and $\frac{2m}{5}$, might occur 20% of the time.

D: (5 points) Describe one advantage of the Count-Min Sketch over the Misra-Gries Algorithm.

One advantage is that the order of the objects in the stream does not matter. The reason is that there is no decrementing of counters when the objects coming in are counted. In Misra-Gries, when an object comes in and there is no counter associated with the object, all counters are decremented. So if an object is spread out in a stream, it's count may be different from the case where it is not spread out.