Lecture 11: Count-Min and Count Sketches

Lecturer: Anshumali Shrivastava
Scribers: Alan Liu, Eric Breyer, Noah Spector

September 26, 2023

1 Count-Min Sketches

Count-min sketches are a probabilistic data structure that allow us to track the occurrences of an event given a stream of data. The structure of a count-min sketch is a 2D matrix M with d rows and R columns. It also maintains a set of d hash functions h_j, one per row. When an event of type i occurs, the count-min sketch stores this occurrence by determining $h_j(i)$ $\forall j = 1, ..., d$, which will increment the values at the buckets that are the outputs of the hash functions.

One can query the total number of occurrences $\hat{c_i}$ of an event i by taking the minimum of the values stored in the buckets for the outputs of $h_j(i)$. This is given by the formula below:

$$\hat{c_i} = min[j, h_j(i)] \tag{1}$$

Note that $\hat{c_i}$ is an estimate for c_i since we can clearly observe that the count-min sketch can return a value greater than the true number of occurrences of i. This is the key difference between count-min sketches and count sketches. Unlike count sketches, count-min sketches will always overestimate the count. However, it guarantees that the estimate will be within the following range with probability $1 - \delta$,

$$c_i \leq \hat{c_i} = c_i + \sum_{j=1, j \neq i}^{N} c_j \cdot \mathbb{1}_{\{h(i)=h(j)\}}, \tag{2}$$

where $\mathbb{1}_{\{h(i)=h(j)\}}$ is an indicator variable such that

$$\mathbb{1}_{\{h(i)=h(j)\}} = \begin{cases} 1, & \text{if } h(i) = h(j) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We also get that

$$\hat{c_i} \leq c_i + \epsilon \cdot \Sigma, \tag{4}$$

where Σ is the total number of events that were passed in to the sketch.

2 Count Sketches

Count sketches are another variant of the sketch data structure to track event occurrences. In this case, every bucket in the count sketch has a sign $s_j(i) \in \{-1, 1\}$, rather than buckets only adding to the count. Thus, we obtain a different guarantee for the estimate of the occurrences for an event i, given by

$$\hat{c}_i = s_k(i)c_i + \sum_{j=1, j\neq i}^{N} s_k(j)c_j \cdot \mathbb{1}_{\{h(i)=h(j)\}} \cdot s_k(i) \tag{5}$$

$\forall k = 1, ..., d$. In a table with one row.

Querying for the total occurrences of an event i will return

$$\frac{median}{k}[s_k(i) \cdot M[k, h_k(i)]]. \tag{6}$$

2.1 Expected Value of c_i

We can derive an important result for $\hat{c}_i$,

$$s_k(i)E[\hat{c}_i] = c_i. \tag{7}$$

We know that $E[s_k(i)] = 0$ since it takes the value -1 or 1 at random, so we can write

$$E[\hat{c}_i] = s_k(i)c_i + E[\sum_{j=1, j\neq i}^{N} s_k(j) \cdot s_k(i) \cdot c_j \cdot \mathbb{1}_{\{C\}}] \tag{8}$$

where $\mathbb{1}_{\{C\}} = \mathbb{1}_{\{h(i)=h(j)\}}$

as

$$\begin{aligned} E[\hat{c}_i] &= s_k(i)c_i + E[\sum_{j=1, j\neq i}^{N} c_j \cdot \mathbb{1}_{\{C\}}] \\ &= s_k(i)c_i. \end{aligned}$$

We can then multiply by $s_k(i)$ to get

$$\begin{aligned} E[\hat{c}_i] &= s_k(i)c_i \\ s_k(i)E[\hat{c}_i] &= (s_k(i))^2 c_i. \\ s_k(i)E[\hat{c}_i] &= 1 \cdot c_i \\ s_k(i)E[\hat{c}_i] &= c_i \end{aligned}$$

2.2 Variance Analysis of c_i

We first note that $E[\hat{c}_i{}^2]$ is a dependency for determining the variance. We have

$$E[\hat{c}_i{}^2] = c_i^2 + \frac{1}{R}\sum_{j=1, j\neq i}^{N} c_j^2. \tag{9}$$

Thus,

$$\begin{aligned}
Var(\hat{c}_i) &= E[\hat{c_i}^2] - E[\hat{c}_i]^2 \\
&= E[\hat{c_i}^2] - (s_k(i)c_i)^2 \\
&= c_i^2 + \frac{1}{R}\sum_{j=1, j\neq i}^{N} c_i^2 - c_i^2 \\
&= \frac{1}{R}\sum_{j=1, j\neq i}^{N} c_i^2,
\end{aligned}$$

which gives us a bound on the variance

$$Var(\hat{c}_i) \leq \frac{1}{R}\sum_{j=1}^{N} c_i^2 = \frac{1}{R}\Sigma^2. \tag{10}$$

Where $[1, N]$ is the possible values of an event and R is the number of rows in the table

2.3 Using the power of k choices

We can make the variance even better by repeating the above process k times and taking the median.
We want to find the probability that a median estimator is farther away than ϵ. Using Chebyshev's inequality, we find

$$Pr(|\hat{c}_i - c_i| \geq \epsilon c_i) \leq \frac{Var(\hat{c}_i)}{\epsilon^2 c_i^2} \tag{11}$$

Now put $f = \sum_{j=1}^{N} c_j^2 / c_i^2$. We now observe

$$\frac{Var(\hat{c}_i)}{\epsilon^2 c_i^2} \leq \frac{f}{R\epsilon^2} \tag{12}$$

Now, we choose $2k$ items and take the median. In order for the estimator to be off, at least k items must be outside the range to one side. Without observing anything about the distribution of $\hat{c}_i$, we find

$$Pr(Median_k(|\hat{c}_i - c_i| \geq \epsilon c_i)) \leq (\frac{f}{R\epsilon^2})^k = \frac{f^k}{R^k \epsilon^{2k}} \tag{13}$$

Therefore, if $\frac{f}{R\epsilon^2} < 1$, as one increases the number of hash functions, the probability of the median estimator falling outside of a given fraction ϵ falls exponentially. If c_i is a heavy hitter, then f is an appreciable fraction, so we can choose R to be large enough to counterbalance ϵ.

References

[1] Anirban Dasgupta (2018) *Frequent Element: Count Sketch*, Youtube.

[2] Shusen Wang *Count Sketch*, Github