COMP 480/580 — Probabilistic Algorithms and Data Structure Sep 22, 2022

Lecture 10

Lecturer: Anshumali Shrivastava *Scribe By: Doug Welsch and Andre Wasem*

1 Count-Min Sketch: A Primer

1.1 Motivation

Rather than a batch and/or fixed collection of inputs, consider the scenario where data is received sequentially. Such a sequence is referred to as a stream and typically cannot be stored accessibly. Streams appear in a variety of applications over large data sets such as trending social media topics and sensor networks.

1.2 Description

Consider the stream of items received in order $(i_1, \Delta_1), (i_2, \Delta_2), ..., (i_t, \Delta_t), ...,$ where each element i_j represents some item from stream, Δ_j to be its corresponding increment, and t indicates the current stream element. The Count-Min Sketch is a data structure of d (depth) hash functions each of size R (range) such that for each item received i_j, the associated count at $h_k(i_j)$ is incremented by Δ_j for $k = 1...R$. (This operation is known as an update.). The count associated with i_j, c_j, is then approximated by taking the minimum of each array entry of array at index $h_k(i_j)$ for $k = 1...R$; i.e., $c_j = min(h_1(i_j), h_2(i_j), ..., h_R(i_j))$. (This operation is known as a query.)

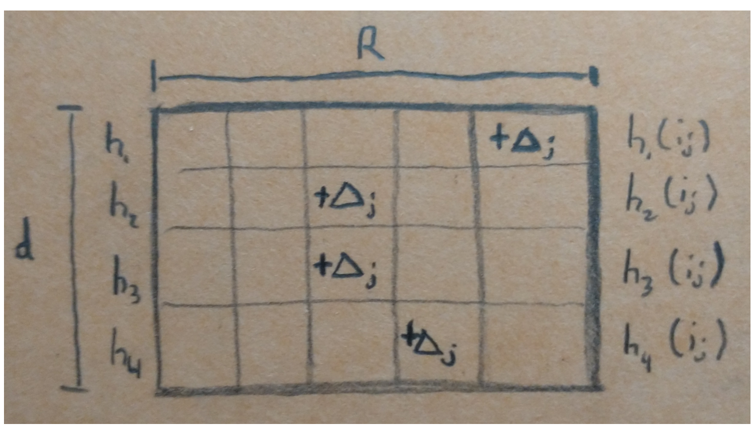

Figure 1: Example Count-Min Sketch. This sketch includes $d = 4$ hash functions of size $R = 5$ and shows the increment of each item in each array.

1.3 Analysis

Consider one hash function $h(x)$. For an item i_j, the value of i_j's count is the count itself plus any items counts that were hashed to the same entry:

$$h(c_{i_j}) = c_{i_j} + \sum(\mathbb{1}_{h_1(i_k) = h_1(i_j)} * c_{i_k}) \tag{1}$$

The expectation value of the count variable indicator variable $\hat{c}_{i_j}$ is the following, where Σ is the sum of all element counts (Aside: $\epsilon\Sigma$ as a whole is considered the error or overestimate.):

$$E[\hat{c}_{i_j}] = c_{i_j} + \frac{1}{R}(\Sigma - c_{i_j}) < c_{i_j} + \frac{\Sigma}{R} = c_{i_j} + \epsilon\Sigma \tag{2}$$

With $R = \frac{1}{\epsilon}$, and using Markov's Inequality:

$$c_{i_j} < \hat{c}_{i_j} < c_{i_j} + 2\epsilon\Sigma \tag{3}$$

The probability that $\hat{c}_{i_j}$ is within this range is $> \frac{1}{2}$ for the one hash function. So, for the d independent hash functions, the probability that $\hat{c}_{i_j}$ is within the range is $> 1 - (1/2)^d$.

1.4 Negative Counts

However, the above derivation (and the Count-Min Sketch) relies on the assumption that all increments are positive. It is possible for there to be negative counts, for example, if there was a malicious stream. Or, more generally, a packet can be regarded as just a number as in machine learning scoring.

2 Count Sketches

2.1 Description

Consider a new hash sign function $S(x)$ that hashes the input to either -1 and 1 with probability $\frac{1}{2}$. Now, within the Count Sketch, each item i updates each array at index $h_k(i_j)$ with $\Delta_j * S_k(i_j)$ for each hash function k. Updates are performed by incrementing each entry for an item i_j in a hash function k by $S_k(i_j * \Delta_j)$, and queries are determined by taking the median, rather than minimum, of counts across hash function arrays.

2.2 Analysis

Considering again one hash function, Equation (1) becomes:

$$S_1(i_j) * h_1(i_j) = S_1(i_j) * c_j + \sum(\mathbb{1}_{h_1(i_k) = h_1(i_j)} * c_k * S_1(i_k) * S_1(i_j)) \tag{4}$$

Similarly, the expectation value becomes:

$$E[\hat{c}_{i_j}] = c_{i_j} * S_1(i_j)^2 + E[\sum(\mathbb{1}_{h_1(i_k) = h_1(i_j)} * c_k * S_1(i_k) * S_1(i_j))] \tag{5}$$

Additionally, $S_1(i_j)^2 = 1$ and, since the sign hash functions were generated independently, $E[Error] = 0$, so:

$$E[\hat{c}_{i_j}] = c_{i_j} \tag{6}$$

Since the counts are no longer positive, Chebyshev's inequality is used to approximate probabilities rather than Markov's, so we need to calculate the variance of $\hat{c}_{i_j}$

$$Var(Error) = E[Error^2] - E[Error]^2 = E[Error^2] \tag{7}$$

Then:

$$E[Error] = E[\sum_{k=1}^{N} \mathbb{1}_{h(i_k) = h(i_j)} * c_{i_k}^2] + \sum_{k=1}^{N}\sum_{l=1}^{N} \mathbb{1}_{h(i_k) = h(i_j)} * \mathbb{1}_{h(i_l) = h(i_j)} * c_{i_l} * c_{i_k} * S(i_k) * S(i_l)] \tag{8}$$

By the linearity of expectations, the two terms can be separated and the latter becomes 0 by similar reasoning of the sign functions being independent. After resolution:

$$Var(Error) = E[\sum_{k=1}^{N} \mathbb{1}_{h(i_k) = h(i_j)} * c_{i_k}^2] \leq \frac{1}{R}\Sigma^2 \tag{9}$$