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 \((i_1, \Delta_1), (i_2, \Delta_2), \ldots, (i_t, \Delta_t), \ldots\), where each element \(i_j\) represents some item from stream, \(\Delta_j\) 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 \(\Delta_j\) for \(k = 1 \ldots R\). (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 \(k = 1 \ldots R\); i.e., \(c_j = \min(h_1(i_j), h_2(i_j), \ldots, h_R(i_j))\). (This operation is known as a query.)

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:

\[ h(c_{ij}) = c_{ij} + \sum (1_{h_1(i_k) = h_1(i_j)} \ast c_{ik}) \]  

(1)
The expectation value of the count variable indicator variable \( \hat{c}_{ij} \) is the following, where \( \Sigma \) is the sum of all element counts (Aside: \( \epsilon \Sigma \) as a whole is considered the error or overestimate.):

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

(2)

With \( R = \frac{1}{\epsilon} \), and using Markov’s Inequality:

\[
c_{ij} < \hat{c}_{ij} < c_{ij} + 2\epsilon \Sigma
\]

(3)

The probability that \( \hat{c}_{ij} \) is within this range is \( > \frac{1}{2} \) for the one hash function. So, for the \( d \) independent hash functions, the probability that \( \hat{c}_{ij} \) 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 \( h_k(ij) \) with \( \Delta_j * S_k(ij) \) for each hash function \( k \). Updates are performed by incrementing each entry for an item \( ij \) in a hash function \( k \) by \( S_k(ij * \Delta_j) \), 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:

\[
S_1(ij) * h_1(ij) = S_1(ij) * c_j + \sum (\mathbb{1}_{h_1(ik) = h_1(ij)} * c_k * S_1(i_k)^* S_1(ij))
\]

(4)

Similarly, the expectation value becomes:

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

(5)

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

\[
E[\hat{c}_{ij}] = c_{ij}
\]

(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 \( \hat{c}_{ij} \):

\[
Var(Error) = E[Error^2] - E[Error]^2 = E[Error^2]
\]

(7)

Then:

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

(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:

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