CS60021: Scalable Data Mining
Autumn 2021

Practice questions: Streaming algorithms and Locality Sensitive Hashing

Q1. Consider a trucking company which transports fruits. It is known that they spoil a random fraction of the fruits between 40% – 10%, with uniform probability distribution. Say X is the percentage of unspoilt fruits delivered. What can we say about P (X > 80) ?

Q2. Consider a sampling algorithm which uniformly samples real numbers in the range 0 – −100. Given that you have sampled 10 numbers, what is the probability that at least two of them has the same integral part ? Consider 0 as the integral part of number 0.5.

Q3. Consider the search procedure where an input string of length n is matched one symbol at a time with a known string of the same length. Assuming that alphabet size is 10, what is the probability that an unsuccessful search lasts for k timesteps, assuming the input string is generated uniformly randomly from the alphabet.

Q4. In a bloom filter table with 10 slots and 3 hash functions, what is the probability that a bit (chosen randomly) is not set to 1 after the first insertion ?

Q5. Deletion of elements from Bloom filter is not allowed. Why?

Q6. You have a stream of n elements. You run the Misra-Gries heavy-hitter algorithm on this set of items with k = 1 (i.e. a single heavy hitter). Which of the following statements is correct.

(a) The output is independent of the order in which the elements arrive, irrespective of the frequencies.
(b) If there is one element with clear majority, that element will be output.
(c) The output is always the middle element of the stream, i.e. the element that appears at (m/2)th position in the stream, m equals length of the stream.
(d) The output is always the last element of the stream.

Q7. Which of the following statements is correct. Suppose you run Misra-Gries algorithm on a stream of length n (with repetitions) and with k = 2. Suppose the Misra-Gries data structure at the end has elements x and y and m_x and m_y are the two counts. Let c_x and c_y be the true counts of x and y in the stream. Which of the following is true.

(a) $m_x \leq c_x$ and $m_y \leq c_y$.
(b) It is possible that either $m_x > c_x$ and $c_x > c_y$.
(c) Either $m_x > c_x$ and $c_x > c_y$ happens, but both are not true.
(d) Both $m_x > c_x$ and $c_x > c_y$ always happens.

Q8. Consider the count-min sketch with w hash functions and d buckets. If you increase d, how will the estimation error change?

Q9. Which of the statements is true? Assume both the Count-Min and the Count-Sketch are using same parameters (same size of sketch and same number of hash functions).

(a) CountSketch always provides a smaller error than CountMin sketch.
(b) CountSketch can handle frequencies becoming negative, whereas CountMin does not have useful guarantees if that happens.
(c) CountMin always provides a smaller error than CountSketch.
(d) Both give exactly the same result for every query.

Q10. Suppose the stream is $x_1, x_2, \ldots, x_n$. You have designed a Count-Min sketch for this stream. But due to a mistake in your code, you used the same hashing functions for all the rows of the Count-Min sketch (i.e. $h_1,\ldots,h_k$ are all same function). For query x, your code still returns Which of the following statements is true.

(a) There are no guarantees that can be given on the output.
(b) The error guarantees on the output are better than what would obtain if we were keeping k different hash functions.
(c) If each hash function has range B, then the returned estimate does not exceed the true value by n/4B with probability 0.75.
(d) For any query it will return the same value.

Q11. Suppose you have a (k, L) LSH, k-anding rows and L such tables. As you increase k, while keeping L fixed, how is the recall likely to change?

Q12. Consider the probability of two points p, q being hashed to the same bucket being 0.7. Suppose we use a LSH scheme with k=5, L=10. If p is present in the data, and q is given as query, what is the probability that p will not come up in the list of candidates using this LSH scheme.