DS702 Assignment 3

Release Date: 27 February 2022
Due Date: 13 March 2022

- Submit your answers as an electronic copy on Moodle (pdf, jupyter notebook).
- No unapproved extension of deadline is allowed. For emergencies and sickness, extensions must be requested as soon as possible.
- Cite your sources if you are taking help (papers, websites, students etc.).
- Plagiarism is strictly prohibited. Negative mark will be assigned for plagiarism.
- Remember to comment your code. And your answers should be detailed.

1 Sampling Data in a Stream

Suppose we have a stream of tuples with the schema

`Grades(university, courseID, studentID, grade)`

Assume universities are unique, but a `courseID` is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., "DS702") and likewise, `studentID's` are unique only within a university (different universities may assign the same ID to different students). Suppose we want to answer certain queries approximately from a 1/20th sample of the data. For each of the queries below, indicate how you would construct the sample. That is, tell what the key attributes should be.

(a) For each university, estimate the average number of students in a course

(b) Estimate the fraction of students who have a GPA of 3.5 or more

(c) Estimate the fraction of courses where at least half the students got "A."

2 Filtering Streams

Exercise 2.1: For the situation of the running example in the book (8 billion bits, 1 billion members of the set S) [Section 4.3.1][1], calculate the false-positive rate if we use three hash functions? What if we use four hash functions?

Exercise 2.2: As a function of n, the number of bits and m the number of members in the set S, what number of hash functions minimizes the false positive rate?

3 Distinct Elements

Exercise 3.1: Suppose our stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Our hash functions will all be of the form h(x) = ax+ b mod 32 for some a and b. You should treat the result as a 5-bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is:

(a) h(x) = 2x + 1 mod 32.

(b) h(x) = 3x + 7 mod 32.

(c) h(x) = 4x mod 32.

Exercise 3.2: Do you see any problems with the choice of hash functions in Exercise 3.1? What advice could you give someone who was going to use a hash function of the form h(x) = ax + b mod 2^k?

4 Counting Ones in a Window

Exercise 4.1: Suppose the window is as shown in Fig. 1. Estimate the number of 1's the the last k positions, for k = (a) 5 (b) 15. In each case, how far off the correct value is your estimate?

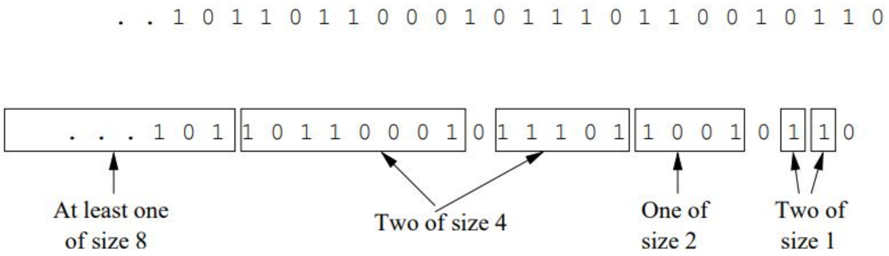

Figure 1: A bit-stream divided into buckets following the DGIM rules

Exercise 4.2: There are several ways that the bit-stream 1001011011101 could be partitioned into buckets. Find all of them.

References

[1] Jure Leskovec et al. Mining of Massive Datasets. 2019