IS622 Week5

James Quacinella

10/04/2015

Exercise 4.2.1 (section 4.2.5)

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 uni- versity (i.e., different universities may have different courses with the same ID, e.g., "CS101") 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."

Answer

The question sounds like it is just asking which part of the tuple should be considered the key.

(a) For the average number of users in a course, we need to break down samples by university and courseID. This is because courseID is not unique in the samples, since courseIDs may be duplicated across universities. Therefore they key should be the **university and courseID** portions of the input tuples.

(b) Since we want data per student, we need to identify samples by student. However, studentIDs are not unique across universities, so the key would have to include both **university and studentID**.

(c) I think similar to (a), we need to group by course, which means we need the **university and courseID** portions of the input tuples.

Exercise 4.3.3 (section 4.3.4)

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?

Answer

From the book: In general, the probability of a false positive is the probability of a 1 bit, which is $1 - e^{-km/n}$, raised to the kth power, i.e., $(1 - e^{-km/n})^k$. With respect to k, we need to find the first derivative of this:

$$FPR_k(n, m) = (1 - e^{-km/n})^k$$

We can express this in a different way to start the process:

$$FPR_k(n, m) = e^{ln((1-e^{-km/n})^k)}$$

The full derivative of this is difficult to compute, and includes quite a few steps, so I'll write it out in final form for now:

$$FPR'_k(n, m) = \frac{d}{dk}\left(e^{ln((1-e^{-km/n})^k)}\right)$$
$$= (1 - e^{-km/n})^k) * \left(\frac{(km/n)e^{-km/n}}{1 - e^{-km/n}} + log(1 - e^{-km/n})\right)$$

I do not know how to solve the equation by setting the derivative to zero, so the best I can say is that the solution is the k value which makes the derivative 0 (with proper concavity checks). Using code, we can investigate specific k values that minimize the FPR for a given m and n value, but that doesn't seem to be what the question asks for.

TODO

Exercise 4.5.3 (section 4.5.6)

Suppose we are given the stream of Exercise 4.5.1, to which we apply the Alon-Matias-Szegedy Algorithm to estimate the surprise number. For each possible value of i, if X_i is a variable starting position i, what is the value of $X_i.value$?

Answer

The stream in question is: 3, 1, 4, 1, 3, 4, 2, 1, 2. Therefore, i can be an element from 1 to n, in this case is 9. Therefore:

$$X_1.element = 3; X_1.valuemustbe2$$

$$X_2.element = 1; X_2.valuemustbe3$$

$$X_3.element = 4; X_3.valuemustbe2$$

$$X_4.element = 1; X_4.valuemustbe2$$

$$X_5.element = 3; X_5.valuemustbe1$$

$$X_6.element = 4; X_6.valuemustbe1$$

$$X_7.element = 2; X_7.valuemustbe2$$

$$X_8.element = 1; X_8.valuemustbe1$$

$$X_9.element = 2; X_9.valuemustbe1$$