IS622 Homework

Ben Arancibia

September 26, 2015

4.2.1

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.

Since we want a per university estiamte of the number of studnets in a course we'll need 2 keys: univerisy, courseID.

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

GPA is a average across different courses so we'll use: studentID and courseID as the keys (no need for university because we only need an estimate).

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

We need university, courseID, and grade as keys. A student would not have their grade listed twice for the same course.

4.3.3

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?

$(1 - exp(-km/n))^k$

so I think this equation : log(m) - log(n) -1 shows the number of has functions that minimizes the false-positive rate.

4.5.3

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 Xi is a variable starting position i, what is the value of Xi.value?

The Alon-Matias-Szegedy Algorithm

Starting Postion i	Xi.element	Xi.value
1st	3	2
2nd	1	3
3rd	4	2
4th	1	2

Starting Postion i	Xi.element	Xi.value
5th	3	1
6th	4	1
7th	2	2
8th	1	1
9th	2	1

An estimate for the value can be found by the following formula.

$F = Sum(n) * (2 * X.value - 1)/(numberofkeptvariables)$

n = length of the stream.

The second moment of the stream is estimated as follows: $F2 = Sum(9) * (2 * X.value - 1)/(9) = 21$.

This result utilizes all 9 possible starting positions for all variables. If less than 9 variables are used to save computational cost, the result will be slightly different from the true value but still within acceptable error.