Big Data Analytics

ESSEC

Olga Klopp

HomeWork 4 Solution: Mining Data Streams, part 2

1. (**Exercise 4.2.1 MMDS book**) Suppose we have a stream of tuples with the schema (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., "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."

   Solution:

   (a) The query wants to generate average number of students in a course. For each tuple, the "university" field is unique, then we chose "university" as the key. To take a sample of 1/20th, we hash the key for each tuple to an integer from 0 to 19, and accept the tuple for the sample if the hash value is 0. Thus, we store only 1/20th of the tuples as the sample, and discard others. For each university in the sample, we can easily count the average number of students in a course.

   (b) The query wants to estimate the fraction of students who have a GPA of 3.5 or more. Since the "studentID" is unique only within a university, it cannot be alone used to identify one tuple in the stream. Here we need to build a composite key, (university, studentID) to identify each couple. We hash the composite key to an integer from 0 to 19, and accept the tuple for the sample if the hash value is 0. Thus, we store only 1/20th of the tuples as the sample, and discard others.

   (c) Same solution as for (b) taking as composite key (university, courseID).

2. (**Exercise 4.3.1 : MMDS book**) For the situation of our running example (8 billion bits, 1 billion members of the set S), calculate the false-positive rate if we use three hash functions? What if we use four hash functions?
   Solution: As we discussed during the lecture, the probability that a given bit will be 1, using one hash function, is $1 - e^{-1/8} \approx 0.1175$. Now, suppose that we use three different hash functions. This situation corresponds to throwing three billion darts at eight billion targets, and the probability that a bit remains 0 is $e^{-3/8}$. In order to be a false positive, a nonmember of S must hash thrice to bits that are 1, and this probability is $(1 - e^{-3/8})^3 \approx 0.03$. Adding a fourth hash function we will get $(1 - e^{-1/2})^4 \approx 0,024$.

3. (**Exercise 4.4.1 MMDS book**) 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 $h(x) = ax + b \bmod 32$ 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) $h(x) = 2x + 1 \bmod 32$;

(b) $h(x) = 3x + 7 \bmod 32$;

(c) $h(x) = 4x \bmod 32$;

Solution:

(a)

Element	Hashed value	Binary representation	Tail length
3	7	00111	0
1	3	00011	0
4	9	01001	0
1	3	00011	0
5	11	01011	0
9	19	10011	0
2	5	00101	0
6	13	01101	0
5	11	01011	0

For the maximum tail length, R, we have $R = 0$ and the number of distinct elements is estimated to be $2^0 = 1$.

(b)

Element	Hashed value	Binary representation	Tail length
3	16	10000	4
1	10	01010	1
4	19	10011	0
1	10	01010	1
5	22	10110	1
9	2	00010	1
2	13	01101	0
6	25	11001	0
5	22	10110	1

For the maximum tail length, R, we have $R = 4$ and the number of distinct elements is estimated to be $2^4 = 16$.

(c)

Element	Hashed value	Binary representation	Tail length
3	12	01100	2
1	4	00100	2
4	16	10000	4
1	4	00100	2
5	20	10100	2
9	4	00100	2
2	8	11000	3
6	24	11000	3
5	20	10100	3

For the maximum tail length, R, we have $R = 4$ and the number of distinct elements is estimated to be $2^4 = 16$.

If we take the mean of these estimators, we get 11 as an estimate of the number of distinct elements which overestimates the true number, 6.