Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 3: Mining Data Streams (Chapter 4)

Due: Apr 13, 09:30 AM

Reminders

- The points of this homework add up to 100.
- Like all homeworks, this has to be done individually.
- Lead T.A.: Jinhong Jung (montecast9@gmail.com)
- Please type your answers in English. Illegible handwriting may get no points, at the discretion of the graders.
- If you have a question about assignments, please upload your question in eTL.
- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date. That is, after Apr 20, we will NOT receive your submission for this assignment.

Remember that:

- Whenever you are making an assumption, please state it clearly

Question 1

Suppose you use a bloom filter for 8 billion bits and 1 billion keys. Calculate the false-positive rate when we use the following number of hash functions. [25 points]

(a) Three hash functions.

(b) Four hash functions.

Question 2

Suppose you are given the following stream:

b, c, a, d, a, c, d, b, a, b, a, c, d

Answer the following questions. [25 points]

(a) What is the surprise number (second moment) of the stream?

(b) What is the third moment of the stream?

Question 3

Suppose we are given the stream of *Question 2*, to which we apply the Alon-Matias-Szegedy (AMS) Algorithm to estimate the surprise number. Suppose we keep four variables X_1, X_2, X_3 and X_4. Assume that we randomly pick the 2nd position for X_1, the 5th position for X_2, the 7th position for X_3, and the 8th position for X_4 to define the four variables. After looking at the stream, answer the following questions using the AMS algorithm. For simplicity, assume that you know the length of the stream (in this case, it's 13) in advance. [25 points]

(a) What are $\mathrm{X_i}.element$ and $\mathrm{X_i}.value$ for the four variables?

(b) What is the estimated surprise number?

Question 4

Suppose the window is as shown in Figure 1. Estimate the number of 1's the last k positions, for k = (a) 5 (b) 15. In each case, how far off the correct value is your estimate? [25 points]

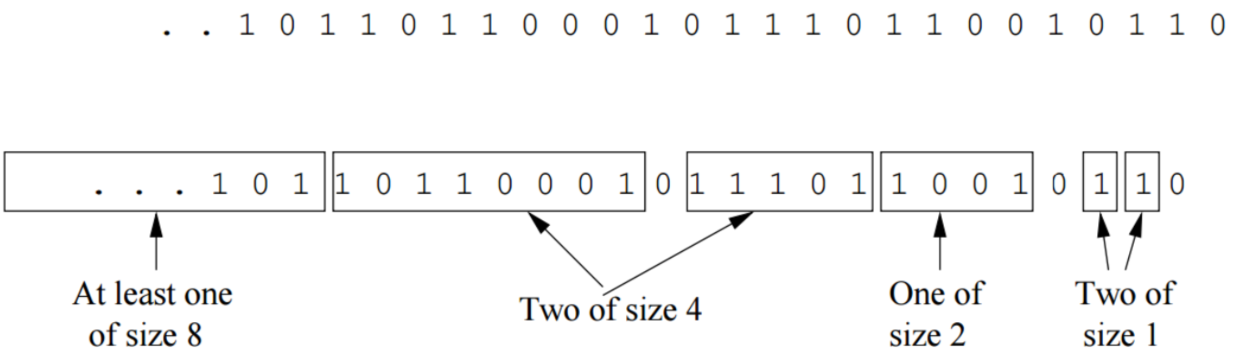

Figure 1. A bit-stream divided into buckets following the DGIM rules