Machine Learning for Descriptive Problems

Exercises 3

Sampling Stream Data

Exercise 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 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."

Counting Ones in a Window

Exercise 4.6.1: Suppose the window is as shown in Fig. 4.2. 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?

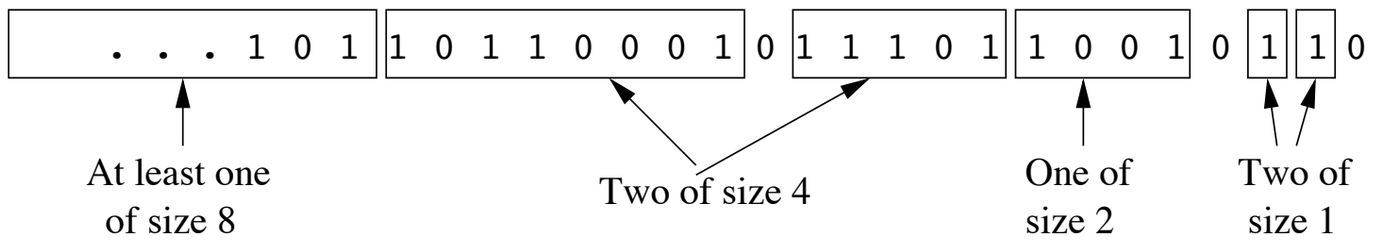

Figure 4.2: A bit-stream divided into buckets following the DGIM rules

Exercise 4.6.3: Describe what happens to the buckets if three more 1's enter the window represented by Fig. 4.3. You may assume none of the 1's shown leave the window.

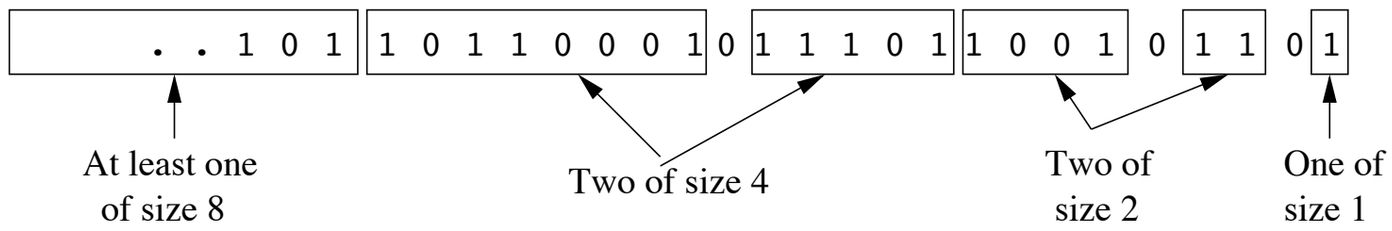

Figure 4.3: Modified buckets after a new 1 arrives in the stream