

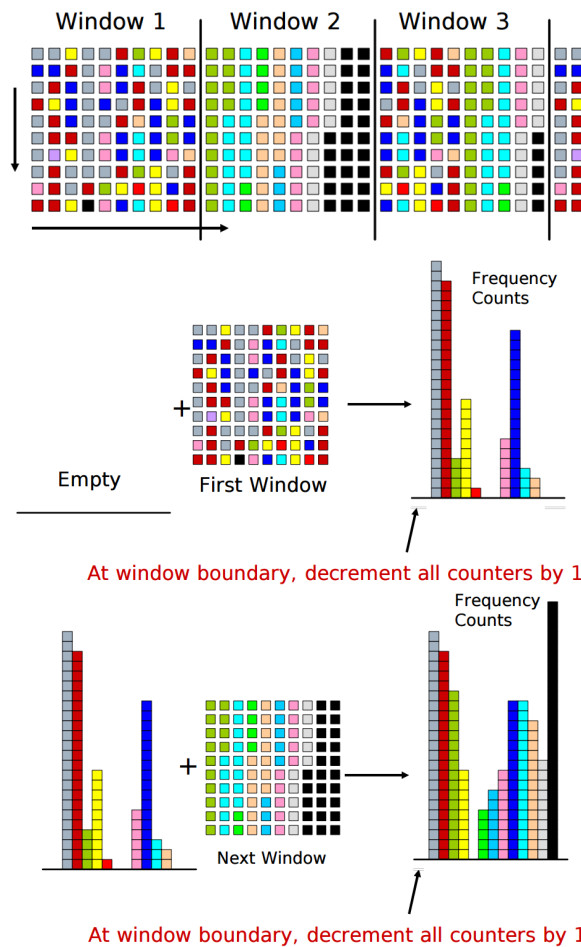
Knowledge Discovery in Databases II
 SS 2017

Exercise 7: Data Stream Clustering

Exercise 7-1 Lossy Counting

Before stream clustering, let's take a look at a more fundamental task in stream: count the occurrence of objects in a stream and output objects with a count larger or equal to some given threshold: $minSup \times L$, where L is the length of the stream up to now and $minSup$ is the given threshold (minimum support).

Lossy Counting is one of the basic algorithms that solve this problem. Given the windows size as $w = \frac{1}{\epsilon}$, the lossy counting algorithm works as the follows: cut stream into windows, process one window a time and prune histogram entries with 0 counts at each window boundary. The illustration is given below:



Please prove that

- (a) the maximum count error (the maximum difference between the real count and the estimated count) of the lossy counting algorithm is ϵL

(b) the memory consumption, i.e., the number of entries stored in the histogram, is $O(\frac{1}{\epsilon} \log(\epsilon L))$. (Optional)

Exercise 7-2 Damped Window Model

Assume a special microcluster decaying mechanism, where all microclusters are fading out after each time stamp according to $f(t) = b^{-\lambda t}$, and only the weight of the microcluster that is currently updated (hit by a point in the current timestamp) increases by 1.

- (a) What is the maximum weight of a microcluster?
- (b) What is the minimum time needed for a newly created microcluster to become potential (weight larger than τ)?
- (c) What is the minimum time needed by a potential microcluster with weight w to become an outlier (weight less than τ)?