Written Assignment #3
CIS5930: Advanced Topics in Data Management
Fall 2009

Assigned: Nov 15, 2008; Due: Dec 5, my mailbox before 5pm, 2008.

Problem 1. [40 points]

A. We modify the sticky sampling algorithm. Instead of sampling on the counter level, we sample on the per-item level, i.e, for each incoming tuple we count it with a probability ρ and discard it with a probability $1 - \rho$. Given the same parameters (ε, δ and s), please derive a required sampling ratio ρ. What's the expected memory consumption of this sampling algorithm?

B. What's the best case and worst case scenarios for the lossy counting algorithm in terms of memory consumption?

Problem 2. [30pts]

Given an input dataset D, how do we find a good value for the k in the k-means algorithm?

Problem 3. [30pts]

A. Given a data set matrix A in n dimension with m points. $SVD(A) = U \times S \times V$. We have kept the first k columns of U and obtained $A_k = A^T \times U_k$. Now, suppose you only have A_k and U_k (no knowledge about A, U, S, V), how do you reconstruct A? What is the reconstruction error? What if you know the covariance matrix $\sum_A$ of A?

B. What is an important underlying condition in order for the FastMap to work?