University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Fall 2014

Assignment 1

This is the first part of Assignment 1. This part of the assignment must be handed in at the beginning of the class in the week of October 27 (Thursday or Friday). You should turn in the code for Question 1, and submit the remaining questions either electronically, or on paper. For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course.

Question 1 (Reservoir Sampling)

In this question you are required to modify the simple (no weights) Reservoir Sampling algorithm to sample $\boldsymbol{k}$ items from a stream of N items.

1. Prove that if I pick k items uniformly at random from a set of N items, the probability of a specific item to be in the selected set is $\frac{k}{N}$.
2. Describe the algorithm for sampling k items uniformly at random from a stream of N items. The algorithm should work in a single pass over the data, reading the items one by one, without prior knowledge of the size of the stream N, and using $O(k)$ of memory (assume the size of an item is fixed).
3. Prove that your algorithm produces a uniform sample, that is, prove that for every $i, 1 \leq i \leq N$, the i-th element has probability k/N to appear in the sample.
4. Write a program that implements the sampling algorithm (in any language you want). Your program should sample k lines from a file. It should be possible to use the program from command line, and it should take command line parameters the value of k, the input file name, and the output file name. Therefore, your program should work as follows: “`sample <k> <inputfile> <outputfile>`”.

Question 2

On the Assignments page of the course there are two files: “dataset1.txt” and “dataset2.txt”. Each file contains two tab-separated columns of 100 values. Each column corresponds to a different time-series. Imagine that you are a data analyst and you want to find out if there is some relationship between the two time-series. For each dataset, analyze the data and write a short report about the kind of analysis that you did and your findings. The report should have convincing evidence (with numbers and/or plots) about the relationship between the two time-series.