COMP5331 Knowledge Discovery in Databases (Fall Semester 2019)
Homework 3
Deadline: 27 Nov, 2019 10:30am
(Please hand in during lecture though there will be project presentations.)
Full Mark: 100 Marks

Coupon Instructions:

1. You can use a coupon to waive any question you want and obtain full marks for this question.
2. You can waive at most one question in each assignment.
3. You can also answer the question you will waive. We will also mark it but will give full marks to this question.
4. The coupon is non-transferrable. That is, the coupon with a unique ID can be used only by the student who obtained it in class.
5. Please staple the coupon to the submitted assignment.
6. Please write down the question no. you want to waive on the coupon.

Q1 [20 Marks]

(a) Consider Sticky Sampling Algorithm.
Prove that Sticky Sampling Algorithm has an ε-deficient synopsis with probability at least $1-\delta$.
(Hints: You can read the paper about Sticky Sampling Algorithm for reference. The proof there is too brief. Please use your own words and elaborate the proof in this question.)
(b) What are the differences between Sticky Sampling Algorithm and Lossy Counting Algorithm?
(c) What are the differences between Lossy Counting Algorithm and Space-Saving Algorithm?

Q2 [20 Marks]

Consider the Lossy Counting Algorithm.
(a) Let $s = 0.4$ and $\varepsilon = 0.2$.
The data streams come in the following ordering. We just show the first 16 data events.
$I_3, I_2, I_4, I_3, I_5, I_4, I_2, I_4, I_2, I_4, I_1, I_6, I_2, I_7, I_1, I_7, \ldots$

(i) Let w be the width of each bucket used in the Lossy Counting Algorithm.
What is the value of w?
(ii) The Lossy Counting Algorithm updates the storage content whenever it reaches the bucket boundary. Let us call this update a cleaning action. Consider that it reads all data events in the first bucket.
(1) Before the cleaning action, what storage content does the algorithm store?
(2) After the cleaning action, what storage content does the algorithm store?
(iii) Consider it proceeds to read all data events in the second bucket.
(1) Before the cleaning action, what storage content does the algorithm store?
(2) After the cleaning action, what storage content does the algorithm store?
(iv) Consider it proceeds to read all data events in the third bucket.
(1) Before the cleaning action, what storage content does the algorithm store?
(2) After the cleaning action, what storage content does the algorithm store?
(v) Continue the scenario from (a)(iv)(2). Suppose we want to obtain the output from the algorithm at this moment. What is the output of the algorithm according to the storage content?

(b) In class, we learnt that the memory consumption of this algorithm is $\lceil 1/\varepsilon \log (\varepsilon N) \rceil$ entries.
In class and in the above question, we learnt that there is only one item at one time instance. Suppose that there is only one transaction instead of one item at one time instance where a transaction is a set of items. Let m be the maximum size of a transaction.

(i) What is the greatest possible memory consumption of this algorithm in this case?

(ii) Is the greatest possible memory consumption in (i) *nicely bounded*? A memory consumption is said to be nicely bounded if the consumption does not include any exponential factor on some variables which may be potentially large. One nicely-bounded example is "$N^2 \log N$" and one non-nicely bounded example is "2^N".

Q3 [20 Marks]

We want to find frequent items over a sliding window.
Suppose that we want to re-use the Space-Saving algorithm.

Assume that we use the batch-based approach for this purpose. Let B be the batch size. The first B-th data points form the first batch. The next B-th data points form the second batch. We can also form other batches for the remaining data points.

(a) Suppose that we are interested in finding frequent items over the 4 recent batches after reading $100 \times B$ data points. Note that $100 \times B$ is a multiple of the batch size B.
(i) Please design an algorithm for this problem such that the error introduced (in terms of the frequency of an item in fraction) is minimized.
(ii) What is the greatest error in any estimated frequency (in fraction) for the above algorithm?

(b) Suppose that we are interested in finding frequent items over the most recent 4B data points after reading $100.5 \times B$ data points. Note that $100.5 \times B$ is *not* a multiple of the batch size. Besides, the most recent 4B data points discussed above come from 4 complete batches and 1 incomplete batch. Specifically, if the 4 complete batches are B_1, B_2, B_3 and B_4 where B_1 is the oldest batch, then *conceptually*, some older data points in B_1 should not be used for finding frequent items but some newer data points in B_1 should be used. Besides, conceptually, all data points in B_2, B_3 and B_4 should be used. All data points in the incomplete batch should be used.
(i) Please design an algorithm for this problem such that the error introduced (in terms of the frequency of an item in fraction) is minimized.
(ii) What is the greatest error in any estimated frequency (in fraction) for the above algorithm?

Q4 [20 Marks]

Suppose that we want to find the skyline with a sliding window of size N where N is a positive integer and a user parameter. In order to achieve this purpose, we can use a data structure D to find the skyline with a sliding window of size N.

Please suggest an algorithm with data structure D for this problem. You can design data structure D with the following requirement.

- Suppose that S is the skyline over the most recent N points. The time complexity to output the skyline over the most recent N points (if needed) is O(|S|).
- D should include a graph where each vertice/node in the graph corresponds to a point in the sliding window (which has not been removed for the previous requirement).

Q5 [20 Marks]

Assume that there are four web sites P, Q, R and S. Suppose that R points to S. Besides, Q and S point to each other, and P and R point to each other.

(a) What is the stochastic matrix created by the page rank method for the four sites?

(b) Is there any spider trap in this example? If yes, please give the set of web sites which form a spider trap. Otherwise, please explain why there is no spider trap in this example.

(c) What equation will be solved to decide on the ranking, taking into the consideration of the possibility of the spider trap problem? There could be a lot of possible equations. For the sake for consistency, please use the equation taught in the lecture.

(d) What will be the resulting ranking according to the equation in Part (c)?