Exam

Algorithms for Data Science
Université Paris-Saclay, M2 Data Science

October 23rd, 2020

This is the final exam for the Algorithms for Data Science class, which will determine 50% of your grade for this class. The duration of the exam is *two hours*. This exam subject consists of 3 exercises and has 2 pages.

The exam is *strictly personal*: any communication or influence between students, or use of outside help, is prohibited. Any violation of the rules may result in a grade of 0 and/or disciplinary action.

Exercise 1 – Frequent Items, Association Rules, A-Priori (8 points)

Assume we have a market basket model, having 100 items, numbered from 1 to 100, and 270 baskets, numbered from 1 to 270. Items i is in basket b if b divides $3i$ with no remainder. For instance, item 1 is present in all baskets which are multiples of 3.

Questions:

1. How many items are in a basket b? How many baskets have items? Give a formula or explain it in English.
2. Which are the items having support thresholds 1 and 5? You do not need to write all of them; you can just give a description and justification for each case.
3. What are the *pairs* of items having support threshold 5? You do not need to write all of them; you can just give a description and justification for each case. *Explain* how you constructed the pairs of items.
4. Give the confidence of the association rules $\{2, 4, 6\} \rightarrow 1$ and $\{2, 3\} \rightarrow 5$.
5. Explain, step by step, the functioning of the *A-Priori algorithm* for support threshold 5. At each step, explain the data structures you used and how they were constructed.

Exercise 2 – Counting Ones in a Window (7 points)

Consider the following stream of bits, where the rightmost element is the most recent one:

```
... 1 1 1 1 0 0 1 0 1 1 1 0 0 1
```

We aim to estimate the number of 1s in the last k bits using the DCIM algorithm.

Questions:

1. Assume DCIM has created the following buckets, where the rectangles represent the buckets of increasing number of 1s inside:

   $$\ldots\quad \boxed{\texttt{1 1 1 1}}\ \texttt{0 0}\ \boxed{\texttt{1 0 1}}\ \boxed{\texttt{1 1}}\ \texttt{0 0}\ \boxed{\texttt{1}}$$

   Explain how DCIM estimates the number of 1s from these buckets, for $k = 5$ and $k = 11$. If it is the cases, explain why the estimation of the number of 1s is wrong.

2. Give one other way to divide the window into buckets, while respecting the DCIM restrictions on buckets.

3. Assume the following bits arrive in the stream in order: 0, 1, 0, 0, 1, 1. Explain how the buckets are updated from the setting in Question 2.1 and what is the final result.

4. Take the following stream of integers between 0 and 3, where the rigthmost element is the most recent one:

   ... 2 3 2 1 1 1

   Explain and exemplify how you can use 2 streams and DCIM to estimate the sum of the last 4 elements of the stream.

Exercise 3 – Counting Distinct Items (5 points)

Consider the following stream:

3 1 1 1 2 3 3 5 5 1

We want to estimate the number of distinct items using several has functions of the form $ax + b \mod 32$, i.e., hash the numbers on 5 bits. We want to use the estimate of Flajolet-Martin, which uses the number of tail 0s in the hash representation to estimate the number of distinct items in the stream.

For each of the following hash functions and the above stream:

- $h_1(x) = 3x + 1 \mod 32$
- $h_2(x) = x + 2 \mod 32$
- $h_2(x) = 5x + 7 \mod 32$

explain the functioning of the Flajolet-Martin algorithm and the estimation of the number of distinct items. Compare it with the real number of distinct items in the above. How would you improve the estimation? Exemplify.