Assignment 4: Scalable Data Mining

Name: **Waghe Shubham Yatindra** Roll No.: **13MF3IM17**

Input Stream : [1 0 1 1 0 0 0 1] 0 [1 1 1 0 1] [1 0 0 1] 0 [1] [1] 0

1. What is the largest possible bucket size for N = 22?
Ans : Largest Possible Bucket size is given by O ($\log_2 22$) = 5

2. What is the estimate of the number of 1's in the latest k = 15 bits of this window?
Ans :

Estimate is given by
[1 0 1 1 0 0 0 1] 0 [1 1 1 0 1] [1 0 0 1] 0 [1] [1] 0
1+1+2+ 4/2 = 6 (1s)

3. The following bits enter the window, one at a time: 1 0 1 1 1 0 0 1. What is the bucket configuration in the window after this sequence of bits has been processed by DGIM?
Ans :

After 1st bit entrance , (1)
[1 0 1 1 0 0 0 1] 0 [1 1 1 0 1] [1 0 0 1] 0 [1 1] 0 [1]
After 2nd Bit , (0)
[1 0 1 1 0 0 0 1] 0 [1 1 1 0 1] [1 0 0 1] 0 [1 1] 0 [1] 0
After 3rd Bit ,(1)
[1 0 1 1 0 0 0 1] 0 [1 1 1 0 1] [1 0 0 1] 0 [1 1] 0 [1] 0 [1]
After 4th Bit ,(1)
[1 0 1 1 0 0 0 1 0 1 1 1 0 1] [1 0 0 1 0 1 1] 0 [1 0 1] [1]
After 5th Bit ,(1)
[1 0 1 1 0 0 0 1 0 1 1 1 0 1] [1 0 0 1 0 1 1] 0 [1 0 1] [1] [1]
After 6th Bit ,(0)
[1 0 1 1 0 0 0 1 0 1 1 1 0 1] [1 0 0 1 0 1 1] 0 [1 0 1] [1] [1] 0
After 7th Bit ,(0)
[1 0 1 1 0 0 0 1 0 1 1 1 0 1] [1 0 0 1 0 1 1] 0 [1 0 1] [1] [1] 0 0
After 9th Bit ,(1)
[1 0 1 1 0 0 0 1 0 1 1 1 0 1] [1 0 0 1 0 1 1] 0 [1 0 1] [1 1] 0 0 [1]

4. After having processed the bits from (3), what is now the estimate of the number of 1's in the latest k = 15 bits of the window?
Ans :

Current Input Stream : [1 0 1 1 0 0 0 1 0 1 1 1 0 1] [1 0 0 1 0 1 1] 0 [1 0 1] [1 1] 0 0 [1]
For k = 15 the estimate is given by : 1 + 2 + 2 + 4/2 = 7 (1s)

5. In the file extension_DGIM.pdf you find 2 slides that explain how to generalize the DGIM algorithm from a bit stream to positive integers. Analogously to the slide example, work out the bit streams for the following stream of 8 numbers (oldest first): (125, 2, 77, 5, 13, 9, 99, 56). Compute the result for k = 3.

Ans :

Convert Numbers to m (7) bit binary format

125 = 1 1 1 1 1 0 1
2 = 0 0 0 0 0 1 0
77 = 1 0 0 1 1 0 1
5 = 0 0 0 0 1 0 1
13 = 0 0 0 1 1 0 1
9 = 0 0 0 1 0 0 1
99 = 1 1 0 0 0 1 1
56 = 0 1 1 1 0 0 0

---

c1 ,c2, c3 ,c4 ,c5, c6, c7 - Seven Different Streams (right to left) , The sum of the integers is calculated as follows:

$$\sum_{i=1}^{7} c_i 2^i$$

First Stream : [1 0 1] [1 1] [1] [1] 0
C1 = 1 + 1
Second Stream: 0 [1] 0 0 0 [1] 0
C2 = 1
Third Stream : [1 0 1] [1] [1] 0 0 0
C3 = 0
Fourth Stream : [1 0 1] 0 [1 1] 0 [1]
C4 = 1 + 2/1 = 2
Fifth Stream : [1] 0 0 0 0 0 0 [1]
C5 = 1
Sixth Stream : [1 0 0 0 0 0 1] [1]
C6 = 1 + 2/2 = 2
Seventh Stream : [1 0 1] 0 0 0 [1] 0
C7 = 1

$$\sum_{i=1}^{7} c_i 2^i = 2(2^0) + 1(2^1) + 0(2^2) + 2(2^3) + 1(2^4) + 2(2^5) + 1(2^6)$$

= 2 + 2 + 0 + 16 + 16 + 64 + 64
= 164