

Assignment - 6 Stream Algorithm

Question 2: Estimate the surprise number of a data stream using the method of RMS.

$2m-1$ m is the number of occurrence of the element of the stream at that timestamp.

Stream has 1, 2, ..., 10 cycles repeatedly.
Timestamp current $[m = 75]$

Upto 75

1-5

↓ Series
8

6-10

↓ Series
7

Range of 1-5 is 5

Range of 6-10 is 5

$$\text{Surprise number} = (5 \times 8^2) + (5 \times 7^2) = 565$$

Let's take $\{24, 44, 65\}$

Our estimate will be the median of three resulting values.

$$n(2m-1)$$

$$24 = 75[2(6)-1] = 825$$

$$44 = 75[2(4)-1] = 525$$

$$65 = 75[2(2)-1] = 225$$

Since its median has 525 nearest to 565 we choose 44.

Question 2: DDM Algorithm

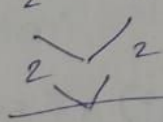
End time	100	98	95	92	87	80	65
Size	1	1	2	2	4	8	8

Sliding window length = 40

Current time stamp = 100

Since there are 3 buckets of 1's we combine earlier buckets of 1's by 2 [according to DDM algorithm]

102	101	100	95	92	87	80	65
1	1	2	2	2	4	8	8



Since 3 2's are in a row, combine to bucket of 4.

103	102	101	100	95	87	80	65
1	1	1	2	4	4	8	8

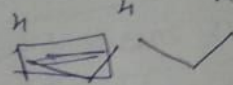
104	103	102	100	95	87	80	85
1	1	2	2	4	4	8	8

As 105th stream arrives, again 3 1's appear

105	104	103	102	100	95	87	80	65
1	1	1	2	2	4	4	8	8

105	104	102	100	95	87	80	65	combine 2 2's to get 4
1	2	2	2	4	4	8	8	

105 104 102 95 87 80 65
1 2 4 4 4 8 8



combine 2
4's 10 get 8

105 104 102 95 80 65
1 2 4 8 8 8

combine 2
9's to get 16

105 104 102 95 80
1 2 4 8 16

~~So size~~

Size of bucket of 105 is 2.

Question 3: Count no. of distinct elements in a stream.

x	$h(x)$	binary
1	10	1010
2	2	0010
3	5	0101
4	8	1000
5	0	0000
6	3	0011
7	6	0110
8	9	1001
9	1	0001
10	4	0100

hash function \rightarrow modulo
 $\hookrightarrow (3x+7) \% 11$

We consider an option

$$\{1, 3, 6, 8\} \Rightarrow \{1+0+0+0\} = 1$$

$$\{2, 6, 8, 10\} = \{1+0+0+2\} = 3$$

$$\{2, 6, 8, 9\} = \{1+0+0+0\} = 1$$

$$\{2, 5, 7, 10\} = \{1+1+1+2\} = 5$$

$\{2, 5, 7, 10\}$ is violated since it has more than 2
0's on right side. so we choose $\{2, 6, 8, 10\}$
since it is closer to our estimate.

Question 4

$$\text{Users} = 10^8$$

$$\text{Sample} = 10^{10} \text{ bytes}$$

User IDs will be hashed to a bucket number from 0 to 999,999.

Threshold (t) \rightarrow 100 byte records for all the users whose IDs hash to t or less will be retained and other users records will not be retained.

Considering $n = 10^9$ then threshold t is

$$t = \frac{10^{14}}{10^{10}} - 1 = 10^5 - 1 = 99999 //$$

Question 5 Bit array length = 100 \rightarrow t

set has = 23 members \rightarrow d

Initially all are 0's and a bit 1 is set whenever a member of set hashes to it.

$$\text{Fraction of 0's} = e^{-hd/t}$$

$$= e^{-23/100}$$

$$\boxed{h = 2}$$

No. of hash functions

$$\text{Fraction of 1's} = 1 - e^{-hd/t}$$

$$= 1 - e^{-23/100}$$