Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Daniyal Kazempour

Munich, 20.12.2016

Big Data Management and Analytics

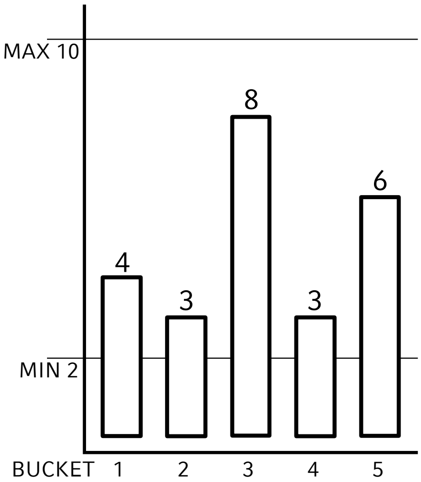
WS 2016/17

Tutorial 9: Stream Algorithms

Assignment 9-1 *K-Buckets*

Given the histogram as seen below, execute the K-Buckets Histogram algorithm for inserts and deletes, assuming the following rules:

- The histogram consists of constantly $k = 5$ buckets.
- The upper threshold (MAX) per bucket is 10, the lower threshold (MIN) is 2.
- For split-and-merge operations: a split occurs when the size of a bucket would otherwise **exceed** MAX; a merge occurs between the two consecutive buckets that were not product of the preceding split with the lowest overall sum of sizes.
- For merge-and-split operations: a merge occurs with the neighbour bucket that has the smallest size, when the size of a bucket would otherwise be below MIN.



INSERTING Insert the items of the given sequence into the histogram, until the first overflow occurs. Execute the resulting split-and-merge and move on to the next section (deleting). Each item is denoted as the index of its respective bucket.

$$\text{Sequence} = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3$$

DELETING Starting with the resulting histogram of the insert section, remove the items of the given sequence from the histogram, until the first underflow occurs. Execute the resulting merge-and-split. Each item is denoted as the index of its respective bucket.

$$\text{Sequence} = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2$$

Assignment 9-2 *CUSUM – Change Detection*

Given a mean value $\omega = 3$ and a threshold value $\alpha = 8$, execute the Cumulative Sum algorithm for change detection on the following sequence:

$$\text{Sequence} = 2, 3, 7, 4, 0, 2, 5, 6, 8, 7$$

n	$x_n - \omega$	G_n
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Assignment 9-3 *Exponential Histograms*

For the given sequence, construct an Exponential Histogram using a window size $N = 8$ and an error parameter $\epsilon = 1/2$.

$$\text{Sequence} = \times, \times, \circ, \times, \circ, \circ, \times, \times, \times, \times, \circ, \times, \times, \circ, \times, \times$$

Estimate the number of $\times$ within the window at time $t = 13$ and compare it to the actual number.

Assignment 9-4 *Hoeffding trees*

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1-2$ years, $2-7$ years, >7 years)
- Gender (male, female)
- Residential area (urban, rural)

These are the first 8 examples.

Person	Time since license	Gender	Area	Risk class
1	$1-2$	m	urban	low
2	$2-7$	m	rural	high
3	>7	f	rural	low
4	$1-2$	f	rural	high
5	>7	m	rural	high
6	$1-2$	m	rural	high
7	$2-7$	f	urban	low
8	$2-7$	m	urban	low

- Incrementally construct a Hoeffding tree for this example.
  Use information gain and $\delta = 0.2$ and $N_{\text{min}} = 2$.
- Compute the value of δ at which the tree would still consist of the leaf only.