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

Munich, 05.12.2018

Big Data Management and Analytics

WS 2018/19

Tutorial 7: Stream Applications and Algorithms

Assignment 7-1 *Exponential Histograms*

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

Sequence = $\times$, $\times$, $\circ$, $\times$, $\circ$, $\circ$, $\times$, $\times$, $\times$, $\times$, $\circ$, $\times$, $\times$, $\circ$, $\times$,$\times$

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

Assignment 7-2 *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 $N_{\text{min}} = 2$. Use log_2 for the computation of information gain and entropy.
- Compute the value of δ at which the tree would still consist of the leaf only.

Assignment 7-3 *Lossy Counting*

Given the following excerpt from a data stream S:

Time t	1	2	3	4	5	6	7	8	9	10	11	12
Item e	A	B	C	C	A	C	B	A	C	C	A	C

Perform the Lossy Counting algorithm with the error threshold $\epsilon = 0.25$. Show after every iteration of the algorithm the content of the lookup table D.