Algorithmic Data Analysis

Esther Galbrun
Autumn 2021
Q3.1: Streaming paradigm

Name the main constraints encountered when mining streams

i) Data points are processed as they come or lost for ever

ii) The number of distinct values is so large that counting them or computing basic statistics is challenging

iii) At peak periods the system has to shed part of the load

iv) The data comes from a non-stationnary distribution
Name these two families of algorithms, which are naturally suited to the streaming setting

i) always have an answer ready, which gets better as they learn

ii) do not require access to the entire training data at once
### Q3.3: Inequalities

Piece the inequalities together

<table>
<thead>
<tr>
<th>Inequality</th>
<th> </th>
<th>Inequality</th>
<th> </th>
<th>Inequality</th>
<th> </th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X \leq (1 - c)E[X]) \leq e^{-E[X]c^2/2}$</td>
<td>Chebychev</td>
<td>$P(X \geq c) \leq E[X]/c$</td>
<td>Chernoff</td>
<td>$P(</td>
<td>X - E[X]</td>
</tr>
<tr>
<td>$P(E[X] - X \geq c) \leq e^{-\frac{2c^2}{\sum r_i^2}}$</td>
<td>Markov</td>
<td>$P(E[X] - X \geq c) \leq e^{-\frac{2c^2}{\sum r_i^2}}$</td>
<td>Markov</td>
<td>$P(E[X] - X \geq c) \leq e^{-\frac{2c^2}{\sum r_i^2}}$</td>
<td>Markov</td>
</tr>
</tbody>
</table>

$X$ is a random variable  

$X$ takes only nonnegative values  

$X = \sum Y_i$, $Y_i \sim B(1, p_i)$ (Bernoulli)  

$Y_i$ has range of bounded size $r_i$  

$Y_i$s are independent random variables  

$c \in [0, 1]$  

$c$ is a constant  

$c \geq 0$  

$c$ satisfies $E[X] \leq c$  

This is a bound on both tails  

This is an upper-tail bound  

This is a lower-tail bound  

There is a similar upper-tail bound  

There is a similar lower-tail bound
Q3.4: Bloom-filters

Consider a Bloom filter of size $m$ storing $n$ distinct values. Assume that the number of hash functions $k$ is increased.

What can you say about the probability of false positives?

What can you say about the probability of false negatives?
## Q3.5: Tools for purposes

Associate tools and purposes

### Tools
- Alon–Matias–Szegedy sketch
- Approximate counting
- Bloom filters
- Count-min sketch
- Flajolet–Martin algorithm
- Hoeffding trees
- Lossy counting algorithm
- Reservoir sampling

### Purposes
- Count distinct values
- Count item occurrences
- Estimate average value
- Estimate the join size of two sets
- Estimate the intersection size of two sets
- Estimate quantiles
- Estimate zeroth-order moment
- Estimate first-order moment
- Estimate second-order moment
- Identify frequent items
- Test set-membership
- Train a classifier
Q3.6: Hash functions

Which of these methods use hash functions as an ingredient?

- Alon–Matias–Szegedy sketch
- Approximate counting
- Bloom filters
- Count-min sketch
- Flajolet–Martin algorithm
- Hoeffding trees
- Lossy counting algorithm
- Reservoir sampling
What is the mean-median trick?

i) A method to estimate the mean of a stream of data by combining the medians of multiple samples

ii) A method to obtain a robust estimate of a random variable from a collection of weaker estimates