MS&E 317/CS 263: Algorithms for Modern Data Models, Spring 2014
`http://msande317.stanford.edu`.
Instructors: Ashish Goel and Reza Zadeh, Stanford University.

Lecture 6, 04/21/2014. Scribed by Jean-Yves Stephan.

7 Sketches applications

7.1 Estimating the number of distinct elements in a stream

Sketch definition For a stream $S = a_0, a_1, \dots, a_t$ we defined $F_k(t) = \sum_{a \in V} |f_t(a)|^k$, and particularly $F_0(t)$ the number of distinct elements seen in the stream at time t: $F_0(t) = |\{a_0, \dots, a_t\}| = |S|$.

The sketch $Min - Sketch(S) =< m_1(S), \dots, m_J(S) >$ where $m_j(S) = min_{a \in S}\ h_j(a)$ can help us estimate that number.

Recall the consistent hash functions are such that $h_j(a)$ follows a uniform distribution over $[0, 1]$ and the random variables $h_i(a), h_j(b)$ are independent if either $i \neq j$ or $a \neq b$. They can be implemented as $h_j(a) = h(j, a)$ with the following pseudo-code:

```
def h(j, a)
srand (j, a)
return rand()
```

Sketches can be combined together by:

$$\tau(Min - Sketch(S_1), Min - Sketch(S_2)) =< min(m_1(S_1), m_1(S_2)), \dots, min(m_J(S_1), m_J(S_2)) >$$

Example $S = \{1, 5, 7, 8, 9, 1\}$, $J = 3$

a	h_1	h_2	h_3
1	**.085**	.138	...
5	.865	.464	...
7	.274	.841	...
8	.399	.833	...
9	.368	**.109**	...

$$\Longrightarrow Min - Sketch(S) =< .085, .109, ... >$$

Estimating F_0 from the sketch The most natural estimator involves the mean:

$$Estimator_{mean} = mean(\frac{1}{m_1(S)}, \dots, \frac{1}{m_J(S)})$$

but this estimator has a large variance, while taking the median can offset some of the variability:

$$Estimator_{median} = \frac{\ln(2)}{median(m_1(S), \ldots, m_J(S))}$$

The median lemma gives a confidence interval for the median of iid. random variables. It will also explain the origin of the $\ln(2)$ normalization factor.

7.2 The Median lemma

For a random variable Z, let $G(x) = \mathbf{Pr}(Z > x)$ denote the residual density function.
With these notations, $Median(Z) = G^{-1}(1/2)$

Theorem 7.1 ***The median lemma***

There exists a constant c such that:

for all $\delta \in (0, 1/2)$ (accuracy) and for all $\epsilon \in (0, 1/2)$ (error probability), for all $J > \frac{c}{\delta^2} \ln(\frac{1}{\epsilon})$
If $Z_1, \ldots, Z_J$ are iid random variables of residual density G,

$$\mathbf{Pr}\Big(median(Z_1, \ldots, Z_J) \in [G^{-1}(1/2 + \delta), G^{-1}(1/2 - \delta)]\Big) \geq 1 - 2\epsilon$$

Proof: Let's proove that $\mathbf{Pr}\Big(median(Z_1, \ldots, Z_J) < G^{-1}(1/2 + \delta)\Big) \leq \epsilon$

We define the iid Bernouilli variables :

$$Y_j = \begin{cases} 1 & \text{if } Z_j < G^{-1}(1/2 + \delta) \\ 0 & \text{otherwise} \end{cases}$$

By definition of G we have:

$$\mathbf{Pr}(Z_j < G^{-1}(1/2 + \delta)) = 1 - G(G^{-1}(1/2 + \delta)) = 1/2 - \delta$$

and therefore:

$$\mu = \mathbf{E}(Y_1 + ... + Y_J) = J(1/2 - \delta) = \frac{J}{2}(1 - 2\delta)$$

By definition of a median we have:

$$median(Z_1, \ldots, Z_J) < G^{-1}(1/2 + \delta) \Longleftrightarrow Y_1 + \ldots + Y_J \geq \frac{J}{2}$$

which implies the necessary condition:

$$median(Z_1, \ldots, Z_J) < G^{-1}(1/2 + \delta) \Longrightarrow Y_1 + \ldots + Y_J \geq \frac{J}{2}(1 - 2\delta)(1 + 2\delta) = \mu(1 + 2\delta)$$

Hence applying the Chernoff bound yields:

$$\mathbf{Pr}\Big(median(Z_1, \ldots, Z_J) < G^{-1}(1/2 + \delta)\Big) \leq \exp(-.38(\frac{J}{2}(1 - 2\delta))(2\delta)^2) \leq \exp(-\frac{J\delta^2}{c})$$

For some constant c. Given our choice of J we finally get the desired result:

$$\mathbf{Pr}\Big(median(Z_1, \ldots, Z_J) < G^{-1}(1/2 + \delta)\Big) \leq \exp(-\ln(\frac{1}{\epsilon})) = \epsilon$$

■

Remark 7.1 *The condition $J > \frac{c}{\delta^2}\ln(\frac{1}{\epsilon})$ shows that the median lemma is efficient to bound the probability error ($\ln(\frac{1}{\epsilon})$ term) but not to get a precise accuracy ($\frac{1}{\delta^2}$ term).*

Application to the F_0 estimate $Estimator = \frac{\ln(2)}{median(m_1(S),\ldots,m_J(S))}$

$$\begin{aligned} G(x) = \mathbf{Pr}(m_1(S) > x) &= \mathbf{Pr}(\forall a \in S, h_1(a) > x) && (1)\\ &= \mathbf{Pr}(h_1(a_0) > x)^{|S|} && (2)\\ &= (1-x)^{|S|} && (3) \end{aligned}$$

where in (1) we used $m_1(S) = min_{a\in S}\ h_1(a)$, in (2) we used the independency of the random hash functions, and in (3) we use the fact that $h_1(a_0)$ follows a uniform $[0,1]$ distribution.

Then since $(1-x)^{|S|} = 1/2 \iff x = \frac{\ln(2)}{|S|}$, the median of $m_1(S)$ is $G^{-1}(1/2) = \frac{\ln(2)}{|S|}$. It follows from the median lemma that with high probability $median(m_1(S),\ldots,m_J(S))$ will be close to $\frac{\ln(2)}{|S|}$, which in turns implies that our estimator will be close to $|S| = F_0(S)$.

Possible concrete applications

- Estimating the number of unique viewers of tweets you make. Taking advantage of the fact that sketches can be computed in parallel on different machines and then combined together.
- Estimating the number of visits on your website that bidding on a collection of adwords would bring you.

7.3 The Min-Hash technique (for computing Jacquard Similarity)

Definition of the Jacquard similarity

$$JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Estimating the Jacquard Similarity

$$m_j(S) = < min_{a\in S} h_j(a), argmin_{a\in S} h_j(a) >$$

This random variable has the following desirable property: $\mathbf{Pr}(m_1(S_1) = m_1(S_2)) = JS(S_1, S_2)$. Therefore we can estimate the Jacquard Similarity from our previous $Min-Sketch$:

$$JS_{estimate}(Min-Sketch(S_1), Min-Sketch(S_2)) = \frac{1}{J}|\{j \text{ such that } m_j(S_1) = m_j(S_2)\}|$$

References

[1] A. Broder. *On the resemblance and containment of documents.* IEEE Computer Society, 1997.