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 10, 5/5/2014. Scribed by Liren Peng.

10.1 Uniform Sampling over windows

Given a stream $S = a_0, a_1, \ldots, a_t$, all elements are unique
Query(t): Unif(w) $\longrightarrow$ need to return a uniform sampling from $\{a_t, a_{t-1}, \ldots, a_{t-w+1}\}$

Let h be a consistent uniform hashing function, we want to find

$$argmin_{t-w+1<i<t} h(a_i)$$

We need to store $< i, a_i >$ at time t if $h(a_i) < h(a_j)$ for all $j > i$

$$S(t) = \{< i, a_i : h(a_i) < h(a_j) \forall j > i\}$$

At time t, add $< t, a_t >$ to $S(t-1)$ and delete every $< i, a_i > s.t. h(a_i) >= h(a_t)$ from $S(t-1)$

$E[|S(t)|]$ **:** Define an indicator variable

$$Z_{i,t} = \begin{cases} 1 & \text{if } < i, a_i > \in S(t) \\ 0 & \text{otherwise} \end{cases}$$

$$|S(t)| = \sum_{i=0}^{t} Z_{i,t}$$

$$E[|S(t)|] = \sum_{i=0}^{t} E[Z_{i,t}] \tag{1}$$

$$= \sum_{i=0}^{t} \frac{1}{t-i+1} \tag{2}$$

$$= \sum_{i=1}^{t+1} \frac{1}{i} \tag{3}$$

$$= ln(t+1) \tag{4}$$

10.2 Johnson-Lindenstrauss Dimensionality Reduction

Given N points in D-dimension space, $x \in \mathbb{R}^D$
Define $R_i =< z_{i,1}, z_{i,2}, \ldots, z_{i,D}$ where $z_{i,j}$ are i.i.d. $N(0,1)$

$$f(x) =< xR_1, xR_2, \ldots, xR_k >, k > \frac{c}{\delta^2} log(N)$$

$$M = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \end{bmatrix} \tag{5}$$

$$f(x) = Mx : \mathbb{R}^D \longrightarrow \mathbb{R}^k$$

$$f_{(i)}(x) - f_{(i)}(y) = (x - y) * R_i \tag{6}$$

$$= \sum_{j=0}^{D} (x_j - y_j) * z_{i,j} \tag{7}$$

$$= ||x - y||_2 \tag{8}$$

$$f(x) - f(y) = ||x - y||_2 \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_k \end{bmatrix}, Z_i \text{ are i.i.d. } N(0,1)$$

$||f(x)-f(y)||_2 = (||x-y||_2) \sum_{i=1}^{k} {Z_i}^2 \longrightarrow$ highly concentrated, so we can use "Median Lemma" for "mean"
$||f(x) - f(y)||_2 \in k * ||x - y||_2 * [1 \pm \delta]$ with probability at least $1 - \epsilon$ if $k > \frac{c}{\delta^2} log(N)$

Formally $d(f(x), f(y))$ is a good approximation of $d(x, y)$
$\epsilon = \frac{1}{N^3}, k = \frac{c * logN}{\delta^2}$
$x_1, x_2, \ldots, x_N$, for $k > \frac{c * logN}{\delta^2}$, with probability $1 - \frac{1}{N}$
$\forall i, j : 1 \leq i, j \leq N, ||f(x_i) - f(x_j)||_2 \in k||x - y||_2 * [1 \pm \delta]$

10.3 Locality Sensing Hashing

Data base of N objects from U, $S = \{x_1, x_2, \ldots, x_n\}, S \in U and w \in U$
Distance metric $d(\bullet, \bullet)$ on U
$Query(w)$: find an object in S similar to w

(C, R) near-neighbor problem Given S
$Query(w)$: if $\exists x \in S$ s.t. $d(w,x) <= R$ then return an object $y \in S$ s.t. $d(w,y) \leq cR$
Never return y s.t. $d(w,y) > cR$

A family of hash function is said to be $(c, R, P_L, P_U) - LSH$ for distance metric $d(\bullet, \bullet)$

- For all $x, y \in U$ s.t. $d(x,y) \leq R, \mathbf{Pr}_{h \sim H}[h(x) = h(y)] \geq P_L$
- For all $x, y \in U$ s.t. $d(x,y) > cR, \mathbf{Pr}_{h \sim H}[h(x) = h(y)] \leq P_U$