MTH 995-003: Intro to CS and Big Data	Spring 2014
Lecture 9 — Feb 4th, 2014	
Inst. Mark Iwen	*Scribe: Ruochuan Zhang*

1 Last time

Theorem 1. *Let* $h(\vec{x}) = \lfloor \frac{<\vec{y},\vec{x}>+u}{w} \rfloor$, *for* $\vec{y} \sim N(0, I_{D\times D})$, $u \sim U([0,w])$, *and* $w \in \mathbb{R}^+$. *Let* $r \in \mathbb{R}^+, c \in (1, \infty)$. *Then h is a LSH function with respect to* l_2 *- distance. It has*

$$p_1 = p_w(r) > p_2 = p_w(cr)$$

where
$p_w(n) = \text{erf}\left(\frac{w}{\sqrt{2}n}\right) + \sqrt{\frac{2}{\pi}}\frac{n}{w}\left[e^{-(\frac{w}{\sqrt{2}n})^2} - 1\right]$

2 This time

- Let $h : \mathbb{X} \to \mathbb{Z}$ be a *LSH* function for metric $d(\cdot, \cdot)$
  1) $d(\vec{x}, \vec{y}) < r \Rightarrow \mathbb{P}[h(\vec{x}) = h(\vec{y})] \geq p_1$
  2) $d(\vec{x}, \vec{y}) \geq rc \Rightarrow \mathbb{P}[h(\vec{x}) = h(\vec{y})] \leq p_2 < p_1$

Definition 1. *Let* $g_k : \mathbb{X} \to \mathbb{Z}^k$ *be a locally sensitive hash function created via k i.i.d. LSH functions* $h_1, h_2, ...h_k$ *defined by* $g_k(\vec{x}) = (h_1(\vec{x}), h_2(\vec{x}), ...h_k(\vec{x}))$

Definition 2. $g_k : \mathbb{X} \to \mathbb{Z}^k$ *will be "good" for a* $\vec{x} \in \mathbb{X}$ *if*
(1) $g(\vec{x}) \neq g(\vec{y}) \quad \forall \vec{y} \in \mathbb{X}$ *with* $d(\vec{x}, \vec{y}) \geq rc$
(2) $g(\vec{x}) = g(\vec{y}) \quad$ *for at least one* $\vec{y} \in \mathbb{X}$ *with* $d(\vec{x}, \vec{y}) \leq r$

Definition 3. *For* $\vec{x} \in \mathbb{X}$, *let* $\vec{x}^* = \arg\min_{\vec{y} \in \mathbb{X} - \{\vec{x}\}}(d(\vec{x}, \vec{y}))$

Fix $\vec{x} \in \mathbb{X}$. Note that

$$\begin{aligned} \mathbb{P}[(1) \; fails \; for \; \vec{x} \in \mathbb{X}] &\leq (|\mathbb{X}| - 1)\mathbb{P}[g_k(\vec{x}) = g_k(\vec{y}) \; for \; some \; \vec{y} \in \mathbb{X} \; with \; d(\vec{x}, \vec{y}) \geq rc] \\ &\leq (|\mathbb{X}| - 1)p_2^k \end{aligned}$$

$$\begin{aligned} \mathbb{P}[(2) \; fails \; for \; \vec{x} \in \mathbb{X}] &\leq 1 - \mathbb{P}\left[g_k(\vec{x}) = g_k(\vec{x^*}) \; and \; d(\vec{x}, \vec{x^*}) < r\right] \\ &\leq 1 - p_1^k \end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}\left[g_k \textit{ is ``good'' for } \vec{x} \in \mathbb{X}\right] &\geq 1 - \mathbb{P}[(1) \; fails] - \mathbb{P}[(2) \; fails] \\ &\geq p_1^k - (|\mathbb{X}| - 1)p_2^k \\ &\geq p_1^k \left(1 - |\mathbb{X}|(p_2/p_1)^k\right)\end{aligned}$$

Setting $k = \log_{\frac{p_1}{p_2}}(2|\mathbb{X}|)$, $\left(\frac{p_1}{p_2} > 1\right)$, we see that

$$\begin{aligned}\mathbb{P}\left[g_k \textit{ is good for } \vec{x} \in \mathbb{X}\right] &\geq \frac{1}{2} p_1^{\log_{\frac{p_1}{p_2}}(2|\mathbb{X}|)} \\ &= \frac{1}{2}(2|\mathbb{X}|)^{\frac{\rho}{\rho-1}}\end{aligned}$$

where $\rho := \frac{\log p_1}{\log p_2}$. (Note $\rho < 1$). We have just proven the following lemma

Lemma 1. *If we set $k \geq \log_{\frac{p_1}{p_2}}(2|\mathbb{X}|)$, then g_k will be good for $\vec{x} \in \mathbb{X}$ with probability at least $\frac{1}{2}(2|\mathbb{X}|)^{\frac{\rho}{\rho-1}}$*

The next lemma bounds the number of i.i.d. hash functions, g_k, one must pick before one can be sure that at every element of $\mathbb{X}$ will have a "good" LSH function.

Lemma 2. *If we generate*

$$L \geq 2(2|\mathbb{X}|)^{\frac{\rho}{1-\rho}} \cdot \log\left(\frac{|\mathbb{X}|}{1-\sigma}\right) \quad \textit{i.i.d.}$$

hash functions $g_k^j : \mathbb{X} \to \mathbb{Z}^k$, $j = 1, ..., L$, with $k \geq \log_{\frac{p_1}{p_2}}(2|\mathbb{X}|)$, then the following will hold with probability at least σ:

$\forall \vec{x} \in \mathbb{X} \; \exists l \in [L]$ s.t. g_k^l is a "good" LSH function for $\vec{x} \in \mathbb{X}$.

Proof. Let $\delta = \frac{1}{2}\left(\frac{1}{2|\mathbb{X}|}\right)^{\frac{\rho}{1-\rho}}$ and fix $\vec{x} \in \mathbb{X}$. All $g_k^1, ..., g_k^L$ will fail to be good for $\vec{x}$ with probability $\leq (1-\delta)^L \leq e^{-\delta L} \leq e^{\log\left(\frac{1-\sigma}{|\mathbb{X}|}\right)} = \frac{(1-\sigma)}{|\mathbb{X}|}$.
The result now follows from a union bound over all $\vec{x} \in \mathbb{X}$. □

- We can now solve the $(c, r) - NN$ (Nearest Neighbor) problem using these g_k^l, $l = 1, ..., L$.
- Let $\mathbb{X} = \{\vec{x_1}, ..., \vec{x_P}\} \subseteq \mathbb{R}^D$ and $d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|_2$

2.1 Algorithm

1. For each $\vec{x_j} \in \mathbb{X}$
2. compute $g_k^l(\vec{x_j})$ for $l = 1, ..., L$.
3. end for

4. Set $f(\vec{x_j}) = (\infty, ..., \infty)$ for $j = 1, ..., P$
5. For each g_k^l, $l = 1, ..., L$
6. For each $n \in g_k^l(\mathbb{X}) \subseteq \mathbb{Z}^k$, with $|(g_l^k)^{-1}(n)| \geq 2$ (at least two $\mathbb{X}$ elements hashed to n)
7. For each $\vec{x} \in (g_k^l)^{-1}(n)$, choose $\vec{y} \neq \vec{x}$, $\vec{y} \in (g_k^l)^{-1}(n)$
8. If $\|\vec{x} - \vec{y}\| < \min\{cr, \ \ \|\vec{x} - f(\vec{x})\|_2\}$
9. set $f(\vec{x}) = \vec{y}$
10. end for
11. end for
12. end for

The runtime from 1 to 3 is $O(PLkD)$

The runtime of 4 is $O(P)$

The runtime of lines 7 through 10 is $O\left(D\left|(g_k^l)^{-1}(n)\right|\right)$

The runtime from 6 to 11 is $O\left(DP\right)$

The runtime from 5 to 12 is $O(DPL)$

This algorithm is GOOD if it beats the simple $O(P^2D)$ - time NN algorithm.

The total runtime is : $O\left(PD\left(\log_{\frac{p_1}{p_2}} 2P\right) 2(2P)^{\frac{\rho}{1-\rho}} \log(\frac{P}{1-\sigma})\right)$.

If $\frac{\rho}{1-\rho} < 1$, we are faster!

Theorem 2. *Choose $\sigma \in (0,1)$, let $\mathbb{X} = \{\vec{x}, ..., \vec{x_P}\} \subseteq \mathbb{R}^D$. Then the $(c,r) - NN$ problem can be solved for $\mathbb{X}$ w.r.t. Euclidean distance with probability at least σ in*

$$O\left(D(2P)^{\frac{\rho}{1-\rho}+1} \cdot \log\left(\frac{P}{1-\sigma}\right) \cdot \log_{\frac{p_1}{p_2}}(2P)\right) - time$$

3 Homework

6. Let $f_{NN}(\vec{x}) = \arg\min_{\vec{y}\in\mathbb{X}-\{\vec{x}\}} \|\vec{y}-\vec{x}\|_2$ and set $\Delta := (\min_{\vec{x}\in\mathbb{X}} \|(f_{NN}(\vec{x})-\vec{x}\|_2)/(\max_{\vec{x}\in\mathbb{X}} 2\|\vec{x}\|_2)$, Prove that we can compute a function $f_{NN}^{A} : \mathbb{X} \to \mathbb{X}$, satisfying

$$\|\vec{x} - f_{NN}^{A}(\vec{x})\|_2 \leq 4\|\vec{x} - f_{NN}(\vec{x})\|_2, \quad \forall \vec{x} \in \mathbb{X}$$

with probability $\geq \sigma$ in time

$$O\left(D(2|\mathbb{X}|)^{\frac{3}{2}} \cdot \log\left(\frac{|\mathbb{X}| \cdot \log_{4/3}(\Delta^{-1})}{1-\sigma} \right) \cdot \log_{3/2}(2|\mathbb{X}|) \cdot \log_{4/3}(\Delta^{-1}) \right).$$

References

[1] Piotr Indyk, Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. *Proceeding STOC '98 Proceedings of the thirtieth annual ACM symposium on Theory of computing*, Pages 604-613, 1998.