COMMONWEALTH OF AUSTRALIA

Copyright Regulations 1969

WARNING

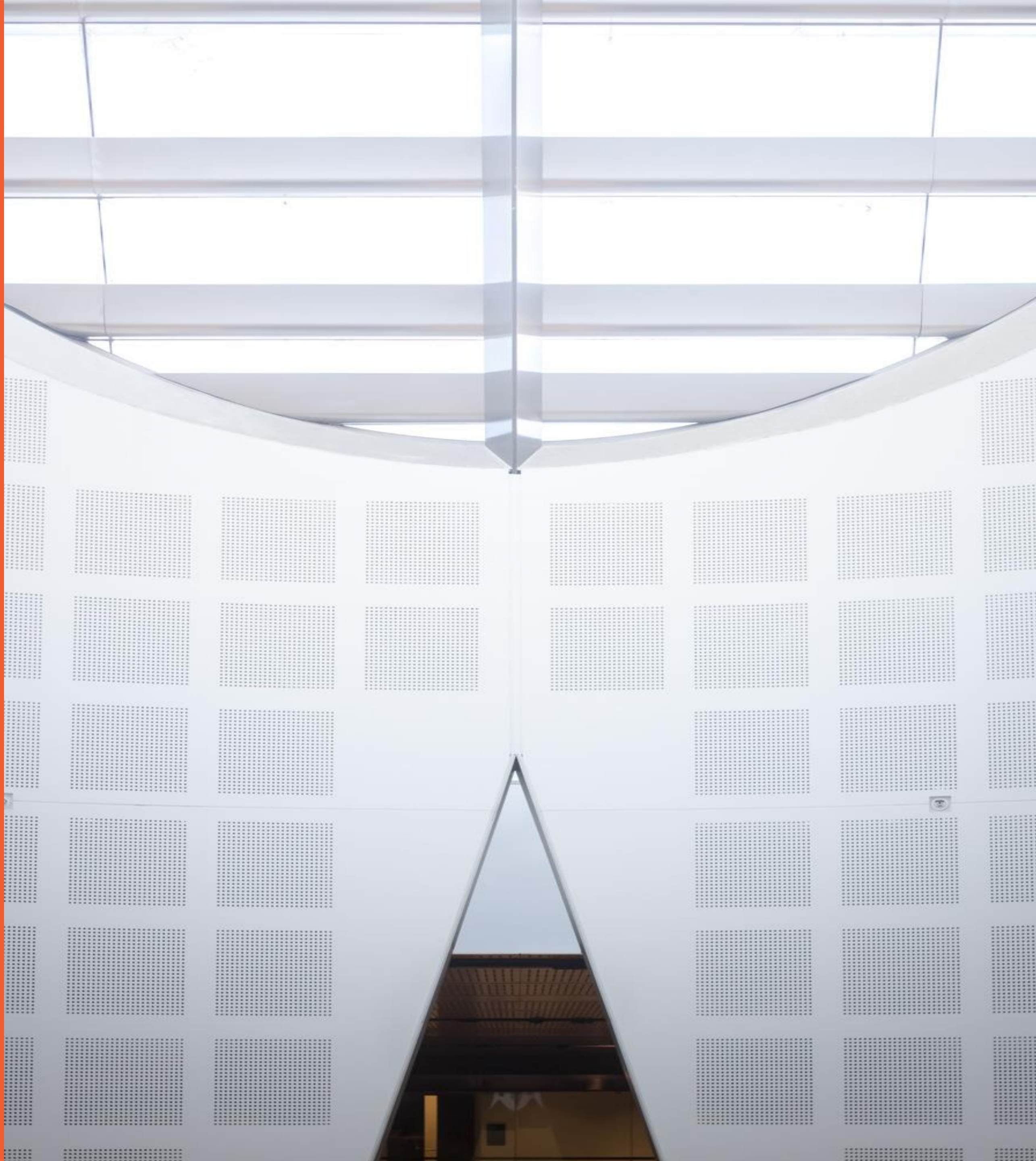
This material has been reproduced and communicated to you by or on behalf of the University of Sydney pursuant to Part VB of the Copyright Act 1968 (**the Act**). The material in this communication may be subject to copyright under the Act. Any further copying or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice.

COMPx270: Randomised and Advanced Algorithms
Lecture 7: Nearest Neighbours and dimensionality reduction

Clément Canonne
School of Computer Science

THE UNIVERSITY OF SYDNEY

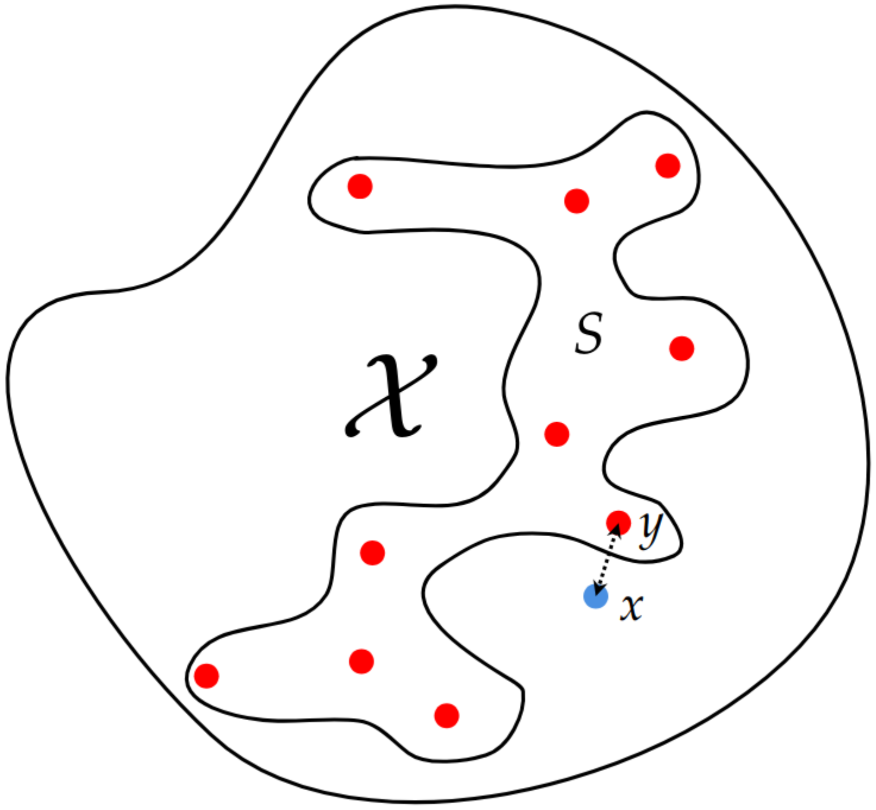


A question

You have n pictures, each 4096x4096 pixels, of venomous spiders. Someone finds a spider in their kitchen and sends you a photo, asking which type of spider it is and if it is venomous, **because they just have been bitten.**

How long will it take you?

Nearest Neighbour Search



$\cancel{|\mathcal{X}| = m}$

$\mathcal{X} = \{0,1\}^d$

or

$\mathcal{X} = \mathbb{R}^d$

+ metric d

$(d(x, y) \equiv$ how similar x, y are$)$

Nearest Neighbour Search

$d: \mathcal{X} \times \mathcal{X} \to [0, \infty)$

① $d(x,y) = 0 \iff x = y$

② $d(x,y) = d(y,x)$

③ $d(x,y) \leq d(x,z) + d(z,y)$ $\forall x, y, z$

On x: Find $y \in S$ s.t. $y = \underset{y' \in S}{\operatorname{argmin}}\, d(x,y)$

① SPACE Would like $O(n\,d)$.

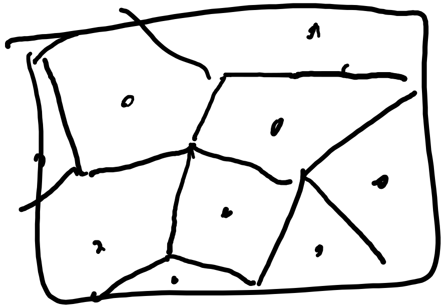
n: # of points, d: dimension

② QUERY TIME Ideally $O(1)$, but... $o(n)$ good

Examples
- $\mathcal{X} = \{0,1\}^d$ Hamming distance: $d(x,y) = \#(\text{bits } x \text{ and } y \text{ differ on})$
- $\mathcal{X} = \mathbb{R}^d$, Euclidean distance: $d(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} = \|x-y\|_2$ (ℓ_2)
- $\mathcal{X} = \mathbb{R}^d$, Manhattan distance: $d(x,y) = \sum_{i=1}^{d} |x_i - y_i| = \|x-y\|_1$ (ℓ_1)

Lists? Voronoi? K-d trees? Hash tables?

- List: Query time $O(nd)$, space $O(nd)$
- (For $\{0,1\}^d$): Query time $O(2^d)$, Space $O(2^d)$
- Voronoi



Query time $O(nd)$, Space $O(n^{\lceil d/2 \rceil})$

- ~~Hash tables~~

$d \gg 1$, $n \gg 1$

large

even larger

$1 \ll d \ll n \leq 2^{O(d)}$

Bad news...

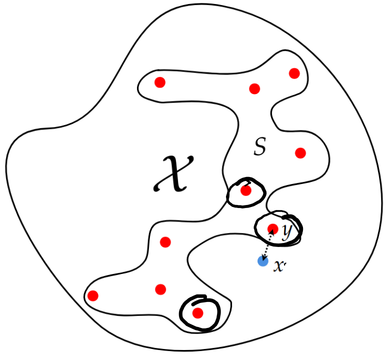
NN: we don't know how.

Either query time or space is

$\Omega(\min(2^d, nd))$

(for everything we know, even randomized algorithm)

Approximate Nearest Neighbour Search



$\textsc{Query}(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$ sort-of-minimising $\mathrm{dist}(x, y)$, that is, $\mathrm{dist}(x, y) \leq C \cdot \min_{y' \in S} \mathrm{dist}(x, y')$.

Dimensionality Reduction: the JL Lemma (Euclidean space)

$$(\mathbb{R}^d, \|\cdot\|_2) \xrightarrow{\Phi} (\mathbb{R}^k, \|\cdot\|_2) \qquad k << d$$

$$\|\phi(x) - \phi(y)\|_2 \overset{C}{\approx} \|x-y\|_2$$

Solve NN on $\mathbb{R}^k$

$\downarrow$ $(O(nk), O(nk))$

Solve ANN on $\mathbb{R}^d$

<u>JL Lemma</u> Can do this with

$C = 1+\varepsilon$

and $k = O\left(\frac{\log n}{\varepsilon^2}\right)$

s.t. if $|S| = n$, $\forall x, y \in S$ $\|\phi(x) - \phi(y)\|_2 = (1 \pm \varepsilon)\|x-y\|_2$

$\in \mathbb{R}^k$ (with prob. $\geq \frac{99}{100}$)

What is $\phi : \mathbb{R}^d \to \mathbb{R}^k$? It's linear.

$M \in \mathbb{R}^{d \times k}$

$$k \updownarrow \overset{d}{\longleftrightarrow}\left(\quad \right) \cdot \frac{1}{\sqrt{k}}$$

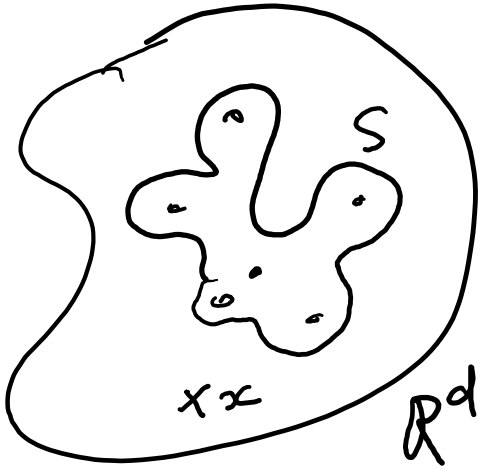
$M_{ij} \sim N(0,1)$

$\phi(x) = Mx$

JL Lemma and ANN

$\phi: \mathbb{R}^d \to \mathbb{R}^k$

$k = O\left(\frac{\log(n+1)}{\varepsilon^2}\right)$



Apply to $T = S \cup \{x\}$

Cost: <u>space</u> $O(nk) = O\left(\frac{n \log n}{\varepsilon^2}\right)$

<u>query</u>: $O(nk) = O\left(\frac{n \log n}{\varepsilon^2}\right)$

Good, but neither is $o(n)$ ...

Beyond JL Lemma: Hashing!

Spoiler: can do query time $O(n^{\rho} d)$ (expected) for some $0<\rho<1$

space $O(n^{1+\rho} d)$

(Hamming/ℓ_1: $\rho \approx \frac{1}{c}$

Euclidean $\rho \approx \frac{1}{c^2}$)

"sublinear"

"nearly" linear

ANN:

find y st

$d(x,y) \le \boxed{C} \cdot \min_{y'} d(x,y')$

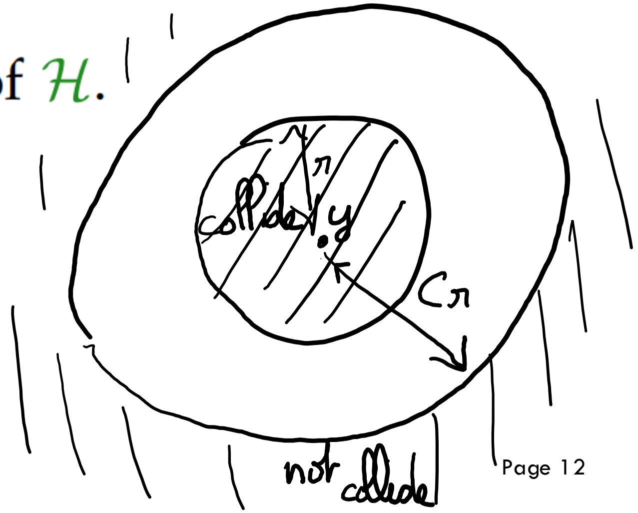
Locality-Sensitive Hashing

Definition 36.1. Let $0 \le q < p \le 1$, $r > 0$, $C > 1$, and $(\mathcal{X}, \mathrm{dist})$ be a metric space. Then a family of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathcal{Y}$ is a (r, C, p, q)-*Locality Sensitive Hash family* (LSH) if, for every $x, x' \in \mathcal{X}$,

WANT collision

- If $\mathrm{dist}(x, x') \le r$, then $\Pr_{h \sim \mathcal{H}}[h(x) = h(x')] \ge p$;
- If $\mathrm{dist}(x, x') \ge Cr$, then $\Pr_{h \sim \mathcal{H}}[h(x) = h(x')] \le q$; Do not want collision

and we say $\rho := \frac{\log(1/p)}{\log(1/q)} < 1$ is the *sensitivity parameter* of $\mathcal{H}$.

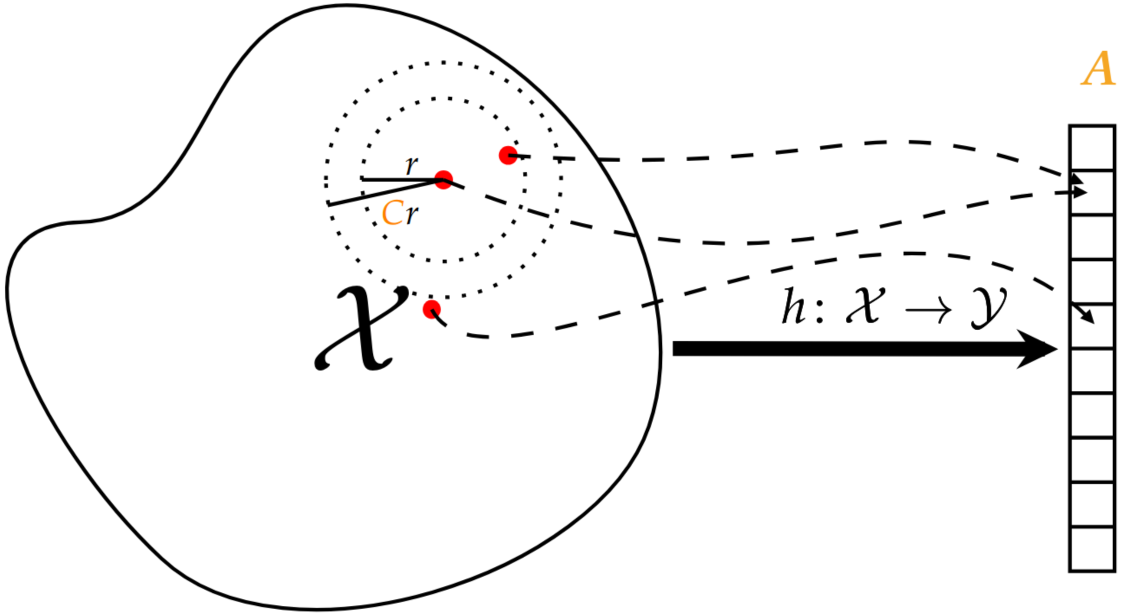


Locality-Sensitive Hashing

Definition 36.1. Let $0 \leq q < p \leq 1$, $r > 0$, $C > 1$, and $(\mathcal{X}, \mathrm{dist})$ be a metric space. Then a family of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathcal{Y}$ is a (r, C, p, q)-*Locality Sensitive Hash family* (LSH) if, for every $x, x' \in \mathcal{X}$,

- If $\mathrm{dist}(x, x') \leq r$, then $\Pr_{h \sim \mathcal{H}}[\, h(x) = h(x') \,] \geq p$;
- If $\mathrm{dist}(x, x') \geq Cr$, then $\Pr_{h \sim \mathcal{H}}[\, h(x) = h(x') \,] \leq q$;

and we say $\rho := \frac{\log(1/p)}{\log(1/q)} > 1$ is the *sensitivity parameter* of $\mathcal{H}$.



Locality-Sensitive Hashing: "Baby version"

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\mathrm{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\mathrm{dist}(x, y^*) \leq C \cdot r$;
- If $\mathrm{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Locality-Sensitive Hashing: "Baby version" (1/4)

QUERY$_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\mathrm{dist}(x, y^*) \leq r$, then, with probability at least 9/10, QUERY$_r(x)$ returns an element $y \in S$ such that $\mathrm{dist}(x, y^*) \leq C \cdot r$;
- If $\mathrm{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, QUERY$_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

r is fixed
$C > 1$ fixed
p, q given
$0 < q < p < 1$
$\ell = \frac{\log(1/p)}{\log(1/q)}$

Claim. From $\mathcal{H}$, can get $\mathcal{H}^{(\ell)}$ $(\ell \geq 1)$ s.t. $\mathcal{H}^{(\ell)}$ is a (r, C, p^ℓ, q^ℓ)-LSH family (and $|\mathcal{H}^{(\ell)}| = |\mathcal{H}|^\ell$)

Pf: $h(x) = (h_1(x), h_2(x), \ldots, h_\ell(x)) \in \mathcal{Y}^\ell$ $\quad (\mathcal{H}: \mathcal{X} \to \mathcal{Y})$

Suppose $d(x, x') \leq r$

$$\Pr_{h \sim \mathcal{H}^{(\ell)}}[h(x) = h(x')] = \Pr_{h_1, h_2, \ldots, h_\ell}[(h_1(x), \ldots, h_\ell(x)) = (h_1(x'), \ldots, h_\ell(x'))]$$

$$= \underbrace{\Pr_{h_1}[h_1(x) = h_1(x')]}_{\geq p} \cdots \underbrace{\Pr_{h_\ell}[h_\ell(x) = h_\ell(x')]}_{\geq p} \geq p^\ell$$

$\mathcal{H}$ is a (r, C, p, q)-LSH

Suppose $d(x, x') \geq C \cdot r$

$$= \underbrace{\quad}_{\leq q} \cdots \underbrace{\quad}_{\leq q} \leq q^\ell$$

Locality-Sensitive Hashing: "Baby version" (2/4)

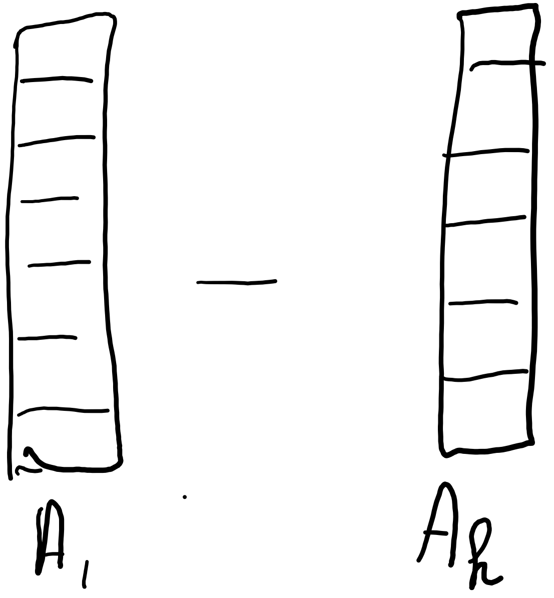
QUERY$_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\mathrm{dist}(x,y^*) \le r$, then, with probability at least $9/10$, QUERY$_r(x)$ returns an element $y \in S$ such that $\mathrm{dist}(x,y^*) \le C \cdot r$;
- If $\mathrm{dist}(x,y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, QUERY$_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

ℓ

k

Get k hash tables $A_1, \ldots, A_k$
using good standard hash functions $h_1, \ldots, h_k$
(not LSH) + chaining



In each A_t, insert the hashes of S by $g_t^{(\ell)}$
where $g_1^{(\ell)}, \ldots, g_k^{(\ell)} \sim \mathcal{G}^{(\ell)}$

$$\Big[\underbrace{S \xrightarrow{g_t^{(\ell)}}}_{\substack{\text{new} \\ \text{(for "similarity")}}} \underbrace{LSH(S) \xrightarrow{h_t}}_{\substack{\text{standard hashing} \\ \text{(for efficiency)}}} A_t\Big]$$

PREPROCESS

$\forall x \in S$
$\quad \forall\, 1 \le t \le k$
$\quad\quad A_t.\mathrm{INSERT}(g_t^{(\ell)}(x))$

QUERY

$\forall\, 1 \le t \le k$
$\quad L_t \leftarrow A_t.\mathrm{LOOKUP}(g_t^{(\ell)}(x))$
$\quad \forall y \in L_t$
$\quad\quad$ if $d(x,y) \le Cr$, return y
return $\perp$

Hope: If x, y are close at least one t of the k LSH $g_1^{(\ell)}, \ldots, g_k^{(\ell)}$ will make them collide, and so $y \in L_t$ that t

Locality-Sensitive Hashing: "Baby version" (3/4)

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\mathrm{dist}(x, y^*) \le r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\mathrm{dist}(x, y^*) \le C \cdot r$;
- If $\mathrm{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Space: k Hash tables, each n elements of size $O(d)$

$\rightarrow O(knd)$

k LSH hash functions $\quad k \times \underbrace{\ell}_{\text{one LSH function from } \mathcal{H}^{(\ell)}} \times \underbrace{O(d)}_{\text{one LSH function from } \mathcal{H}} = O(k\ell d)$

$\Big\} \; O(knd + k\ell d)$

Query time
- Evaluate $g_t^{(\ell)}(x) \quad \forall\, 1 \le t \le k : \quad O(k\ell)$
- $\mathbb{E}\left[\sum_{t=1}^{k} |L_t| \cdot O(d)\right] \approx k \cdot (nq^{\ell}) \cdot O(d) = O(kdnq^{\ell})$

only far ones (bad collisions)

$\Big\} \; O(k\ell + kdnq^{\ell})$

Locality-Sensitive Hashing: "Baby version" (4/4)

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\mathrm{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\mathrm{dist}(x, y^*) \leq C \cdot r$;
- If $\mathrm{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Correctness

When unlucky: there is $y^* \in S$ st $d(x, y^*) \leq r$

but $g_t^{(\ell)}(x) \neq g_t^{(\ell)}(y^*)$ $\forall\, 1 \leq t \leq k$

$\Pr[\text{unlucky}] = (1 - p^\ell)^k \leq \frac{1}{10}$ $(\ast)$

WANT

k, ℓ?

WANT: $(\ast)$ + $n q^\ell \leq 1$ (for query time)

set $k = O(n^\rho)$ ← set $\ell = O\left(\frac{\log n}{\log 1/q}\right)$

Space $O(n^{1+\rho} d)$

$\mathbb{E}[\text{Query time}] = O(n^\rho d)$

Locality-Sensitive Hashing: "Baby version" (👶)

Locality-Sensitive Hashing: "They grow up so fast" (🧑)

"Binary search"

↳ solves ANN from baby version losing only a log d factor.

Locality-Sensitive Hashing: But... do they exist?

Hamming $\{0,1\}^d$

Given c, r

If $d(x,x') \le r$,

$$\Pr_h[h(x)=h(x')] \ge 1 - \frac{r}{d}$$

If $d(x,x') \ge c \cdot r$

$$\Pr_h[h(x)=h(x')] \le 1 - \frac{cr}{d}$$

$h_1(x) = x_1 \in \{0,1\}$

$h_2(x) = x_2$

$\vdots$

$h_d(x) = x_d$

$\mathcal{H} = \{h_i : i \in [d]\}$

$$\rho = \frac{\log(1/p)}{\log(1/q)} = \frac{\log\left(1 - \frac{r}{d}\right)}{\log\left(1 - \frac{cr}{d}\right)} \approx \frac{1}{c}$$

Locality-Sensitive Hashing: But... do they exist?