Approximate Nearest Neighbor in High Dimensions

what if d is not constant?

[lots of apples]
["curse of dimensionality"
all methods thus far have exp. dep. on d]

Method 1: Dimension Reduction (Indyk-Motwani '98)

Def Given n pts in $\mathbb{R}^d$,
an **ε-embedding** into k dims is a map $f: P \to \mathbb{R}^k$
s.t. $\forall p, q \in P$,

$$c_0(1-\varepsilon)\|p-q\| \leq \|f(p)-f(q)\| \leq c_0(1+\varepsilon)\|p-q\|$$

Johnson-Lindenstrass Lemma (1984) For any n pts in $\mathbb{R}^d$,
$\exists$ ε-embedding into $O\left(\frac{1}{\varepsilon^2}\log n\right)$ dims.

Pf: Very Simple Randomized Algm:
just choose a random projection f!
more precisely, let $f: \mathbb{R}^d \to \mathbb{R}^k$ be

$$f(q_1, \ldots, q_d) = \left(\sum_{j=1}^{d} X_{1j} q_j, \ldots, \sum_{j=1}^{d} X_{kj} q_j\right)$$

where coefficients X_{ij} are indep rand vars
of normal distribution $\sim N(0,1)$
(mean, variance)

Analysis: (sketch)
Fix $p, q \in P$.
Then $\|f(p)-f(q)\|^2 = \sum_{i=1}^{k} Y_i^2$

where Y_i = i^{th} coord of $f(p)-f(q)$

$$= \sum_{j=1}^{d} X_{ij}(p_j - q_j)$$

(linear comb. of normal vars is normal)
($\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$)

is normally distributed $\sim N\left(0, \sum_{j=1}^{d}(p_j-q_j)^2\right)$

$$= N(0, \|p-q\|^2)$$

Let $\mu = E\left[\sum_{i=1}^{k} Y_i^2\right] = \sum_{i=1}^{k} \mathrm{Var}[Y_i]$

$= k\|p-q\|^2$

By Chernoff bound (for squared normal distribution),

$\Pr\left\{\sum_{i=1}^{k} Y_i^2 \notin [(1-\varepsilon)\mu, (1+\varepsilon)\mu]\right\} \le e^{-k\varepsilon^2/4}$

Set $k = \frac{12}{\varepsilon^2}\log n$ $\le e^{-3\log n}$

$= \frac{1}{n^3}$

$\Rightarrow \Pr\{\|f(p)-f(q)\|^2 \notin [(1-\varepsilon)k\|p-q\|^2, (1+\varepsilon)k\|p-q\|^2\}$ for some $p,q \in P] \le \frac{1}{n}$ □

To solve approx decis problem for fixed r:

1. apply embedding
2. then use grid method with

query time $\tilde{O}(1)$ ($\tilde{O}$ hides poly(d) factors)

space $O\left(\left(\frac{1}{\varepsilon}\right)^k n\right)$

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

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

polynomial space for any const ε regardless of d!

Rmk - extends to ANN queries

- many other embedding results for diff. metrics (extends to L_p, $1 \le p \le 2$)
- Ailon-Chazelle '06: faster JL transform ($O(kd)$ time $\to \tilde{O}(d)$)

Method 2: Locality Sensitive Hashing (Indyk-Motwani)

Consider special case of Hamming space $\{0,1\}^d$:

given $p, q \in \{0,1\}^d$, $d(p,q)$ = # bits that are different

Consider approx decis. problem for fixed r, with approx factor c

Preproc. Algm:

Fix k to be set later

define hash function $h : \{0,1\}^d \to \{0,1\}^k$ by projecting to k random dims

build hash table

$\Rightarrow$ space $O(n)$

Query(q):

for each $p \in P$ with $h(p) = h(q)$

if $d(p,q) \le cr$ return yes

return no

Facts

0. $\Pr[h(p) = h(q)] = \left(1 - \frac{d(p,q)}{d}\right)^k$

#= $d(p,q)$

p = 10 1 1 10 0 1 1

q = 10 0 1 1 1 0 0 1

all k chosen dims are not among circled ones

1. If $d(p,q) > cr$, then

$\Pr[h(p) = h(q)] \le \left(1 - \frac{cr}{d}\right)^k$

$\lesssim e^{-\frac{cr}{d}k}$ by picking $k = \frac{d \ln n}{cr}$

$= \frac{1}{n}$.

2. If $d(p,q) \le r$, then

$$\Pr[h(p) = h(q)] \ge \left(1 - \frac{r}{d}\right)^k$$

$$\approx e^{-\frac{rk}{d}}$$

$$= \frac{1}{n^{1/c}}$$

expected query time

$$= \tilde{O}\left(E[\#\text{ pts } p \in P \text{ with } h_i(p) = h(q) \text{ \& } d(p,q) > cr] \right)$$

$$\le \tilde{O}\left(n \cdot \frac{1}{n} \right) = \tilde{O}(1)$$

Error Analysis:

Case 1. $d(p,q) > cr$ for all $p \in P$

alg'm is always correct

Case 2. $d(p,q) \le r$ for some $p \in P$

$\Pr[\text{alg'm is correct}]$

$\ge \Pr[h(p) = h(q)] = \frac{1}{n^{1/c}}$ tiny!

∴ final idea - repeat $t = 100 n^{1/c}$ times to lower error prob

$$\Rightarrow \text{err prob} \le \left(1 - \frac{1}{n^{1/c}}\right)^t \le e^{-100}$$

$$\Rightarrow \text{space } \boxed{\tilde{O}(n^{1+1/c})}$$

$$\text{query time } \boxed{\tilde{O}(n^{1/c})}$$

for approx factor c

(e.g. $c = 2$: space $\tilde{O}(n^{3/2})$, query $\tilde{O}(\sqrt{n})$)

Rmks - immediately extends to L_1 metric in $[U]^d$

$$d(p,q) = \sum_{i=1}^{d} |p_i - q_i|$$

by first mapping $[U]^d \to \{0,1\}^{dU}$

$p = (2,3) \to (11000, 11100)$

$q = (5,2) \to (11111, 11000)$

L_1-dist $3+1=4$

3/18/16

- extends to Euclidean metric by modifying hash fn (Datar-Indyk-Immorlica-Mirrokni '04)

take rand. projection $f: \mathbb{R}^d \to \mathbb{R}^k$

$$f(a_1, \ldots, a_d) = \left(\sum_{j=1}^{d} X_{1j} a_j, \ldots, \sum_{j=1}^{d} X_{kj} a_j \right)$$

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

take randomly shifted grid in $\mathbb{R}^k$ of suff. large side length m

define $h(a)$ = grid cell containing $f(a)$

Then $\Pr[f(p), f(q) \text{ in diff grid cell in } i^{th} \text{ coord}] \approx \frac{E[|Y_i|]}{m} = \frac{\sqrt{2/\pi}\,\|p-q\|}{m}$

where $Y_i = i^{th}$ coord of $f(p) - f(q)$

$$= \sum_{j=1}^{d} X_{ij}(p_j - q_j)$$

$$\sim N(0, \|p-q\|^2)$$

$$\Rightarrow \Pr[h(p) = h(q)] \approx \left(1 - \sqrt{2/\pi}\,\frac{d(p,q)}{m}\right)^k$$

similar to Fact 0

rest is as before

$\Rightarrow$ space $O^*(n^{1+1/c})$

query time $O^*(n^{1/c})$ dyn $O^*(n^{1/c})$

Andoni-Indyk '06 improve to space $O^*(n^{1+1/c^2})$
time $O^*(n^{1/c^2})$

Andoni-Razenshteyn '15 (data-dependent LSH) space $O^*(n^{1+1/(2c^2-1)})$
time $O^*(n^{1/(2c^2-1)})$
(Hamming $n^{1/(2c-1)}$)

Pagh '16 / Wei '19: Las Vegas

Andoni et al. '17: space/time trade-offs

A12