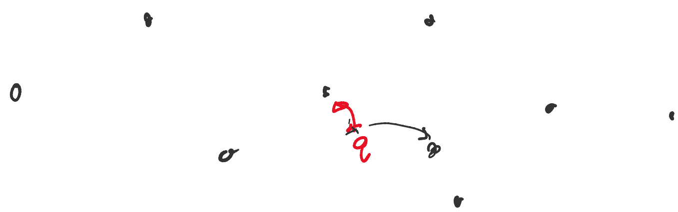
April 7, 2021 9:10 AM

Approx Nearest Neighbor Search

build data structure for set S of n pts in high-dimensional space $\mathcal{S}$ (d = dim)

s.t. given query pt q,

can find a pt $p \in S$ "close" to q



exact nearest neighbor: find $p^* \in S$ minimizing $\delta(p^*, q)$

e.g. Euclidean distance

"curse of dimensionality"

(query time $\rightarrow$ $O(dn)$)

approx nearest neighbor: find $p \in S$ s.t.

$$\delta(p, q) \leq c \min_{p^* \in P} \delta(p^*, q)$$

approx factor $c > 1$.

approx fixed radius search: given fixed r,

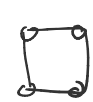
find $p \in S$ with $\delta(p, q) \leq cr$

or conclude that $\min_{p^* \in P} \delta(p^*, q) > r$.

(ANN reduces to fixed radius search by binary search)

Hamming Space Case

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



pts = binary strings of length d

pts = binary strings of length d

for $p = p_1 \cdots p_d$ and $q = q_1 \cdots q_d$,

$$\delta(p,q) = |\{ i : p_i \neq q_i \}|$$

(Hamming distance)

e.g. 10110
11100

Locality-Sensitive Hashing (LSH) (Indyk-Motwani '98)

approach. design family of hash function $h: S \to T$ s.t.

if $\delta(p,q) \le r$, $\Pr_h[h(p) = h(q)]$ is large

if $\delta(p,q) > cr$, $\Pr_h[h(p) = h(q)]$ is small.

how?

by random projection!

Pick rand sample $I = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, d\}$ where each index is chosen w. prob α indep'ly ($E[k] = \alpha d$).

Define $h(p_1 \cdots p_d) = p_{i_1} \cdots p_{i_k}$.

Obs For fixed p, q,

$$\Pr_h[h(p) = h(q)] = (1-\alpha)^{\delta(p,q)}$$

e.g. $p = 101110011$

$q = 100111001$

$h(p) = 010$

$h(q) = 010$

$\delta(p,q) = 3$

$d = 9$.

Cor

1. if $\delta(p,q) > cr$, then

$$\Pr_h[h(p) = h(q)] \le (1-\alpha)^{cr}$$
$$\le e^{-\alpha cr}$$
$$= \frac{1}{n} : \text{ small}$$

Pick $\alpha = \frac{\ln n}{cr}$

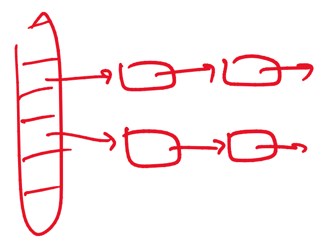
2. if $\delta(p,q) < r$, then

$$\Pr_h(h(p) = h(q)) \ge (1-\alpha)^{r}$$
$$\gtrsim e^{-\alpha r}$$
$$= e^{-\frac{\ln n}{c}} = \frac{1}{n^{1/c}}.$$

not as small

insert(p):

add p to the linked list $A[h(p)]$.



delete(p): similar

Search(q): for each p in the list $A[h(q)]$

if $\delta(p,q) \le cr$, stop & return p

return no

Analysis: Fix q. (assume oblivious adversary)

expected query time

$$= O\left(d \cdot E\left[\,\left|\{p \in S : h(p) = h(q) \;\&\; \delta(p,q) > cr\}\right|\,\right]\right)$$

$$= O\left(d \sum_{\substack{p \in S: \\ \delta(p,q) > cr}} \Pr(h(p) = h(q))\right)$$

$\underset{\substack{p \in S: \\ \delta(p,q) > cr}}{}$

$$\leq O\left(d \cdot n \cdot \frac{1}{n}\right)$$

$$\leq O(d)$$ excellent!

Err prob?

let $p^* \in S$ minimize $\delta(p^*, q)$.

Case 1. $\delta(p^*, q) > cr$.

always return no. Correct.

(if $\delta(p^*, q) \in (r, cr)$, allowed to go either way)

Case 2. $\delta(p^*, q) \leq r$.

$$\Pr[\text{correct}] \geq \Pr[h(p^*) = h(q)]$$

$$\geq \frac{1}{n^{1/c}}.$$ tiny!

Final idea - repeat $t = n^{1/c} \ln n$ times

(i.e. use t hash fns & t hash tables)

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

$$\leq e^{-\frac{1}{n^{1/c}} \cdot t} \leq \frac{1}{n}.$$

$\Rightarrow$ query time $\boxed{\tilde{O}(dn^{1/c})}$

insert time $\boxed{\tilde{O}(n^{1/c})}$ for approx factor c

delete "

insert time $\boxed{\tilde{O}(n^{1/c})}$
delete "
Space $\boxed{\tilde{O}(n^{1+1/c})}$

for approx factor c

(e.g. $c=2$: $\tilde{O}(\sqrt{n})$)

Rmks - improved to $\tilde{O}(n^{\frac{1}{2c-1}})$ (data-dependent LSH)

- derand??
- Monte Carlo $\to$ Las Vegas

Other spaces?

e.g. L_1 metric space

$\mathcal{S} = \{0, \ldots, U-1\}^d$

for $p = (p_1, \ldots, p_d)$, $q = (q_1, \ldots, q_d)$

$\delta_1(p,q) = |p_1 - q_1| + \ldots + |p_d - q_d|$

(Manhattan dist.)

idea - embedding

can map L_1 into Hamming space ...