Nearest Neighbor Search and Other Problems in High Dimensional Computational Geometry

Piotr Indyk

Stanford University

What is this course about ?

Solving Computational Geometry problems in high dimensional spaces

- recent progress
- future work

Nearest Neighbor Search (NNS)

Given: metric space $D = (X, d)$, n points $P \subset X$.

Goal: a data structure which for any query $q \in X$ returns the point(s) in P closest to q.

Metric space:

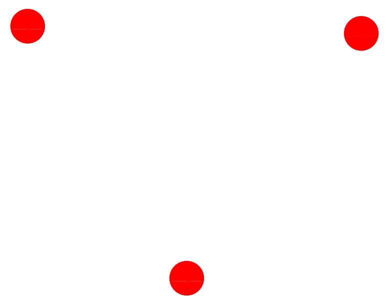
- reflexive: $d(x, y) = 0$ iff $x = y$
- symmetric: $d(x, y) = d(y, x)$
- triangle ineq.: $d(x, y) \leq d(x, z) + d(z, y)$

High dimensional metrics = "many degrees of freedom" (e.g. R^d with l_2, l_1 or l_∞ norm)

Low dimensional case

- $d = 1$ - Binary Search

- $d = 2$ - Post Office Problem



2D algorithms

Query	Storage	Paper
$\log n$	n^2	Dobkin, Lipton'76
$\log^2 n$	n	Shamos'75
$\log n$	n	Lipton, Tarjan'77

High dimensions

Query*	Storage*	Paper
$2^d \log n$	$n^{2^{d+1}}$	Dobkin,Lipton'76
$d^d \log n$	$n^{\lceil d/2 \rceil (1+\delta)}$	Clarkson'88
$d^5 \log n$	$n^{d+\delta}$	Meiser'93 improvements in Agarwal, Erickson'98
$\log^{d-1} n$	$n \log^{d-1} n$	range searching techniques (l_1, l_∞)

Exponential dependence on d !

* bounds sometimes "simplified" for clarity

Approximate NN

Query	Storage	Paper
$d^2 \log n$ $(\epsilon \geq d)$	$d^2 n \log n$	Bern '93, Chan'97
$\epsilon^{-(d-1)/2} \log n$	$\epsilon^{-(d-1)/2} n \log n$	Arya, Mount'93 Clarkson'94, Chan'97
$d^d \epsilon^{-d}$	dn	Arya et al'94
n	dn	Cohen, Lewis '97 (positive dot product)
$d^2 \log^2 d$ n	n^{2d} $dn \log^2 n \log^2 d$	Kleinberg '97

Either linear query time, or exponential storage

High dimensional applications

Pattern recognition(60's):

- Nearest Neighbour Rule:
  - given: set of classified objects S
  - goal: classify unknown object x
  - NN method: return the class of $y \in S$ closest to x

- the error at most twice the optimal [Cover, Hart'67]

- large storage/query time [Minsky, Papert'69]

Applications ctd.

Retrieval:

- text
  - vector space model [Salton, Wong, Yang'74]
  - Latent Semantic Indexing [Berry, Dumais, O'Brien al'88]
- images [QBIC, Photobook, Virage]
  - histograms
  - texture parameters
  - shape information
- audio, video ...
- genetic sequences
- drug molecules

Applications ctd.

- Compression/Vector Quantization
- Data Mining/Computational Statistics
- Recycling computation
- ...

Corollaries

1. High dimensional NNS is crucial to a large body of applications

2. There are many other interesting high dimensional problems

    - proximity problems
    - clustering
    - regression
    - ...

Proximity problems

- close(st) pair
- furthest pair/diameter
- minimum enclosing ball
- minimum spanning tree

Static/dynamic versions

Clustering

- clustering:
  - k-center: minimize

$$\max_p d(p, C(p))$$

where $C(p)$ is the center of p's cluster

  - k-median[2]: max replaced by sum of squares
  - k-clustering: minimize

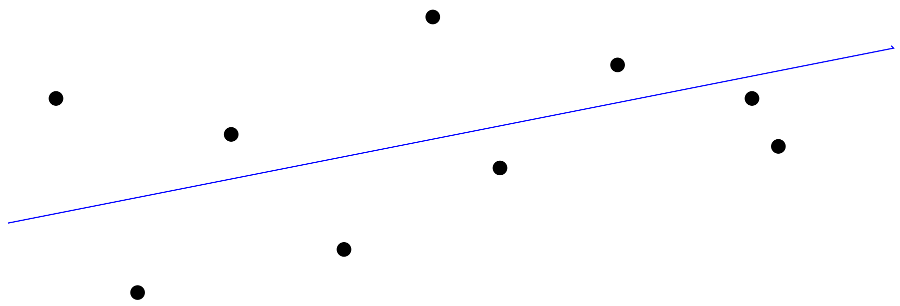
$$\sum_i \sum_{p,q \in C_i} d(p, q)$$

where C_i is the ith cluster

  - ...

Regression

Find a (hyper)plane fitting a given set of points



Proximity problems

Problem	Time	Paper
closest pair	exp$(d)n \log n$	Bentley,Shamos'76, Rabin
diameter (approximate)	$n^{2-2^{-(d+1)}} \log n$ $(1/\epsilon)^{2d} + n$	Yao'82 folklore
smallest ball (approximate)	exp$(\sqrt{d})n$ poly$(d)n$	Gartner'92 linear programming
MST (approximate)	$n \log^d n$ exp$(d)n \log n$	range queries Vaidya'84

Clustering problems

Problem	Time	Factor	Paper
k-center	$n \log k + p/\epsilon^{dk}$	$1+\epsilon$	Agarwal, Procopiuc'98
	$\exp(d) n \log k$	2	Feder, Greene'90
metric k-center	nk	2	Hochbaum,Shmoys'85
k-median	$\exp(1/\epsilon^d) n \log^2 n$	$1+\epsilon$	Kolliopoulos, Rao'99
metric k-median	n^2	6	Jain, Vazirani'99
	n^3	4	Charikar,Guha'99
k-median^2	$\epsilon^{-k^2 d} n \log^k n$	$1+\epsilon$	Matousek'99
metric k-median^2	n^2	25	like k-median
k-clustering^2	$n^{dk^2} k^{k^d}$	1	Inaba et al; Schulman
metric k-clustering	n^{k+1}	2	Guttman-Beck, Hassin'99

Regression problems

Measure	Time	Factor	Paper
max error	$f(1/\epsilon)^d$ n	$1+\epsilon$	Duncan et al'99
sum of squares	$d^2 n$	1	SVD

For more see `www.ics.uci.edu/~eppstein/280`

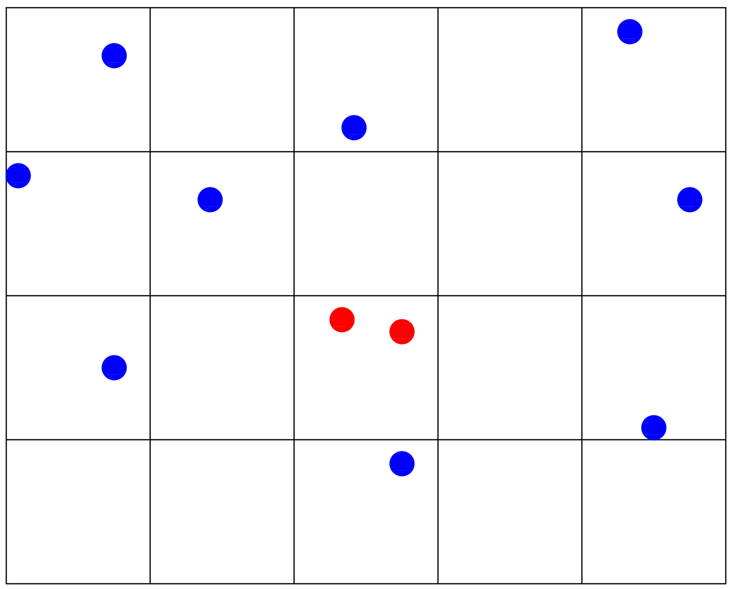
Why exponential dependence ?

An example: Closest Pair (CP):

- promise problem: assume that CP distance $t \in [\Delta, c\Delta]$
- can be used to solve CP with $\log\log R$ overhead, where R =diameter/closest pair

CP: algorithms

Randomized, $d^{O(d)} n$ time.



1. Impose a randomly translated grid of side $2c\Delta$
2. Hope both CP points fall to the same cell
3. Check all pairs in all cells

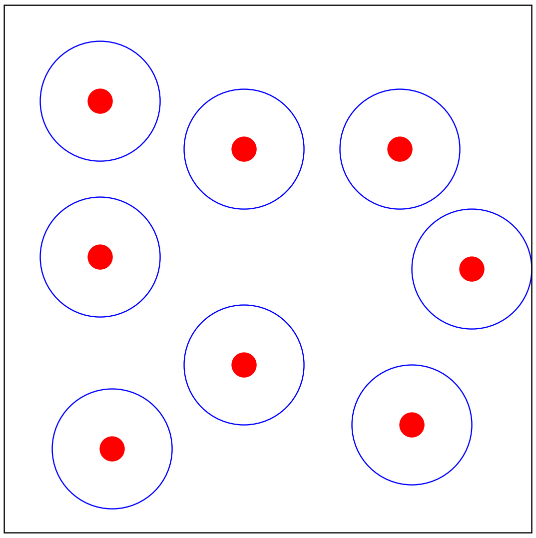
CP: correctness

- grid randomly translated in each dimension
- for each dimension we have $> \frac{1}{2}$ probability of success
- therefore $> \frac{1}{2^d}$ overall success probability

Repeat 2^d times for "hope amplification".

CP: complexity

"How many devils can sit on a pinhead ?"



Finite number ($d^{O(d)}$)

Overall complexity: $2^d \cdot n \cdot d^{O(d)}$