Geometric Problems in High Dimensions: Sketching

Piotr Indyk

Dimensionality Reduction in Hamming Metric

Theorem: For any r and eps>0 (small enough), there is a distribution of mappings G: $\{0,1\}^d \rightarrow \{0,1\}^t$, such that for any two points p, q the probability that:

- If D(p,q)< r then D(G(p), G(q)) < (c+eps/10)t
- If D(p,q)>(1+eps)r then D(G(p), G(q)) >(c+eps/20)t

is at least 1-P, as long as $t=C*\log(2/P)/eps^2$, C large constant.

- Given n points, we can reduce the dimension to O(log n), and still approximately preserve the distances between them
- The mapping works (with high probability) even if you don't know the points in advance

Proof

- Mapping: $G(p) = (g_1(p), g_2(p),\ldots,g_t(p))$, where

$$g_j(p)=f_j(p_{|I_j})$$

  - I: a multiset of s indices taken independently uniformly at random from {1…d}
  - $p_{|I}$: projection of p
  - f: a random function into {0,1}

- Example: p=01101, s=3, I={2,2,4} $\rightarrow p_{|I} = 110$

Analysis

- What is $Pr[p_{|I} = q_{|I}]$?
- It is equal to $(1-D(p,q)/d)^s$
- We set $s=d/r$. Then $Pr[p_{|I} = q_{|I}] = e^{-D(p,q)/r}$, which looks more or less like this:
- Thus
  - If $D(p,q) < r$ then $Pr[p_{|I} = q_{|I}] > 1/e$
  - If $D(p,q) > (1+eps)r$ then $Pr[p_{|I} = q_{|I}] < 1/e - eps/3$

Analysis II

- What is Pr[g(p) <> g(q)] ?
- It is equal to $Pr[p_{|I} = q_{|I}]*0 + (1- Pr[p_{|I} = q_{|I}]) *1/2 = (1- Pr[p_{|I} = q_{|I}])/2$
- Thus
  - If D(p,q)< r then Pr[g(p) <> g(q)] < (1-1/e)/2 = c
  - If D(p,q)>(1+eps)r then Pr[g(p) <> g(q)] > c+eps/6

Analysis III

- What is D(G(p),G(q)) ? Since $G(p)=(g_1(p), g_2(p),\ldots,g_t(p))$, we have:

$$D(G(p),G(q))=\Sigma_j [g_j(p)<> g_j(q)]$$

- By linearity of expectations

$$E[D(G(p),G(q))]= \Sigma_j \Pr[g_j(p) <> g_j(q)] = t \Pr[g_j(p) <> g_j(q)]$$

- To get the high probability bound, use Chernoff inequality

Chernoff bound

- Let X_1, $X_2 \ldots X_t$ be independent random 0-1 variables, such that $Pr[X_i=1]=r$. Let $X= \Sigma_j X_j$. Then for any $0<b<1$:

$$Pr[\ |X - t\,r| > b\,t\,r] < 2e^{-b^2 tr/3}$$

- Proof I: Cormen, Leiserson, Rivest, Stein, Appendix C
- Proof II: attend one of David Karger's classes.
- Proof III: do it yourself.

Analysis IV

- In our case $X_j=[g_j(p)<> g_j(q)]$, $X=D(G(p),G(q))$. Therefore:
  - For r=c:

$$\Pr[X>(c+eps/20)t] < \Pr[|X-tc|>eps/20\ tc] <2e^{-(eps/20)^2tc/3}$$

  - For r=c+eps/6:

$$\Pr[X<(c+eps/10)t]<\Pr[|X-(c+eps/6)t|>eps/20\ tc]<2e^{-(eps/20)^2t(c+eps/6)/3}$$

- In both cases, the probability of failure is at most $2e^{-(eps/20)^2tc/3}$

Finally…

$$2e^{-(eps/20)^2 tc/3} = 2e^{-(eps/20)^2\ c/3\ C^*\ \log(2/P)/eps^2} = 2e^{-\log(2/P)c^*C/1200}$$

- Take C so that c*C/1200 = 1. We get

$$2e^{-\log(2/P)c^*C/1200} = 2e^{-\log(2/P)} = P$$

- Thus, the probability of failure is at most P.

Algorithmic Implications

- Approximate Near Neighbor:
  - Given: A set of n points in $\{0,1\}^{d,}$ eps>0, r>0
  - Goal: A data structure that for any query q:
    * if there is a point p within distance r from q, then report p' within distance (1+eps)r from q
- Can solve Approximate Near*est* Neighbor by taking r=1,(1+eps),…

Algorithm I - Practical

- Set probability of error to 1/poly(n) → $t=O(\log n/\mathrm{eps}^2)$
- Map all points p to G(p)
- To answer a query q:
  - Compute G(q)
  - Find the nearest neighbor G(p) of G(q)
  - If D(p,q) < r(1+eps), report p
- Query time: $O(n \log n/\mathrm{eps}^2)$

Algorithm II - Theoretical

- The exact nearest neighbor problem in $\{0,1\}^t$ can be solved with
  - 2^t space
  - O(t) query time

  (just store pre-computed answers to all queries)
- By applying mapping G(.), we solve approximate near neighbor with:
  - $n^{O(1/eps^2)}$ space
  - $O(d \log n/eps^2)$ time

Another Sketching Method

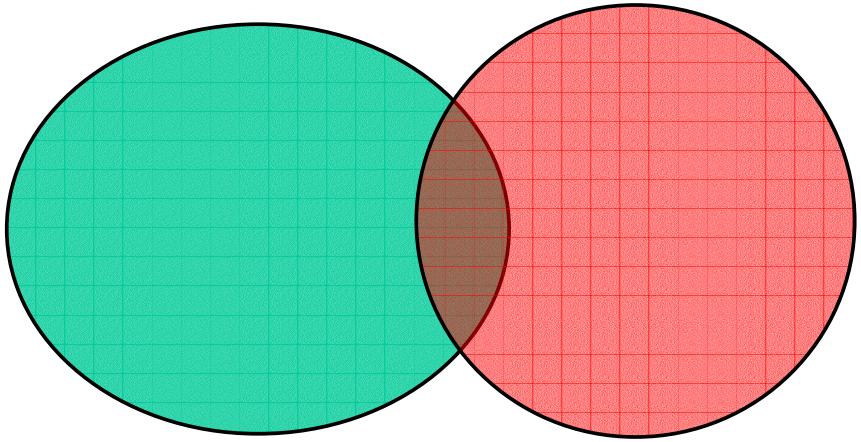
- In many applications, the points tend to be quite sparse
  - Large dimension
  - Very few 1's
- Easier to think about them as sets. E.g., consider a set of words in a document.
- The previous method would require very large s
- For two sets A,B, define Sim(A,B)=|A ∩ B|/|A U B|
  - If A=B, Sim(A,B)=1
  - If A,B disjoint, Sim(A,B)=0
- How to compute short sketches of sets that preserve Sim(.) ?

“Min Approach”

- Mapping: $g(A)=\min_{a \text{ in } A} h(a)$, where h is a random permutation of the elements in the universe
- Fact:

$$Pr[g(A)=g(B)]=Sim(A,B)$$

- Proof: Where is min(h(A) U h(B)) ?



Min Sketching

- Define $G(A)=(g_1(A), g_2(A),\ldots, g_t(A))$
- By Chernoff bound, we can conclude that if $t=C \log(1/P)/eps^2$, then for any A,B, the number of j's such that $g_j(A)= g_j(B)$ is equal to

$$t\,[Sim(A,B) +/- eps\,]$$

with probability at least 1-P