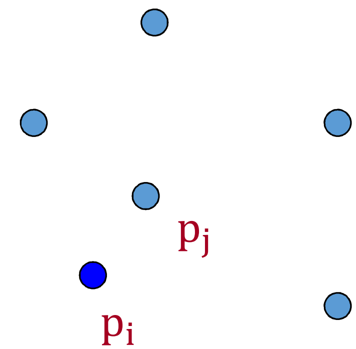
Lecture 27: Algorithmic Applications of Embeddings

David Woodruff

Finding Similar Items

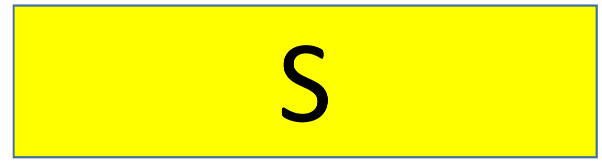
- In recommendation systems, want to find users that have similar buying patterns so can recommend items to users
- In data imputation, may be missing entries in a database and fill them in based on your “nearest neighbor”
- In a document collection, want to find similar documents to detect multiple versions of the same article, mirror websites, plagiarism, etc.
- Special case: Closest Pair Problem

Closest Pair Problem



- Given n points in R^{d}, find the pair p,q with minimum distance dist(p,q)
- dist(p,q) could be $\left(\sum_{j=1,\ldots,d}(p_j - q_j)^2\right)^{1/2}$
- Can solve in n^2d time, but not good if n and d are large
- Divide-and-conquer algorithms depend on 2^d, too slow if d > log n
  - Often referred to as the "Curse of Dimensionality"

Embedding Paradigm



- Choose a random s x d matrix S for a small value $s \ll d$

- Replace the n points $p_1, \dots, p_n \in R^d$ with n points $S \cdot p_1, \dots, S \cdot p_n \in R^s$

- Compute a function $f\left(S \cdot p_i, S \cdot p_j\right) \approx \text{dist}(p_i, p_j)$ between all pairs $S \cdot p_i$ and $S \cdot p_j$ and output the pair p_i and p_j for which $f\left(S \cdot p_i, S \cdot p_j\right)$ is minimal

- Time: $O(nd \cdot s + n^2 \cdot s)$ if f computable in O(s) time

- Example: if n = d and s = $\Theta(\log n)$, get $O(n^2 \log n)$ time instead of $O(n^2 d) = O(n^3)$

A Randomized Embedding

- Let s= $O(\frac{1}{\epsilon^2})$ for an accuracy parameter constant $\epsilon > 0$
- Choose a random s x d matrix S

$$\frac{1}{\sqrt{s}} \cdot \begin{matrix} 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 \end{matrix} \cdot p$$

- Each entry of S is $1/\sqrt{s}$ with pr. ½, and is $-1/\sqrt{s}$ with pr. ½
- For a point $p \in R^d$, the vector $S \cdot p \in R^s$ is much lower dimensional
- Claim: $E[|S \cdot p|_2^2] = |p|_2^2$

Expectation

- Claim: $E[|S \cdot p|_2^2] = |p|_2^2$

S_1

$\frac{1}{\sqrt{s}}$	1 -1 1 1 -1 1 -1 1 1 -1 1 1
	-1 -1 -1 1 -1 1 1 -1 1 -1 1

- Proof: Let S_i be the i-th row of S

- Since each row of S is identically distributed, $E[|S \cdot p|_2^2] = s \cdot E[< S_1, p >^2]$

- $E[< S_1, p >^2] = E[\left(\sum_{j=1,\ldots,d} \sigma_j p_j\right)^2] = \sum_{j_1, j_2} E[\sigma_{j_1} \sigma_{j_2}] \cdot p_{j_1} p_{j_2}$

- If $j_1 = j_2$, then $E[\sigma_{j_1} \sigma_{j_2}] = \frac{1}{s}$, otherwise $E[\sigma_{j_1} \sigma_{j_2}] = E[\sigma_{j_1}] \cdot E[\sigma_{j_2}] = 0$

- So $E[< S_1, p >^2] = \frac{|p|_2^2}{s}$

Variance

- Claim: $\mathrm{Var}[|S \cdot p|_2^2] = O(\frac{|p|_2^4}{s})$
- Proof: $\mathrm{Var}[|S \cdot p|_2^2] = E[|S \cdot p|_2^4] - E^2[|S \cdot p|_2^2]$
- $E[|S \cdot p|_2^4] = E[\left(\sum_{i=1,\ldots,s} < S_i, p >^2\right)^2] = \sum_{i,i'} E[< S_i, p >^2 < S_{i'}, p >^2]$

$$= \sum_i E[< S_i, p >^4] + \sum_{i \neq i'} E[< S_i, p >^2] \cdot E[< S_{i'}, p >^2]$$

- Hence,

$$\mathrm{Var}[|S \cdot p|_2^2] \leq s \cdot E[\left(\sum_{j=1,\ldots,d} \sigma_j p_j\right)^4] = s \cdot \sum E[\sigma_{j_1}\sigma_{j_2}\sigma_{j_3}\sigma_{j_4}] \cdot p_{j_1}p_{j_2}p_{j_3}p_{j_4}$$

Variance Continued

- $\mathrm{Var}[|S \cdot p|_2^2] \le s \cdot E[(\sum_{j=1,\ldots,d} \sigma_j p_j)^4] = s \cdot \sum E[\sigma_{j_1}\sigma_{j_2}\sigma_{j_3}\sigma_{j_4}] \cdot p_{j_1}p_{j_2}p_{j_3}p_{j_4}$
- If $E[\sigma_{j_1}\sigma_{j_2}\sigma_{j_3}\sigma_{j_4}] \neq 0$, the set $\{j_1, j_2, j_3, j_4\}$ has 4 equal indices, or 2 pairs of equal indices
- If $j_1 = j_2 = j_3 = j_4$, then $E[\sigma_{j_1}\sigma_{j_2}\sigma_{j_3}\sigma_{j_4}] = 1/s^2$
  - Contribution is $s \cdot \left(\frac{1}{s^2}\right) \cdot \sum_j p_j^4 \le s \cdot \left(\frac{1}{s^2}\right) \cdot \left(\sum_j p_j^2\right)^2 = \left(\frac{1}{s}\right) \cdot |p|_2^4$
- If say, $j_1 = j_2$ and $j_3 = j_4$, then $E[\sigma_{j_1}\sigma_{j_2}\sigma_{j_3}\sigma_{j_4}] = 1/s^2$
  - Contribution is $s \cdot \left(\frac{1}{s^2}\right) \cdot \left(\sum_j p_j^2\right)^2 = \left(\frac{1}{s}\right) \cdot |p|_2^4$
- Thus, $\mathrm{Var}[|S \cdot p|_2^2] \le \frac{4}{s} |p|_2^4$

Recap

- $\mathrm{E}[|S \cdot p|_2^2] = |p|_2^2$
- $\mathrm{Var}[|S \cdot p|_2^2] \leq \frac{4}{s} |p|_2^4$
- Chebyshev's inequality: for a random variable X,

$$\Pr[|X - E[X]| \geq \lambda(\mathrm{Var}[X])^{\frac{1}{2}}] \leq 1/\lambda\text{^}2$$

- Proof: $\Pr\left[|X - E[X]| \geq \lambda(\mathrm{Var}[X])^{\frac{1}{2}}\right]$

$$= \Pr\left[(X - E[X])^2 \geq \lambda^2 \mathrm{Var}[X]\right] \leq 1/\lambda^2$$

Applying Chebyshev's Bound

- $E[|S \cdot p|_2^2] = |p|_2^2$
- $Var[|S \cdot p|_2^2] \le \frac{4}{s} |p|_2^4$
- Chebyshev's inequality: for a random variable X,
  $Pr[|X - E[X]| \ge \lambda (Var[X])^{\frac{1}{2}}] \le 1/\lambda\text{^}2$
- $Pr\left[\left||S \cdot p|_2^2 - |p|_2^2\right| \ge \frac{20}{s^{1/2}} |p|_2^2\right] \le \frac{1}{100}$. Set $s = 400/\epsilon^2$
- $Pr[||S \cdot p|_2^2 - |p|_2^2| \ge \epsilon \cdot |p|_2^2] \le \frac{1}{100}$

Recap

- Chose a random s x d matrix S for s= $O(\frac{1}{\epsilon^2})$
- For an individual point p, $\Pr\left[\left||S \cdot p|_2^2 - |p|_2^2\right| \geq \epsilon \cdot |p|_2^2\right] \leq \frac{1}{100}$
- Since S is linear, we can compute $S \cdot (p_i - p_j)$. Setting $p = p_i - p_j$ above,

$$\Pr\left[\left|\left|S \cdot (p_i - p_j)\right|_2^2 - \mathrm{dist}(p_i, p_j)^2\right| \geq \epsilon \cdot \mathrm{dist}(p_i, p_j)^2\right] \leq \frac{1}{100}$$

- But we have $\frac{n(n-1)}{2}$ distinct pairs of points, can't union bound over all of them!

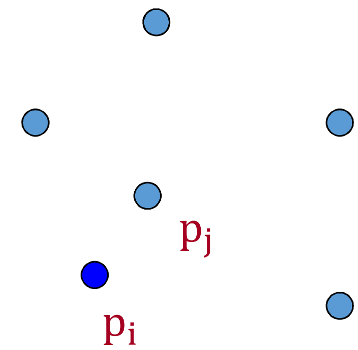
Amplifying the Probability

- Let $r = O(\log n)$
- Choose r independent s x d matrices $S^1, \ldots, S^r$
- $f\left((S^1 p_i, S^2 p_i, \ldots, S^r p_i), (S^1 p_j, S^2 p_j, \ldots, S^r p_j)\right) = \text{median}_{k=1,\ldots,r} \left|S^k(p_i - p_j)\right|_2^2$
- Since $\left|S^k(p_i - p_j)\right|_2^2 \in (1 \pm \epsilon)\text{dist}(p_i, p_j)^2$ with probability 99/100,

  $f(p_i, p_j) \in (1 \pm \epsilon) \cdot \text{dist}(p_i, p_j)^2$ with probability $1 - 1/n^3$
- By a union bound, with probability at least 1-1/n, simultaneously for all i,j:

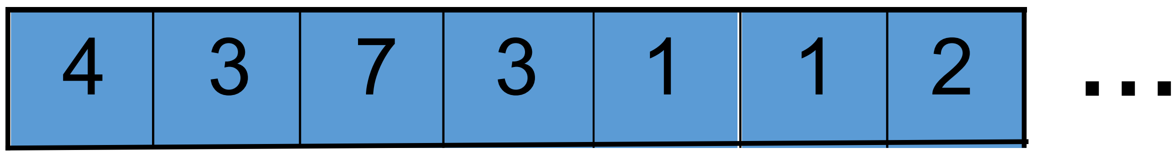
$$f(p_i, p_j) \in (1 \pm \epsilon) \cdot \text{dist}(p_i, p_j)^2$$

Recap



- We are given n points $p_1, \ldots, p_n \in R^d$
- Choose r = O(log n) independent s x d matrices $S^1, \ldots, S^r$, for s= $O(\frac{1}{\epsilon^2})$
- Compute $S^1 \cdot p_i, \ldots, S^r \cdot p_i$ for each i. Total time is O(n d log n /ϵ^2)
- Compute $f(p_i, p_j) = \text{median}_{k=1,\ldots,r} |S^k(p_i - p_j)|_2^{\ 2}$ for each pair i,j, and output the minimum-valued pair. Total time is $O(n^2 \log n / \epsilon^2)$
- Overall time is O(n d log n /ϵ^2 + $n^2 \log n / \epsilon^2$)

Application to Data Streams



- We are given a stream of items $i_1, i_2, \ldots, i_n$ from a universe U of size u
- Let f be the frequency vector of length u, so f_i is the number of occurrences of item i
- Want to approximate $|f|_2^2 = \sum_i f_i^2$, which is an indication of the "skew" of the stream

How much memory does a streaming algorithm need?

Streaming from Embeddings

- Choose a random s x u matrix S for s= $O(\frac{1}{\epsilon^2})$
- Initialize $S \cdot f = 0^s$
- Given an occurrence of item i, $S \cdot f \leftarrow S \cdot f + S \cdot e_i$, where e_i is i-th standard basis vector
- At end of the stream, output $|S \cdot f|_2^2 \in (1 \pm \epsilon)|f|_2^2$ with probability > 99/100
- Can maintain $S \cdot f$ with s= $O(\frac{1}{\epsilon^2})$ words of memory
- S can be chosen from a 4-universal hash family, so O(1) words to store S