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 $\mathbb{R}^d$, find the pair $p, q$ with minimum distance $\text{dist}(p, q)$

- $\text{dist}(p, q)$ could be $\left(\sum_{j=1}^{d} (p_j - q_j)^2\right)^{1/2}$

- Can solve in $n^2 d$ 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 \( s \times d \) matrix \( S \) for a small value \( s \ll d \)

- Replace the \( n \) points \( p_1, \ldots, p_n \in \mathbb{R}^d \) with \( n \) points \( S \cdot p_1, \ldots, S \cdot p_n \in \mathbb{R}^s \)

- Compute a function \( f(S \cdot p_i, S \cdot p_j) \approx \text{dist}(p_i, p_j) \) between all pairs \( S \cdot p_i \) and \( S \cdot p_j \) and output the pair \( p_i \) and \( p_j \) for which \( f(S \cdot p_i, S \cdot p_j) \) 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 \( O(n^2d) = O(n^3) \)
A Randomized Embedding

- Let $s = O\left(\frac{1}{\epsilon^2}\right)$ for an accuracy parameter constant $\epsilon > 0$

- Choose a random $s \times d$ matrix $S$

- Each entry of $S$ is $1/\sqrt{s}$ with pr. $\frac{1}{2}$, and is $-1/\sqrt{s}$ with pr. $\frac{1}{2}$

- For a point $p \in \mathbb{R}^d$, the vector $S \cdot p \in \mathbb{R}^s$ is much lower dimensional

- Claim: $\mathbb{E}[|S \cdot p|^2] = |p|^2$
Expectation

• Claim: $E[|S \cdot p|^2] = |p|^2$

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

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

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

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

• Claim: \( \text{Var}[|S \cdot p|_2^2] = O\left(\frac{|p|_2^4}{s}\right) \)

• Proof: \( \text{Var}[|S \cdot p|_2^2] = E[|S \cdot p|_2^4] - \text{E}^2[|S \cdot p|_2^2] \)

\[ E[|S \cdot p|_2^4] = E\left[\left(\sum_{i=1,\ldots,s} < S_i, p >^2\right)^2\right] = \sum_{i,i'} E[< S_i, p >^2 < S_{i'}, p >^2] \]
\[ = \sum_i E[< S_i, p >^4] + \sum_{i \neq i'} E[< S_i, p >^2] \cdot \text{E}[< S_{i'}, p >^2] \]

• Hence,

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

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

- 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 $s \cdot \left(\frac{1}{s^2}\right) \cdot \sum_j p_j^4 \leq s \cdot \left(\frac{1}{s^2}\right) \cdot (\sum_j p_j^2)^2 = \left(\frac{1}{s}\right) \cdot |p|^2$

- 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 $s \cdot \left(\frac{1}{s^2}\right) \cdot (\sum_j p_j^2)^2 = \left(\frac{1}{s}\right) \cdot |p|^4$

- Thus, $\text{Var}[|S \cdot p|_2^2] \leq \frac{4}{s} |p|^4$
Recap

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

- $\text{Var}[|S \cdot p|^2] \leq \frac{4}{s} |p|^4$

- **Chebyshev’s inequality**: for a random variable $X$,
  \[ \Pr[|X - E[X]| \geq \lambda (\text{Var}[X])^{\frac{1}{2}}] \leq \frac{1}{\lambda^2} \]

- **Proof**: \[
  \Pr\left[|X - E[X]| \geq \lambda (\text{Var}[X])^{\frac{1}{2}}\right] \\
  = \Pr\left[(X - E[X])^2 \geq \lambda^2 \text{Var}[X]\right] \leq \frac{1}{\lambda^2} \]
Applying Chebyshev’s Bound

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

• Chose a random \( s \times d \) matrix \( S \) for \( s = O\left(\frac{1}{\epsilon^2}\right) \)

• For an individual point \( p \), \( \Pr[\left|\|S \cdot p\|^2_2 - |p|^2_2\right| \geq \epsilon \cdot |p|^2_2] \leq \frac{1}{100} \)

• Since \( S \) is linear, we can compute \( S \cdot (p_i - p_j) \). Setting \( p = p_i - p_j \) above,

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

• 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 $s \times d$ matrices $S^1, \ldots, S^r$

• $f((S^1 p_i, S^2 p_i, \ldots, S^r p_i), (S^1 p_j, S^2 p_j, \ldots, S^r p_j)) = \text{median}_{k=1,\ldots,r} |S^k(p_i - p_j)|^2$

• Since $|S^k(p_i - p_j)|^2 \in (1 \pm \varepsilon)\text{dist}(p_i, p_j)^2$ with probability 99/100,

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

  with probability $1 - 1/n^3$

• By a union bound, with probability at least $1 - 1/n$, simultaneously for all $i, j$:

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

• We are given \( n \) points \( p_1, \ldots, p_n \in \mathbb{R}^d \)

• Choose \( r = O(\log n) \) independent \( s \times d \) matrices \( S^1, \ldots, S^r \), for \( s = O\left(\frac{1}{\epsilon^2}\right) \)

• Compute \( S^1 \cdot p_i, \ldots, S^r \cdot p_i \) for each \( i \). Total time is \( O(n d \log n / \epsilon^2) \)

• 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 \( O(n d \log n / \epsilon^2 + n^2 \log n / \epsilon^2) \)
Application to Data Streams

| 4 | 3 | 7 | 3 | 1 | 1 | 2 |

- We are given a stream of items $i_1, i_2, ..., i_n$ 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 $|f|^2 = \sum_i f_i^2$, which is an indication of the “skew” of the stream

*How much memory does a streaming algorithm need?*
Streaming from Embeddings

• Choose a random $s \times u$ matrix $S$ for $s = O\left(\frac{1}{\varepsilon^2}\right)$

• 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 $|S \cdot f|_2^2 \in (1 \pm \varepsilon)|f|_2^2$ with probability $> 99/100$

• Can maintain $S \cdot f$ with $s = O\left(\frac{1}{\varepsilon^2}\right)$ words of memory

• $S$ can be chosen from a 4-universal hash family, so $O(1)$ words to store $S$