COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2021.

Lecture 13

LOGISTICS

- Problem Set 2 is due tomorrow, 11:59pm.
- The exam will be held next Tuesday in class.
- I am holding additional office hours for midterm prep, **tomorrow from 3-5pm** and **Monday, 4-6pm**.

Last Class:

- Finish Up proof of the JL lemma.
- Example application to clustering.
- Discuss connections to high dimensional geometry.

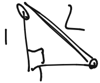
This Class:

- Finish up connection between JL Lemma and high dimensional geometry.
- Midterm review.
- Will do the 'fun' parts of high dimensional geometry after the midterm.

CURSE OF DIMENSIONALITY

Many-Near Orthogonal Vectors: In d-dimensional space, a set of $2^{\Theta(\epsilon^2 d)}$ random unit vectors have all pairwise dot products at most ϵ (think $\epsilon = .01$)

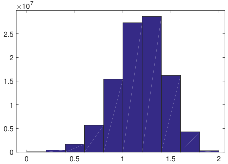
$$\|\vec{x}_i - \vec{x}_j\|_2^2 = \|\vec{x}_i\|_2^2 + \|\vec{x}_j\|_2^2 - 2\vec{x}_i^T\vec{x}_j \in [1.98, 2.02].$$



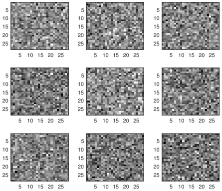
Even with an exponential number of random vector samples, we don't see any nearby vectors.

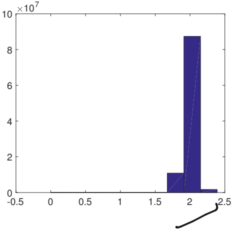
- One version of the 'curse of dimensionality'.
- If all your distances are roughly the same, distance based methods (k-means clustering, nearest neighbors, SVMs, etc.) aren't going to work well.
- Distances are only meaningful if we have lots of structure and our data isn't just independent random vectors.

Distances for MNIST Digits:



Distances for Random Images:





Recall: The Johnson Lindenstrauss lemma states that if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is a random matrix (linear map) with $m = O\left(\frac{\log n}{\epsilon^2}\right)$, for $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ with high probability, for all i, j:

$$(1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2 \leq \|\mathbf{\Pi}\vec{x}_i - \mathbf{\Pi}\vec{x}_j\|_2^2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2.$$

CONNECTION TO DIMENSIONALITY REDUCTION

Recall: The Johnson Lindenstrauss lemma states that if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is a random matrix (linear map) with $m = O\left(\frac{\log n}{\epsilon^2}\right)$, for $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ with high probability, for all i, j:

$$(1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2 \leq \|\mathbf{\Pi}\vec{x}_i - \mathbf{\Pi}\vec{x}_j\|_2^2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2.$$

Implies: If $\vec{x}_1, \ldots, \vec{x}_n$ are nearly orthogonal unit vectors in d-dimensions (with pairwise dot products bounded by $\epsilon/8$), then $\frac{\mathbf{\Pi}\vec{x}_1}{\|\mathbf{\Pi}\vec{x}_1\|_2}, \ldots, \frac{\mathbf{\Pi}\vec{x}_n}{\|\mathbf{\Pi}\vec{x}_n\|_2}$ are nearly orthogonal unit vectors in m-dimensions (with pairwise dot products bounded by ϵ).

Recall: The Johnson Lindenstrauss lemma states that if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is a random matrix (linear map) with $m = O\left(\frac{\log n}{\epsilon^2}\right)$, for $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ with high probability, for all i, j:

$$(1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2 \leq \|\mathbf{\Pi}\vec{x}_i - \mathbf{\Pi}\vec{x}_j\|_2^2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2.$$

Implies: If $\vec{x}_1, \ldots, \vec{x}_n$ are nearly orthogonal unit vectors in d-dimensions (with pairwise dot products bounded by $\epsilon/8$), then $\frac{\mathbf{\Pi}\vec{x}_1}{\|\mathbf{\Pi}\vec{x}_1\|_2}, \ldots, \frac{\mathbf{\Pi}\vec{x}_n}{\|\mathbf{\Pi}\vec{x}_n\|_2}$ are nearly orthogonal unit vectors in m-dimensions (with pairwise dot products bounded by ϵ).

- Algebra is a bit messy but a good exercise to partially work through.

CONNECTION TO DIMENSIONALITY REDUCTION

Claim 1: n nearly orthogonal unit vectors can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In m dimensions, there are at most $2^{O(\epsilon^2 m)}$ nearly orthogonal vectors.

Claim 1: n nearly orthogonal unit vectors can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In m dimensions, there are at most $2^{O(\epsilon^2 m)}$ nearly orthogonal vectors.

- For both these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$.

Claim 1: n nearly orthogonal unit vectors can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In m dimensions, there are at most $2^{O(\epsilon^2 m)}$ nearly orthogonal vectors.

- For both these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$.
- $2^{O(\epsilon^2 m)} = 2^{O(\log n)} \geq n$.

$\log n \lesssim O(\epsilon^2 m)$

$O\left(\frac{\log n}{\epsilon^2}\right) \lesssim m$

Claim 1: n nearly orthogonal unit vectors can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In m dimensions, there are at most $2^{O(\epsilon^2 m)}$ nearly orthogonal vectors.

- For both these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$.
- $2^{O(\epsilon^2 m)} = 2^{O(\log n)} \geq n$. Tells us that the JL lemma is optimal up to constants.

$m \geq \Omega\left(\frac{\log n}{\varepsilon^2}\right)$

Claim 1: n nearly orthogonal unit vectors can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

Claim 2: In m dimensions, there are at most $2^{O(\epsilon^2 m)}$ nearly orthogonal vectors.

- For both these to hold it must be that $n \leq 2^{O(\epsilon^2 m)}$.
- $2^{O(\epsilon^2 m)} = 2^{O(\log n)} \geq n$. Tells us that the JL lemma is optimal up to constants.
- m is chosen just large enough so that the odd geometry of d-dimensional space still holds on the n points in question after projection to a much lower dimensional space.

Midterm Review

Rough Outline: (subject to small changes)

- Question 1: 4 always, sometimes, nevers.
- Question 2: 4 short answers, sort of like quiz questions.
- Question 3: 5 part question with limited proofs.
- Question 4: 5 part question on analyzing an algorithm. Similar to but easier than a homework question.
- Question 5: Extra credit question touching on high dimensional geometry.

Rough Outline: (subject to small changes)

- Question 1: 4 always, sometimes, nevers.
- Question 2: 4 short answers, sort of like quiz questions.
- Question 3: 5 part question with limited proofs.
- Question 4: 5 part question on analyzing an algorithm. Similar to but easier than a homework question.
- Question 5: Extra credit question touching on high dimensional geometry. 4-5 parts.

You only need to know the statement of the Johnson-Lindenstrauss Lemma, not the proof.

Content or Format Questions?

QUESTIONS

$S = \frac{1}{k}\sum_{j=1}^{k} \min_i h_k(x_i)$ $\qquad \mathbb{E} S = \frac{1}{d+1}$ $\qquad \mathrm{Var}(S) \le \frac{1}{(d+1)^2 k}$

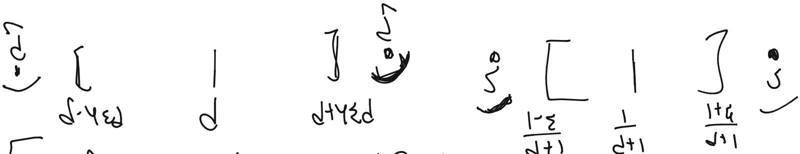
$\hat{d} = \frac{1}{S} - 1$

why does this inequality hold

A $\qquad$ B

① $\Pr(|\hat{d} - d| \ge 4\epsilon d) \le \Pr(|S - \mathbb{E}S| \ge \epsilon \mathbb{E}S) < \text{small}$

② If $|\hat{d} - d| \ge 4\epsilon d$ then $|S - \mathbb{E}S| \ge \epsilon \mathbb{E}S$



if $\hat{d} \ge d + 4\epsilon d$ $\quad$ and if $\hat{d} \le d - 4\epsilon d$

$S \le \frac{1-\epsilon}{d+1}$ $\quad$ then $S \ge \frac{1+\epsilon}{d+1}$

Fully Random Hash Functions

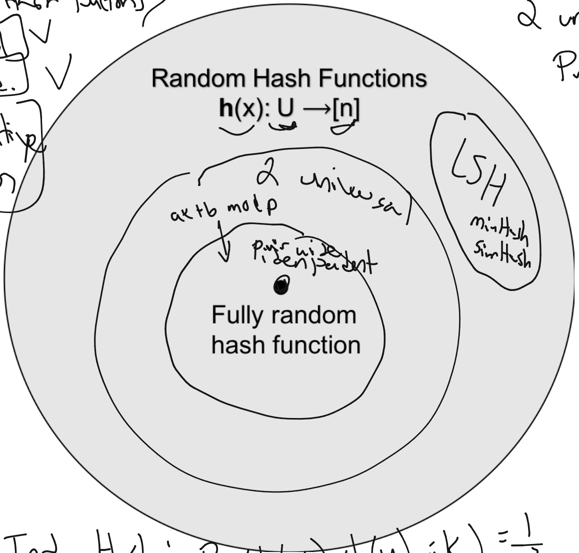
2-universal ✓

pairwise ind. ✓

locality sensitive hash function

2 universal:

$Pr(h(x) = h(y)) \leq \frac{1}{n}$



Pairwise Ind. Hash: $Pr(h(x) = h(y) = k) = \frac{1}{n^2}$

$Pr(h(x) = k) \cdot Pr(h(y) = k) = \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2}$

$Pr(|X - \mu| \le t)$

dis all questions

Concentration Bound Requirements

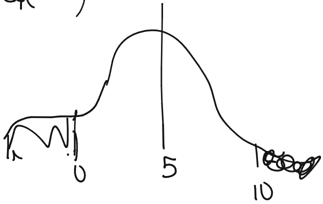
memorize

Markov's	Chebyshev's	Chernoff	Bernstein
non-negative random vars $\mathbb{E}[X]$ $Pr(X > t) \le \frac{\mathbb{E}[X]}{t}$	$\mathbb{E}[X]$ $Var(X)$ $Pr(\lvert X - \mathbb{E}X\rvert \ge t) \le \frac{Var(X)}{t^2}$	X is sum of binary random variables independent $\mathbb{E}[X] = \mu$ $Pr(\lvert X - \mathbb{E}X\rvert \ge \delta\mu) \le 2\exp\left(\frac{-\delta^2\mu}{2+\delta}\right)$	$X_i \in [-M, M]$ $X = \sum X_i$ X_i are independent $Var(X) = \sum Var(X_i)$ $\mathbb{E}(X) = \sum \mathbb{E}(X_i)$

X with $\mathbb{E}[X] = 5$

(A) $Pr(X > 12)$ use chebyshevs.

$\le$ (B) $Pr(|X - \mathbb{E}X| > 7) \le \frac{Var(X)}{49}$



EXAMPLE PROBLEMS

median trick

3. Consider an algorithm $\mathcal{A}$ running in time $T(\mathcal{A})$, that with probability .6 outputs an estimate of the number of triangles in an input graph up to error ± 100, and with probability .4 outputs some bad estimate with worse error. Describe an algorithm that outputs an estimate of the number of triangles in an input graph up to error ± 100 with probability $\geq .99$ and runs in time $O(T(\mathcal{A}))$.

550 $\boxed{880}$ 60 1010 990

median

$\Delta - 100$ $\quad \Delta \quad$ $\Delta + 100$

60

$X > .55t$

t trials $\qquad X = \#$ "succesful trials"

$\mathbb{E}X = .6t \qquad \Pr(X < .55t) < .01$

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2 \mu}{2+\delta}\right)$.

3. Consider an algorithm $\mathcal{A}$ running in time $T(\mathcal{A})$, that with probability .6 outputs an estimate of the number of triangles in an input graph up to error ± 100, and with probability .4 outputs some bad estimate with worse error. Describe an algorithm that outputs an estimate of the number of triangles in an input graph up to error ± 100 with probability $\geq .99$ and runs in time $O(T(\mathcal{A}))$.

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^n X_i\right]$, for any $\delta > 0$, $\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.

   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$.

   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^n X_i\right]$, for any $\delta > 0$, $\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2 \mu}{2+\delta}\right)$.

EXAMPLE PROBLEMS

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.

   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$.

   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

ALWAYS, SOMETIMES, or NEVER:

2. $\Pr[\max(X_1, \dots X_n) \geq t] \leq \sum_{i=1}^{n} \Pr[X_i \geq t]$ for any random variables $X_1, \dots, X_n$.

(c) $\Pr[\mathbf{X} = s \cap \mathbf{Y} = t] = \Pr[\mathbf{X} = s] \cdot \Pr[\mathbf{Y} = t]$.