CS369M: Algorithms for Modern Massive Data Set Analysis **Lecture 1 - 09/23/2009**

The Johnson-Lindenstrauss Lemma

Lecturer: Michael Mahoney *Scribes: Ben Newhouse and Gourab Mukherjee*

**Unedited notes*

1 Background and Motivation

Data can be modeled using many different methods. Two examples being:

- A matrix $A \in \Re^{m*n}$
  m data points
  n features
- A graph G consisting of vertexes V and edges E, which can be represented as (among others):
  1. An adjacency matrix $A_{ij} = \{0, 1\}$
  2. A Laplacian matrix where matrix $L = D - A$ where:
     D - degree matrix (a diagonal matrix representing the number of edges connected to each vertex)
     A - adjacency matrix of above
  3. A normalized Laplacian $\mathscr{L} = D^{\frac{1}{2}} L D^{\frac{1}{2}} = I - D^{\frac{1}{2}} A D^{\frac{1}{2}}$

All of the aforementioned representations are based upon linear algebra. Modern algorithms in data mining take care to employ a number of techniques to ensure that solving problems posed do not become intractable. We can characterize the above data representations in the context of algorithmic complexity using the following tools: for matrices m data points, n features, and M nonzero elements; for graphs $p = |v|, q = |m|$.

2 Taking advantage of dimensionality

In shaping our algorithms to be feasably computable, we can examine our problem space in terms of dimensionality:

Low Dimensionality - Good, computation is inexpensive

High Dimensionality - Good, use measure concentration and the Law of Large numbers (ie. randomness)

Medium Dimensionality - Hard, neither of the above advantages apply

This raises the goal:

Goal: Given points $P = \left\{n \in \Re^d\right\}$, describe P in fewer demensions $k << n, d$. That is, define a function f such that $\forall x_i, x_j \in P$

$$||f(x_i) - f(x_j)||_2 \approx ||x_i - x_j||_2 \tag{1}$$

This dimensionality reduction is import in many different fields, namely:

- Latent Semantic Indexing (LSI) and Information Retrieval (IR) - in addition to reducing storage and computation needs, dimensionality reduction can also be used for clustering of related features (terms).[2]
- Finding nearest neighbors
- Learning mixtures of gaussians
- Compound compressed sensing applications

An important part of understanding dimensionality reduction is the Johnson-Lindenstrauss Lemma. The Johnson-Lindenstrauss Lemma states that any n points in high dimensional euclidian space can be mapped onto k dimensions where $k \geq O(\log n/\epsilon^2)$ without distorting the euclidian distance between any two points more than a factor of $1 \pm \epsilon$. [1]

While the remaining part of these notes pertain to proving that such a mapping exists (rather than how to find it), it is good to keep in mind that it is important to ensure that finding the function f that maps between dimensions is computable itself.

3 Johnson-Lindenstrauss Lemma

Lemma For any $0 < \epsilon < 1$ and any interger n let k be a possitive interger such that

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n \tag{2}$$

then for any set A of n points $\in \Re^d$ there exists a map $f : \Re^d \to \Re^k$ such that for all $x_i, x_j \in A$

$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2 \tag{3}$$

Further this map can be found in randomized ploynomial time. Repeating this projection O(n) times can boost the success probability to any desired constant, giving us a randomized polynomial time algorithm.[1]

General Proof Outline: Construct a random projection over k dimensional subspaces. Prove that the expected value of the euclidian distance of the random projection is equal to the euclidian distance of the original subspace. Prove that the variance of the euclidian distance is greater than the specified error factor only with a probability $\mathscr{F} = 2/n^2$ such that the union bound of this probability across all pairs of points is less than $1 - 1/n$.

3.1 Formal Proof

Definition Let R be a random matrix of order $k \times d$ i.e $R_{ij} \overset{i.i.d}{\sim} N(0,1)$ and u be any fixed vector $\in \Re^d$. Define $v = \frac{1}{\sqrt{k}} R \cdot u$. Thus $v \in \Re^k$ and $v_i = \frac{1}{\sqrt{k}} \sum_j R_{ij} u_j$

Lemma $E\left[||v||_2^2\right] = ||u||_2^2$

Proof

$$\begin{aligned}
E||v||_2^2 &= E[\sum_{i=1}^{k} v_i^2] \quad \text{(Breaking out the summation)} \\
&= \sum_{i=1}^{k} E[v_i^2] \quad \text{(Move factor out)} \\
&= \sum_{i=1}^{k} \frac{1}{k} E[(\sum_{j} R_{ij}u_j)^2] \quad \text{(Introduce a dummy variable } k\text{)} \\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\le j,k\le d} u_j u_k E(R_{ij}R_{ik}) \quad (\delta_{jk} = \{0, j \neq k; 1, j = k\}) \\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\le j,k\le d} u_j u_k \delta_{jk} \\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\le j\le d} u_j^2 \\
&= \sum_{1\le j\le d} u_j^2 \\
&= ||u||_2^2
\end{aligned}$$

Definition Let $X = \frac{\sqrt{k}}{||u||}v$ i.e $x_i = \frac{1}{||u||}R_i^T \cdot u$ for $i = 1(1)k$
$x = \sum_{i=1}^{k} x_i^2 = ||X||_2^2 = \frac{k||v||_2^2}{||u||_2^2}$

Note : $\{v_i : i = 1(1)k\} \overset{i.i.d}{\sim} N(0, ||u||_2^2/k)$. So $X \sim N_k(0, I)$

Lemma The probability $\mathrm{P}\big[||v||_2^2 \geq (1+\epsilon)||u||_2^2\big] > 1 - n^{-2}$

Proof

$$\begin{aligned}
P \quad & \big[||v||_2^2 \geq (1+\epsilon)||u||_2^2||\big] \\
= P \quad & \left[\frac{||u||_2^2 x}{k} \geq (1+\epsilon)||u||_2^2\right] \\
= P \quad & [x \geq (1+\epsilon)k] \\
= P \quad & \left[e^{\lambda x} \geq e^{\lambda(1+\epsilon)k}\right] \quad \text{(for all } \lambda \geq 0 \text{)}
\end{aligned}$$

By Markov's inequality

$$\begin{aligned} P\left[x \geq a\right] &\leq \frac{E\left[x\right]}{a} \\ P\left[e^{\lambda x} \geq e^{\lambda(1+\epsilon)k}\right] &\leq \frac{E\left[e^{\lambda x}\right]}{e^{\lambda(1+\epsilon)k}} \\ &\leq \prod_{i=1}^{k} \frac{E\left[e^{\lambda x_i^2}\right]}{e^{\lambda(1+\epsilon)k}} \quad \text{(as } x_i\text{'s are } i.i.d.\text{)} \\ &\leq \left(\frac{E\left[e^{\lambda x_i^2}\right]}{e^{\lambda(1+\epsilon)k}}\right)^k \\ &\leq \left(\frac{1}{\sqrt{1-2\lambda}\cdot e^{\lambda(1+\epsilon)}}\right)^k \quad \text{(for all } 0<\lambda<1/2 \text{ ; using m.g.f of } \chi^2 \text{ .)} \end{aligned}$$

Setting $\lambda = \frac{\epsilon}{2(1+\epsilon)}$

$$\leq \left[(1+\epsilon)e^{-\epsilon}\right]^{k/2}$$

Using the inequality $log(1+x) < x - x^2/2 + x^3/3$ and (2)

$$\begin{aligned} &\leq e^{-(\epsilon^2/2-\epsilon^3/3)k/2} \leq e^{-2\log n} \\ &\leq n^{-2} \end{aligned}$$

Lemma The probability $\mathrm{P}\left[||v||_2^2 \leq (1-\epsilon)||u||_2^2\right] > 1 - n^{-2}$

Proof

$$\begin{aligned} P &\quad \left[||v||_2^2 \leq (1-\epsilon)||u||_2^2||\right] \\ = P &\quad \left[\frac{||u||_2^2 x}{k} \leq (1-\epsilon)||u||_2^2\right] \\ = P &\quad [x \leq (1-\epsilon)k] \\ = P &\quad \left[e^{-\lambda x} \geq e^{-\lambda(1-\epsilon)k}\right] \quad \text{(for all } \lambda \geq 0\text{)} \end{aligned}$$

By Markov's inequality

$$\begin{aligned} P\left[x \geq a\right] &\leq \frac{E\left[x\right]}{a} \\ P\left[e^{-\lambda x} \geq e^{-\lambda(1-\epsilon)k}\right] &\leq \frac{E\left[e^{-\lambda x}\right]}{e^{-\lambda(1-\epsilon)k}} \\ &\leq \prod_{i=1}^{k} \frac{E\left[e^{-\lambda x_i^2}\right]}{e^{-\lambda(1-\epsilon)k}} \quad \text{(as } x_i\text{'s are } i.i.d.\text{)} \\ &\leq \left(\frac{E\left[e^{-\lambda x_i^2}\right]}{e^{-\lambda(1-\epsilon)k}}\right)^k \\ &\leq \left(\frac{1}{\sqrt{1+2\lambda}\cdot e^{-\lambda(1-\epsilon)}}\right)^k \quad \text{(using the m.g.f. of the } \chi^2 \text{ distribution)} \end{aligned}$$

Setting $\lambda = \frac{\epsilon}{2(1-\epsilon)}$

$$\leq \left[(1-\epsilon)e^{-\epsilon}\right]^{k/2}$$

Using the inequality $log(1-x) < -x - x^2/2$ and (2)

$$\leq n^{-2}$$

Combining the above we get $\mathrm{P}(||v||_2^2 \notin [(1-\epsilon)||u||_2^2, (1+\epsilon)||u||_2^2]) \leq \frac{2}{n^2}$

Since u is an arbitrary d dimensional vector, this probability holds for $u = x_i - x_j$ where x_i, x_j are any two points in A and f is defined as multiplication by the random $k \times d$ matrix. Using the Bonferroni Inequality, taking the disjoint set A consisting of all pairings of points (of count n, thus with $\frac{n(n-1)}{2}$ pairings), the union bound for the probability of any pair of points falling out of the desired error is:

$$\begin{aligned} P(\cup E_i) &\leq \sum_{i=1}^{n} P(E_i) \\ &\leq \frac{n(n-1)}{2}\frac{2}{n^2} \\ &\leq 1 - \frac{1}{n} \end{aligned}$$

Thereby completing our proof.

4 References

1. S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms,* 22(1):60-65, 2003.

2. M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review,* 41(2):335-362, 2001.

3. D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences,* 66(4):671-687, 2003.