Very Sparse Random Projections

Ping Li, Trevor Hastie and Kenneth Church [KDD '06]

Presented by: Aditya Menon

UCSD

March 4, 2009

Dimensionality reduction

- In dimensionality reduction, we want to embed points into a lower dimensional space such that some measure of interest is preserved
- In this talk, we are interested in preserving pairwise Euclidean distances in a specific sense:

The problem

Suppose we have n data points $x_1, \ldots, x_n$, where $x_i \in \mathbb{R}^d$. Given some $k \ll d$ and $\epsilon > 0$, find points $y_1, \ldots, y_n$ where $y_i \in \mathbb{R}^k$ such that for all i, j,

$$(1-\epsilon)||x_i - x_j||^2 \leq ||y_i - y_j||^2 \leq (1+\epsilon)||x_i - x_j||^2$$

- Let us call the y_i's an ϵ-embedding of the input
- How do we find this embedding?

The solution: PCA?

- Simply applying PCA does not necessarily solve this problem
- Recall that PCA focusses on the following global minimization:

$$\min \sum_{i=1}^{n} ||x_i - y_i||^2$$

where $y_i = \sum_{j=1}^{k} (x_i \cdot \hat{u}^{(j)}) \hat{u}^{(j)}$ lies on a linear subspace defined by the basis vectors $\hat{u}^{(j)}$

  - Does not necessarily mean that local pairwise distances are preserved
- So do we have reason to believe such an embedding is easy to find?

Theory: the Johnson-Lindenstrauss lemma

- A classic result of Johnson and Lindenstrauss guarantees the following:

The Johnson-Lindenstrauss lemma [3]

Any n points in $\mathbb{R}^d$ can be ϵ-embedded into $\mathbb{R}^{O(\log n/\epsilon^2)}$

- **Note**: The reduced dimension does not depend on d, but instead on n
- What's the problem with the lemma?
  - It's non-constructive
- **Question**: Is there an easy way to find this mapping?

Construction through randomization

- It turns out we can find an ϵ-approximate embedding through randomization
  - Same guarantee, only probabilistic
- All we need is a random transformation through a Gaussian:

Theorem ([2])

Let the rows of $X \in \mathbb{R}^{n \times d}$ denote the points $x_1, \ldots, x_n$. Let $R \in \mathbb{R}^{d \times k}$ be a matrix whose entries are iid $\mathcal{N}(0, 1)$, and let y_i denote the ith row of the matrix $\frac{1}{\sqrt{k}} XR$. Then, if $k \geq \frac{4 \log n}{\epsilon^2/2 - \epsilon^3/3}$, with high probability,

$$(1 - \epsilon)||x_i - x_j||^2 \leq ||y_i - y_j||^2 \leq (1 + \epsilon)||x_i - x_j||^2$$

- We call this operation a random projection

Why does this projection work?

- It is not difficult to show the following fact:

Theorem

Let x_i, x_j and y_i, y_j be defined as before. Then,

$$\mathbb{E}[||y_i - y_j||^2] = ||x_i - x_j||^2$$

$$Var[||y_i - y_j||^2] = \frac{2}{k}||x_i - x_j||^4$$

- In expectation, distances are preserved with this mapping
- The variance is sufficiently small that certain concentration bounds let us derive a high probability guarantee for all distances
- We will prove the expectation part of this theorem

Preservation of distance: proof

First, recall that by definition

$$y_i = \frac{1}{\sqrt{k}} x_i \cdot R = \frac{1}{\sqrt{k}} \begin{bmatrix} x_i \cdot r_1 & \dots & x_i \cdot r_k \end{bmatrix}$$

So, (focussing on x_1, y_1 for simplicity)

$$\begin{aligned}
\mathbb{E}[||y_1||^2] &= \frac{1}{k} \sum_j \mathbb{E}[(x_1 \cdot r_j)^2] \text{ by linearity of expectation} \\
&= \frac{1}{k} \sum_j \left(\sum_i \mathbb{E}[x_{1i}^2 r_{ij}^2] + 2 \sum_{i<i'} \mathbb{E}[r_{ij} r_{i'j} x_{1i} x_{1i'}] \right) \\
&= \frac{1}{k} \sum_j \sum_i x_{1i}^2 \text{ since the } r_{ij}\text{'s have zero mean and unit variance} \\
&= ||x_1||^2
\end{aligned}$$

Preservation of distance: proof (2)

Now we show that $\mathbb{E}[y_1 \cdot y_2] = x_1 \cdot x_2$:

$$\begin{aligned}
\mathbb{E}[y_1 \cdot y_2] &= \frac{1}{k} \sum_i (x_1 \cdot r_i) \cdot (x_2 \cdot r_i) \\
&= \frac{1}{k} \sum_i \left(\sum_j x_{1j} x_{2j} \mathbb{E}[r_{ij}^2] + \sum_{j \neq j'} x_{1j} x_{2j'} \mathbb{E}[r_{ij} r_{ij'}] \right) \\
&= \frac{1}{k} \sum_i \sum_j x_{1j} x_{2j} \\
&= x_1 \cdot x_2
\end{aligned}$$

Preservation of distance: proof (3)

It is now easy to show the preservation of expectation:

$$\begin{aligned}
\mathbb{E}[||y_1 - y_2||^2] &= \mathbb{E}[||y_1||^2 + ||y_2||^2 - 2y_1 \cdot y_2] \\
&= \mathbb{E}[||y_1||^2] + \mathbb{E}[||y_2||^2] - 2\mathbb{E}[y_1 \cdot y_2] \\
&= ||x_1||^2 + ||x_2||^2 - 2x_1 \cdot x_2 \\
&= ||x_1 - x_2||^2
\end{aligned}$$

- **Surprising observation**: All this required was that the r_{ij}'s were iid with zero mean and unit variance!

Computation time for the projection

- Doing the multiplication XR will take $O(ndk)$ time
- The R that we proposed is that it is dense
  - **Question 1**: Is there is a sparse matrix, perhaps an approximation, we could use instead?
- Perhaps generating Gaussians is too expensive
  - **Question 2**: Can we just use uniform random variables?
- Fortunately, turns out that a much simpler sparse matrix suffices...

A sparse projection matrix

- The following sparse matrix will suffice for a random projection:

Sparse projection matrix [1]

Replace the previous matrix R by a matrix whose entries follow

$$R_{ij} = \sqrt{3} \cdot \begin{cases} +1 & \text{probability } 1/6 \\ 0 & \text{probability } 2/3 \\ -1 & \text{probability } 1/6 \end{cases}$$

Then, the same guarantee as before holds.

- Same asymptotic complexity, but potentially a factor of 3 faster to compute
- **Note**: This distribution matches the first three moments of the Gaussian

This paper: an even sparser projection matrix

- Proves the theoretical viability of the following very sparse distribution

$$R_{ij} = \sqrt{s} \cdot \begin{cases} +1 & \text{probability } 1/2s \\ 0 & \text{probability } 1 - 1/s \\ -1 & \text{probability } 1/2s \end{cases}$$

  for any $s \geq 3$
- In fact, shows that $s = \sqrt{d}$ gives only small additional loss in accuracy!
- Considerable computational savings
  - With $s = \sqrt{d}$, computation is $O(n\sqrt{d}k)$

Preservation of distance

- As before, the very sparse R will preserve distance in expectation
- It does introduce some extra variance in our prediction, however

Theorem

Let R denote the very sparse matrix defined above, with parameter s. Then,

$$\mathbb{E}[||y_1 - y_2||^2] = ||x_1 - x_2||^2$$

$$Var[||y_1 - y_2||^2] = \frac{2}{k}||x_1 - x_2||^4 + \frac{(s-3)}{k}\sum_j (x_{1j} - x_{2j})^4$$

- The extra term in the variance is initially unsettling, but we will be able to prove it converges quickly to the case of a normal distribution

Asymptotic behaviour of distribution

- What is the effect of the extra term in the variance?
- Paper shows that this extra term diminishes as $d \to \infty$, meaning that the very sparse distribution is asymptotically normal

Theorem

Suppose $s = o(d)$. *As* $d \to \infty$, *we have*

$$(s-3)\frac{\sum_j (x_{1j} - x_{2j})^4}{||x_1 - x_2||^4} \to 0$$

at the rate $\sqrt{\frac{s-3}{d}}$. *As a consequence,*

$$Var[||y_1 - y_2||^2] \to \frac{2}{k}||x_1 - x_2||^4$$

- To prove this, we need one piece of sophisticated machinery...

The strong law of large numbers

Theorem

Let $X_1, X_2, \ldots, X_n$ be iid random variables with means μ. Let $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$. Then,

$$(\forall \epsilon > 0) \lim_{n \to \infty} \Pr[|\bar{X} - \mu| < \epsilon] = 1$$

That is, with probability 1, the mean of the variables approaches the true mean.

Proof is out of scope for this talk! But now we can prove the theorem...

Asymptotic behaviour of distribution: proof

Let us look at the extra term in the variance, and show that it tends to zero:

$$
\begin{aligned}
(s-3)\frac{\sum_j (x_{1j}-x_{2j})^4}{||x_1-x_2||^4} &= \frac{(s-3)}{d}\frac{(\sum_j (x_{1j}-x_{2j})^4)/d}{(||x_1-x_2||^2/d)^2} \\
&\to \frac{(s-3)}{d}\frac{\mathbb{E}[(x_{1j}-x_{2j})^4]}{(\mathbb{E}[(x_{1j}-x_{2j})^2])^2} \\
&\to 0
\end{aligned}
$$

where the second line is by the strong law of large numbers. The rate of convergence is $O(\sqrt{(s-3)/d})$.

Implications thus far

- The very sparse distribution preserves distances in expectation
- The variance of the estimator tends to that with a normal distribution at the rate $O(\sqrt{(s-3)/d})$
- Consider $s = \sqrt{d}$, e.g.
  - This converges to the normal estimator at the rate $O(d^{-1/4})$
  - When d is large, this rate is extremely quick
  - Considerable speedup over normal/Achlioptas distribution
- So, we have a theoretically sound way to do dimensionality reduction in $O(n\sqrt{d}k)$ time
  - In fact, even $s = \frac{d}{\log d}$ will work, but with higher error

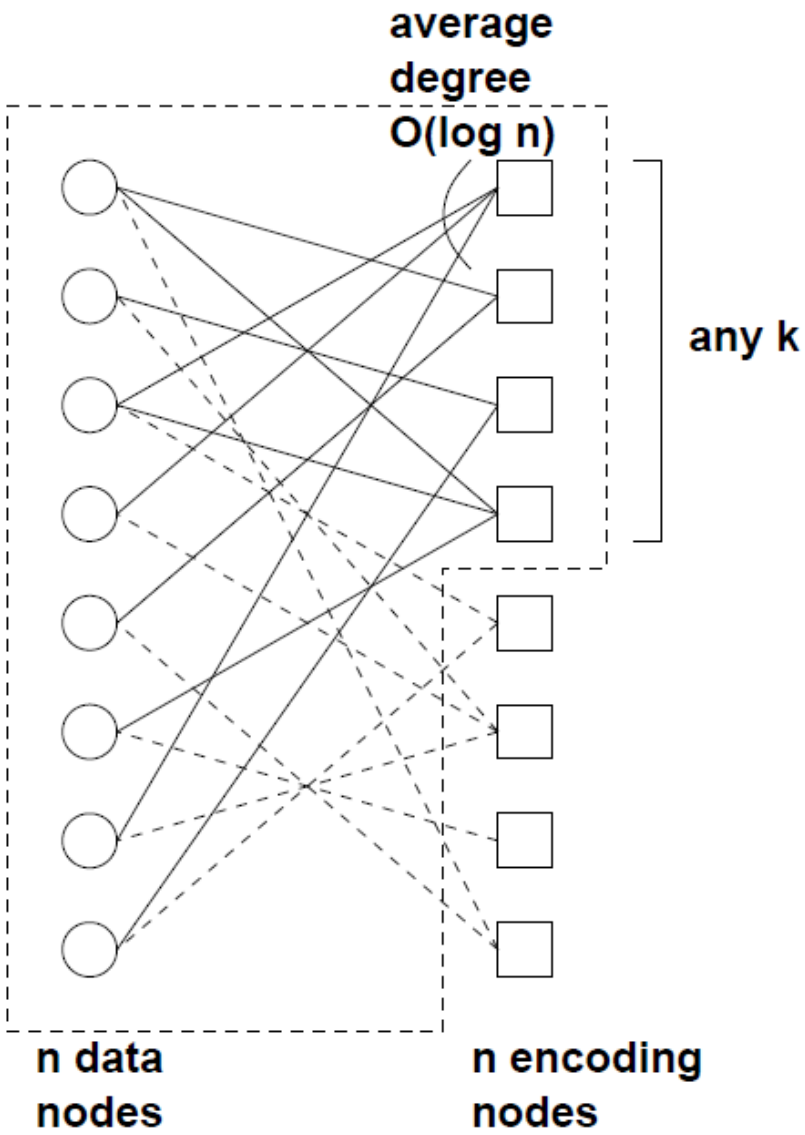
Application: wireless sensor network [4]

- Say we have d sensors, each of which generates some value x_i: combined, this describes a vector $(x_1, \ldots, x_d)$
- **Problem**: We'd like to query a few (say k) of these sensors, and receive a good approximation $(y_1, \ldots, y_k)$ to the true observations
  - Not unreasonable that the observations are tightly compressible
  - Would like to do this without global coordination
- Is there a simple solution?

Application: wireless sensor network

- Can use sparse projections with $s = \frac{d}{\log d}$:
  1. Each sensor i generates $r_{i1}, \ldots, r_{id}$ $\leftarrow$ on average $\log d$ values
  2. For each non-zero r_{ij}, it requests u_j and so computes $u \cdot r_i$
  3. Picking these dot-products for any k sensors gives us a good approximation

Application: wireless sensor network



Questions?


Dimitris Achlioptas.
Database-friendly random projections.
In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, New York, NY, USA, 2001. ACM.


Sanjoy Dasgupta and Anupam Gupta.
An elementary proof of the Johnson-Lindenstrauss Lemma.
Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, 1999.


W.B. Johnson and J. Lindenstrauss.
Extensions of Lipschitz maps into a Hilbert space.
In *Conference in Modern Analysis and Probability (1982, Yale University)*, volume 26 of *Contemporary Mathematics*, pages 189–206. AMS, 1984.


Wei Wang, Minos Garofalakis, and Kannan Ramchandran.
Distributed sparse random projections for refinable approximation.

In *IPSN '07: Proceedings of the 6th international conference on Information processing in sensor networks*, pages 331–339, New York, NY, USA, 2007. ACM.