15-750: Graduate Algorithms

Dimensionality Reduction:
  Johnson-Lindenstrauss Transform
  Principal Component Analysis
High dimensional vectors

Common in many real-world applications
  E.g., Documents, Movie or product ratings by users, gene expression data

Often face the “curse of dimensionality”
  E.g., Nearest-neighbor search

Dimension reduction: Transform the vectors into lower dimension while retaining useful properties
High dimensional vectors

We will study two techniques

(1) Maintaining distances approximates (Johnson-Lindenstrauss Transform)

(2) Maintaining most of the variance in the data points (Principal Component Analysis)
Dimensionality Reduction:

Johnson-Lindenstrauss Transform

Principal Component Analysis
Subsequent material will be on board
Johnson-Lindenstrauss Transform

• Linear transformation
  • Specifically, multiply vectors with a specially chosen matrix
  • Preserves pairwise distances (L2) between the data points

Set of points $X = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^D$
Original dimension $= D$

Want a linear transform $S: \mathbb{R}^D \rightarrow \mathbb{R}^K$
Final dimension after reduction $= k$
**Johnson-Lindenstrauss Transform**

**JL Lemma:**
Let $\varepsilon \in (0, 1/2)$. Given any set of points $X = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^D$, there exists a map $S: \mathbb{R}^D \rightarrow \mathbb{R}^K$ with $k = O(\varepsilon^{-2}\log n)$ s.t
\[(1-\varepsilon) \|x_i - x_j\|^2 \leq \|Sx_i - Sx_j\|^2 \leq (1+\varepsilon) \|x_i - x_j\|^2\]

**Observations:**
- The final dimension is independent of the original dimension $D$ (can be very large) and is dependent only on the number of points $n$ and the accuracy parameter $\varepsilon$
- Log factor! E.g., can map 10 billion points ($\sim 2^{30}$) to $O(30 \varepsilon^{-2})$ dimensions!
Johnson-Lindenstrauss Transform

Construction:

Let A be a \((K \times D)\) matrix, such that every entry of A is filled with an i.i.d. draw from a standard Normal \(N(0,1)\) distribution (a.k.a. the Gaussian distribution)

Define the transformation matrix \(S := \frac{1}{\sqrt{K}} A\).

Transformation: The point \(x \in \mathbb{R}^D\) is mapped to \(Sx\)

- I.e.: Just multiply with a Gaussian matrix and scale with \(\frac{1}{\sqrt{K}}\)
- The construction does not even look at the set of points \(X\)

Proof sketch next.
Proof sketch for JL Transform

Gaussian random variables

Continous R.V.

Notation: $N(\mu, \sigma^2)$

PDF: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- Very good concentration
- Tapers off exponentially

$-\frac{1}{2\sigma^2}$
Proof sketch for JL Transform

Gaussian random variables

Properties:

P1. \( x_i \sim N(\mu_i, \sigma_i^2) \) \& independent

Then \( x_1 + x_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \)

P2. \( x \sim N(\mu, \sigma^2) \Rightarrow aX \sim N(a\mu, a^2 \sigma^2) \)

P3. \( x_i \sim N(0, 1) \Rightarrow \text{“Standard Normal”} \)

\[ \Rightarrow \sum a_ix_i \sim N(0, \sum a_i^2) \]

\[ \Rightarrow N(0, \|a\|^2) \]

\( \Rightarrow \) L2 norm (or Euclidean length) of \( a \)
Proof sketch for JL Transform

**Fact:** \( X \in \mathbb{R}^D \), \( A = \begin{bmatrix} A_1 & \cdots & A_k \end{bmatrix} \in \mathbb{R}^{(k \times D)} \)

\[ S \mathbf{x} = \frac{1}{\sqrt{k}} A \mathbf{x} = \begin{bmatrix} \frac{1}{\sqrt{k}} \langle A_1, \mathbf{x} \rangle \\ \vdots \\ \frac{1}{\sqrt{k}} \langle A_k, \mathbf{x} \rangle \end{bmatrix} \sim N(0, \frac{\| \mathbf{x} \|^2}{k}) \]

\[ \sum_{i=1}^{k} a_i \mathbf{x}_i \sim N(0, \| \mathbf{x} \|^2) \]
**Fact 2**  For any $Y \in \mathbb{R}^D$

$$E \left[ \| SY \|^2 \right] = E \left[ \sum_{i=1}^{k} \frac{1}{k} \langle A_i, Y \rangle^2 \right]$$

$$= \sum_{i=1}^{k} \frac{1}{k} E \left[ \langle A_i, Y \rangle^2 \right]$$

$$= \text{Var} \left( \langle A_1, Y \rangle \right)$$

$$= \langle \text{i.i.d} \rangle$$

$$= \| Y \|^2$$
Proof sketch for JL Transform

Expected value of the distance after transformation

\[ E \left[ \| Sx_1 - Sx_2 \|^2 \right] = E \left[ \| S(x_1 - x_2) \|^2 \right] \]

\[ = E \left[ \| SY \|^2 \right] \quad \text{where} \quad Y = x_1 - x_2 \in \mathbb{R}^d \]

\[ = \| Y \|^2 \quad \text{(from Fact 3)} \]

\[ = \| x_1 - x_2 \|^2 \]

Thus, the expected result (of preserving pairwise L2 distance) holds in expectation.

Now, need to take this to a high probability result
Proof sketch for JL Transform

We will use the following concentration bound

Lemma 1:
Let $U_1, \ldots, U_k$ be i.i.d $N(0, \sigma^2)$

Let $Z = \sum_{i=1}^{k} U_i^2$

$E(Z) = k \sigma^2$

Then for $\epsilon \in (0, \frac{1}{2})$, some constant $c$

$$\Pr \left[ \left| Z - E[Z] \right| \geq \epsilon E[Z] \right] \leq \frac{k \epsilon^2}{c}$$