A Modeling Framework for Scalable Near-Duplicate Detection Using Random Projections and Locality-Sensitive Hashing

Johnson–Lindenstrauss Lemma

Rahul Ghosh (MA25M021)

February 8, 2026

Group 3

- Rahul Ghosh (MA25M021)
- Ankit Gangwar (MM25D950)
- Puneet (ID25S027)
- Tanmoy Ghosh (MA25M026)

Problem Setting

Given n points

$$X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$$

Goal:

- Reduce dimension from d to $k \ll d$
- Approximately preserve all pairwise distances

Johnson–Lindenstrauss Lemma

For any $0 < \varepsilon < 1$, there exists a map

$$f : \mathbb{R}^d \to \mathbb{R}^k$$

with

$$k = O\left(\frac{\log n}{\varepsilon^2}\right)$$

such that for all i, j,

$$(1 - \varepsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1 + \varepsilon)\|x_i - x_j\|^2$$

Random Projection Construction

Let

$$R \in \mathbb{R}^{k \times d}$$

with entries

$$R_{ij} \sim \mathcal{N}(0, 1)$$

Define

$$f(x) = \frac{1}{\sqrt{k}} Rx$$

This is a **random linear map**.

Reduction to Norm Preservation

For any two points x_i, x_j:

$$\|f(x_i) - f(x_j)\| = \|f(x_i - x_j)\|$$

Hence, it suffices to show:

$$\|f(u)\|^2 \approx \|u\|^2 \quad \text{for fixed } u \in \mathbb{R}^d$$

Distribution of One Projection

Let r_i be the i-th row of R.

$$r_i \cdot u = \sum_{j=1}^{d} r_{ij} u_j$$

Since $r_{ij} \sim \mathcal{N}(0, 1)$,

$$r_i \cdot u \sim \mathcal{N}(0, \|u\|^2)$$

Sum of Squares

$$\|Ru\|^2 = \sum_{i=1}^{k} (r_i \cdot u)^2$$

Define

$$Z_i = \frac{r_i \cdot u}{\|u\|} \quad \Rightarrow \quad Z_i \sim \mathcal{N}(0, 1)$$

Then

$$\|Ru\|^2 = \|u\|^2 \sum_{i=1}^{k} Z_i^2$$

Chi-Square Distribution

By definition:

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$$

Hence:

$$\|Ru\|^2 = \|u\|^2 \chi_k^2$$

And:

$$\|f(u)\|^2 = \|u\|^2 \cdot \frac{1}{k} \chi_k^2$$

Goal After Reduction

Since

$$\|f(u)\|^2 = \|u\|^2 \cdot \frac{1}{k}\chi_k^2,$$

norm preservation is equivalent to concentration of $\frac{1}{k}\chi_k^2$ around 1.

We must show:

$$\Pr\left(\left|\frac{1}{k}\chi_k^2 - 1\right| \geq \varepsilon\right) \leq 2e^{-c\varepsilon^2 k}$$

MGF of Chi-Square

If $X \sim \chi^2_k$, its moment generating function is:

$$\mathbb{E}[e^{tX}] = (1 - 2t)^{-k/2}, \quad t < \frac{1}{2}$$

This allows exponential tail bounds via Chernoff's method.

Chernoff Bound: Upper Tail (Step 1)

Let $X \sim \chi^2_k$ and $\varepsilon > 0$.
For any $t > 0$, by Markov's inequality:

$$\Pr(X \geq (1+\varepsilon)k) = \Pr(e^{tX} \geq e^{t(1+\varepsilon)k}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\varepsilon)k}}$$

Using the MGF:

$$\Pr(X \geq (1+\varepsilon)k) \leq \frac{(1-2t)^{-k/2}}{e^{t(1+\varepsilon)k}}$$

Chernoff Bound: Upper Tail (Step 2)

Choose

$$t = \frac{\varepsilon}{2(1+\varepsilon)} < \frac{1}{2}$$

Then:

$$\Pr(X \geq (1+\varepsilon)k) \leq \exp\left(-\frac{k}{2}(\varepsilon - \ln(1+\varepsilon))\right)$$

Using:

$$\varepsilon - \ln(1+\varepsilon) \geq \frac{\varepsilon^2}{2}, \quad 0 < \varepsilon < 1,$$

we obtain:

$$\Pr(X \geq (1+\varepsilon)k) \leq e^{-\frac{k\varepsilon^2}{4}}$$

Chernoff Bound: Lower Tail

Similarly, for $0 < \varepsilon < 1$:

$$\Pr(X \leq (1-\varepsilon)k) \leq e^{-\frac{k\varepsilon^2}{4}}$$

Combining both tails:

$$\Pr\left(\left|\frac{1}{k}\chi_k^2 - 1\right| \geq \varepsilon\right) \leq 2e^{-c\varepsilon^2 k}$$

Norm Preservation for One Vector

Therefore:

$$\Pr\Big((1-\varepsilon)\|u\|^2 \leq \|f(u)\|^2 \leq (1+\varepsilon)\|u\|^2\Big) \geq 1 - 2e^{-c\varepsilon^2 k}$$

How Many Pairs?

Given n points, the number of unordered pairs is:

$$\binom{n}{2} = \frac{n(n-1)}{2} \leq n^2$$

Each pair gives one distance constraint.

Definition of the Bad Event E_{ij}

For each unordered pair (i, j), define:

$$E_{ij} = \left\{ \|f(x_i) - f(x_j)\|^2 \notin \left[(1-\varepsilon)\|x_i - x_j\|^2, \, (1+\varepsilon)\|x_i - x_j\|^2 \right] \right\}$$

That is, E_{ij} occurs if the distance is not preserved.

From previous steps:

$$\Pr(E_{ij}) \leq 2e^{-c\varepsilon^2 k}$$

Union Bound

By the union bound:

$$\Pr\left(\bigcup_{i<j} E_{ij}\right) \leq \sum_{i<j} \Pr(E_{ij}) \leq 2n^2 e^{-c\varepsilon^2 k}$$

Choosing Dimension k

We want:

$$2n^2 e^{-c\varepsilon^2 k} < 1$$

This holds if:

$$k \geq C \frac{\log n}{\varepsilon^2}$$

for a sufficiently large constant C.

Conclusion

With positive probability:

$$(1-\varepsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2$$

for all pairs (i, j).

Hence, a Johnson–Lindenstrauss embedding exists.