Khoury College of Computer Sciences | Fall 2023
Northeastern University | 17 October 2023
CS 7870: Seminar in Theoretical Computer Science | Scribe: Eric Chapdelaine

Outline:

- Motivation
- Random Projection Theorem
- Johnson-Lindenstrauss Transform

The presentation of the material in this lecture is heavily drawn from Chapters 2 and 3 of the text by Blum, Hopcroft, and Kannan [BHK20].

1 Motivation

Let's consider a common task in high-dimensional data: nearest neighbor search. We are given a set S of n points in $\mathbb{R}^d$ and are asked to find a point in S that is nearest a given query point. It is straightforward to solve the problem in time polynomial in n, but in many applications n can be very large, and we cannot afford to spend time dependent on n for every query. In Lecture 2, we explored the use of nets to solve nearest neighbor efficiently for large n, when d is small; our running time was exponential in d and logarithmic in the diameter of S. A challenge we face is that when n and d become very large, the preceding approaches for the nearest-neighbors problem cannot be executed in a reasonable time. A clever approach to combat this issue is to reduce the dimensionality of the dataset by projecting the points to a k-dimensional space with $k \ll d$ while (approximately) preserving the pairwise distances between the points.

To do this, we will explore two ideas: the Random Projection Theorem and the Johnson-Lindenstrauss Transform. Recall the Spherical Gaussian Annulus Theorem from previous lectures.

Theorem 1. Suppose x is drawn from a spherical Gaussian of dimension d with 0-mean and unit variance. There exists a constant $c > 0$ such that with probability $\geq 1 - 3e^{-c\beta^2}$

$$\sqrt{d} - \beta \leq \|x\| \leq \sqrt{d} + \beta$$

2 Random Projection Theorem

Let's first consider what happens when we project a unit vector v with a vector u taken from a 0-mean, unit variance spherical Gaussian. Recall definition: $v \cdot u = \sum_{i=1}^{d} v_i u_i$

$$\mathbb{E}(v \cdot u) = \sum_{i=1}^{d} v_i \cdot \mathbb{E}(u_i) = 0$$

$$\text{Var}(v \cdot u) = \sum_{i=1}^{d} v_i^2 \cdot 1 = 1$$

This implies $v \cdot u$ is also a 0-mean, unit variance spherical Gaussian. Now let's pick vectors $u_1, u_2, \cdots, u_k$ from a 0 mean, unit-variance spherical Gaussian and consider the projection $f : \mathbb{R}^d \to \mathbb{R}^k$:

$$f(v) = (v \cdot u_1, v \cdot u_2, \cdots, v \cdot u_k)$$

By above, we know that each of the coordinates of $f(v)$ is a normally distributed variable with a mean of 0 and variance 1. Therefore, $f(v)$ is a spherical Gaussian distributed variable with a mean of 0 and variance of 1.

Because the random projection above is also a spherical Gaussian, the theorem below follows by the Gaussian annulus theorem.

Theorem 2. Let v be a fixed vector in $\mathbb{R}^d$ and let f be the projection described above. There exists a constant $c > 0$ such that for $\epsilon \in (0, 1)$, with probability $\leq 1 - 3e^{c\epsilon^2 k}$

$$\sqrt{k}(1 - \epsilon) \leq ||f(v)|| \leq \sqrt{k}(1 + \epsilon)$$

3 Johnson-Lindenstrauss Transform

The theorem above states that the length of a projection of a single vector only differs from its expected value with very low probability. We can then apply this to all pairwise distances in a given dataset. We can use a union bound to say that all pairwise distances are preserved with high probability. This idea describes the Johnson-Lindenstrauss Theorem explicitly stated below [JL].

Theorem 3. Let f be the random projection described above. Suppose $v_1, v_2, \cdots, v_n$ are points in d-dimension Euclidean space. For all $\epsilon \in (0, 1)$, if $k \geq \frac{3\ln(n)}{c\epsilon^2}$, then with probability $1 - \frac{3}{n}$, for all i, j

$$\sqrt{k}(1 - \epsilon)||v_i - v_j|| \leq ||f(v_i) - f(v_j)|| \leq \sqrt{k}(1 + \epsilon)||v_i - v_j||$$

Informally, this transforms points from a d-dimensional space to a k-dimensional space while (essentially) preserving distances between points.

Proof. We know that $f(v_i) - f(v_j) = f(v_i - v_j)$ because f is a linear transformation. It follows by the Random Projection Theorem, with probability $\leq 1 - 3e^{-c\epsilon^2 k}$, we have

$$\sqrt{k}(1 - \epsilon)||v_i - v_j|| \leq ||f(v_i) - f(v_j)|| \leq \sqrt{k}(1 + \epsilon)||v_i - v_j||$$

To get the success probability of $1 - \frac{3}{n}$, we can bound the probability that any pair of points fails to be $\frac{3}{n^3}$. Therefore, we want

$$\begin{aligned} 3e^{-c\epsilon^2 k} &= 3e^{-3\ln(n)} \\ &= \frac{3}{n^3} \\ \implies k &= \frac{3c\ln(n)}{\epsilon^2} \end{aligned}$$

□

References

[BHK20] Avrim Blum, John Hopcroft, and Ravi Kannan. *Foundations of Data Science*. Cambridge University Press, Cambridge, 2020.

[JL] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206. Conference in modern analysis and probability.