High Dimensional Space

Jayati Kaushik

St. Joseph's University, Bengaluru

Foundations of Data Science
BDA2121

Random Projection

Consider the following projection $f : \mathbb{R}^d \to \mathbb{R}^k$:
Pick k Gaussian vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k} \in \mathbb{R}^d$ with unit variance coordinates. For any vector $\mathbf{v}$, define the projection $f(\mathbf{v})$ by:

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

Applications

The Random Projection Theorem

Let $\mathbf{v}$ be a fixed vector in $\mathbb{R}^d$ and let f be defined as before. There exists $c > 0$ such that for $\epsilon \in (0, 1)$,

$$Prob(||f(\mathbf{v})| - \sqrt{k}|\mathbf{v}|| \geq \epsilon\sqrt{k}|\mathbf{v}|) \leq 3e^{-ck\epsilon^2}$$

where the probability is taken over the random draws of vectors $\mathbf{u_i}$ used to construct f.

Johnson-Lindenstrauss Lemma

For any $0 < \epsilon < 1$ and any integer n, let $k \geq \frac{3}{c\epsilon^2}\ln n$. For any set of n points in $\mathbb{R}^d$, the random projection f has the proeprty that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $11 - 3/2n$

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

Applications

Separating Gaussians