CSE 5095: Research Topics in Big Data Analytics
Lecture on Apr 17, 2014
Prof. Sanguthevar Rajasekaran

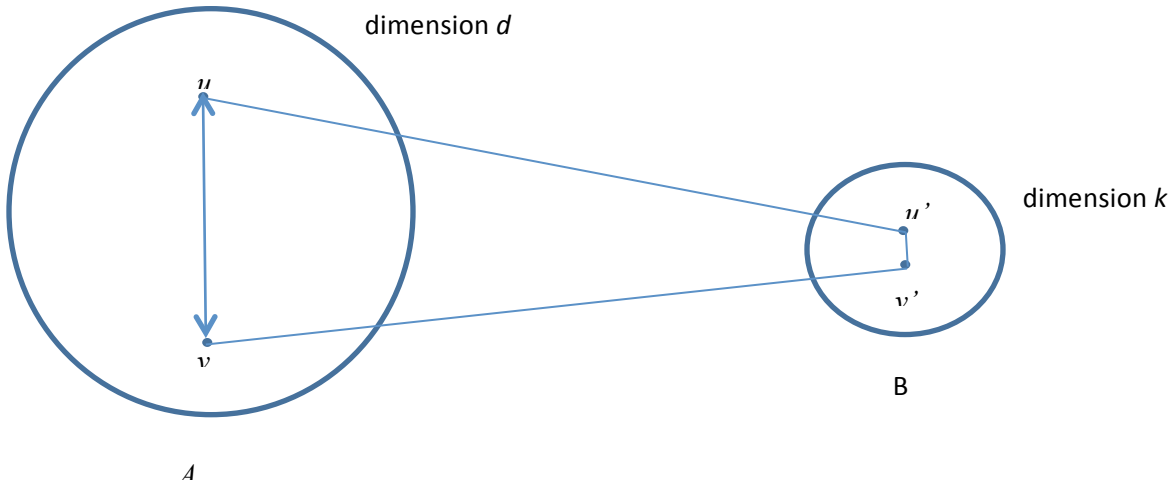
Notes prepared By: Priya Periaswamy

Random Projection:

In this lecture we will discuss data reduction techniques that play a major role while dealing with big data. If there are terabytes of data, it may not be feasible to analyze the entire dataset. If we can reduce the size of the data such that we preserve the information content as much as possible that will be great since we may be able to process the reduced dataset effectively.

For high-dimensional datasets, dimension reduction is usually performed in order to avoid the effects of the curse of dimensionality.

Random Projection was first proposed by Johnson & Lindenstrauss (1984)



The idea here is to project from a high dimensional space to a lower dimensional space.
The distance between any two points should be closely preserved. (In the above figure, points u and v are from a high dimensional space A with dimension d. Their distance should be preserved when projected to a lower dimensional space B with dimension k).

Application: Analyzing genome data, medical image analysis, etc.,

Theorem:

Let S be any set of n points from d-dimensional space.

Let $k \geq 4\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)^{-1} \log_e n$

where ε is a constant, $1 > \varepsilon > 0$.

Then $\exists$ a projection f of points in S into a k-dimensional space such that distance is closely preserved for ay pair of point u and v from S.

$$(1-\varepsilon)\ \|u-v\|^2 \leq \|f(u)-f(v)\|^2 \ \leq\ (1+\varepsilon)\|u-v\|^2$$

(Please note || is the L2 norm).

We can find such a projection in randomized polynomial time.

Proof:

(By Dasgupta & Gupta, 2001)

It suffices to show that the length of a unit vector in the original space O is closely preserved in the random projected space P.

It suffices to show that the length of a random unit vector in O is closely preserved in a fixed k dimensional subspace P.

Use the subspace spanned by first k co-ordinates.

Let $X = X_1, X_2, \ldots, X_d$

Where $X_i = N(0,1)\ \forall\ i$. (Here N is the standard normal distribution with $\mu = 0$ and $\sigma = 1$).

$X_1, X_2, \ldots, X_d$ are independent.

$$Y = \frac{1}{\|X\|}(X_1, X_2, \cdots, X_d).$$

Let Z be the projection of Y in the k-dimensional space. Specifically,

$$Z = \frac{1}{\|X\|}(X_1, X_2, \cdots, X_k).$$

Let L be the norm of Z.

$$E(L) = \frac{k}{d}(norm\ of\ Y)$$

$$Please\ note\ norm\ of\ Y = 1$$

Thus, $E(L) = \frac{k}{d}$.

Claim:

$$Prob\left(L \leq \beta \frac{k}{d}\right) \leq \exp(\frac{k}{2}(1-\beta+\ln\beta)) \text{ for any } \beta < 1$$

$$Prob\left(L \geq \beta \frac{k}{d}\right) \leq \exp\left(\frac{K}{2}(1-\beta+\ln\beta)\right) \text{ for any } \beta > 1.$$

Assume that the claim is true. Let v_i and v_j be any two points in the d-dimensional space and let their projections in the k-dimensional space be v_i' and v_j', respectively. Let $L = \left\|v_i' - v_j'\right\|^2$ *and* $\mu = (\frac{k}{d})\left\|v_i - v_j\right\|^2$. Then it follows that

$$Prob[L \leq (1-\varepsilon)\mu] \leq \exp\left(\frac{k}{2}\left(1-(1-\varepsilon)-\varepsilon-\frac{\varepsilon^2}{2}\right)\right) \text{ using the fact } ln(1-\varepsilon) \leq \left(-\varepsilon-\frac{\varepsilon^2}{2}\right) \text{ for } 0 \leq \varepsilon < 1$$

$$\leq \exp\left(-\frac{k}{4}\varepsilon^2\right)$$

$$\leq \exp(-2\ln n)$$

$$= \frac{1}{n^2}.$$

Similarly,

$$Prob[L \geq (1+\varepsilon)\mu] \leq \exp\left(\frac{k}{2}\left(1-(1-\varepsilon)+\varepsilon-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)\right)$$

$$\text{using the fact } ln(1+\varepsilon) \leq \left(\varepsilon-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)$$

$$= \exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)\right)$$

$$\leq \exp(-2\ln n)$$

$$= \frac{1}{n^2}.$$

We will complete the proof of the JL theorem in the next class.