CS 683 Advanced Design and Analysis of Algorithms April 07, 2008

Lecture 30: High Dimensional Data

Instructor: John Hopcroft *Scribe: Cristian Danescu Niculescu-Mizil, Yao Yue*

Continue on Altman and Tennenboltz's Five Axioms

Axiom 4 (Collapsing) We can collapse two vertices into one without changing the rank of remaining vertices, if the out-going edges of the two vertices lead to the same set of vertices, as in Figure 1.

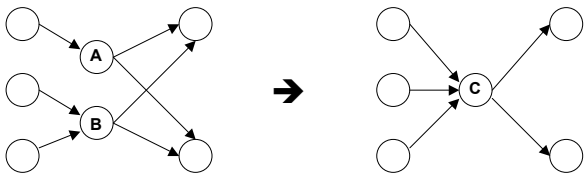


Figure 1: Axiom 4: collapsing A and B into C does not change the rank of the remaining vertices.

Axiom 5 (Proxy) If A, B and C have the same rank, and they all point to the same vertex D then D can be removed by letting A, B and C pointing directly to the successors of D (as in Figure 2); this operation will not change the rank of the remaining vertices.

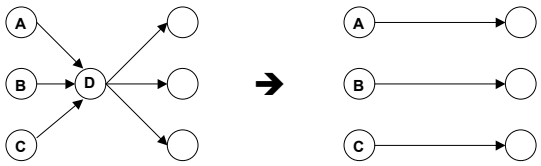


Figure 2: Axiom 5: The proxy node D can be removed without changing the rank of the remaining vertices.

Altman and Tennenboltz [1] proved that pagerank is the only ranking that satisfies all five axioms. However, it is not so interesting in realistic scenarios because it only applies to strongly connected and unweighted graphs (Thus restarting is not allowed).

High Dimensional Data

High dimensional data are different from their low dimensional counterparts in many ways. Consider the following aspects:

Number of Data Points

In low dimensions, the number of data points is usually much more than dimensions. However, in high dimensions, the number of data points might be less than dimensions.

Volume

A unit hypercube always has volume of 1 unit (be it square unit, cubic unit or d-unit). However, this doesn't hold for unit sphere. The volume of a unit sphere goes to 0 when d (dimension) goes to infinity (we will prove this later in the lecture).

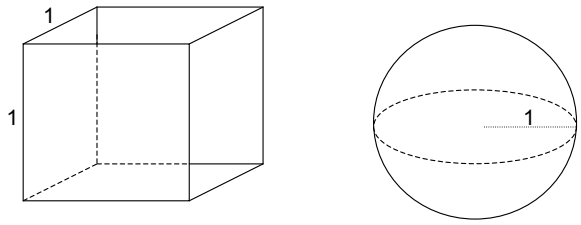


Figure 3: A unit cube and a unit sphere in three dimensions

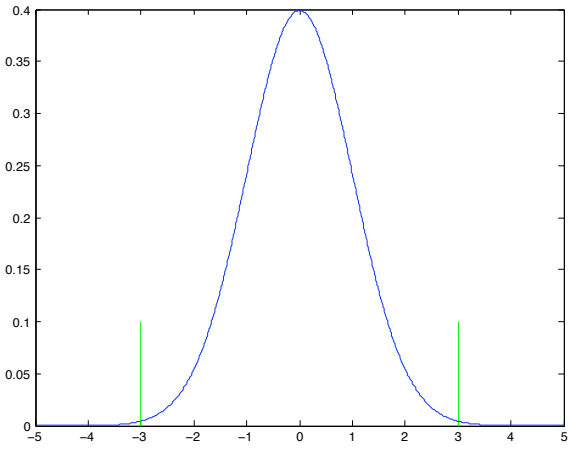


Figure 4: The probability mass of a one-dimensional standard Gaussian lays almost entirely in the interval (-3,3)

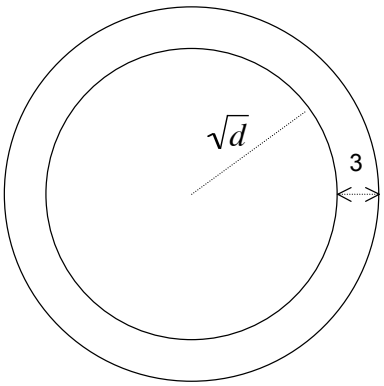


Figure 5: In high dimensions, there is almost no volume in the sphere of radius smaller than $\sqrt{d}$, and, therefore, the probability mass in that region is close to 0.

Gaussian Distribution

One dimensional standard Gaussian distribution has all its probability mass in interval $(-3, 3)$ (Figure 4); in high-dimensional standard Gaussian distribution, the probability mass falls into a annulus of radii $\sqrt{(d)}$ and $\sqrt{d}+3$ (Figure 5). This is consequence of the fact that a sphere of radius smaller than $\sqrt{d}$ has volume 0 when d goes to infinity.

Shortest/Longest Distance Ratio

Consider a set of uniformly distributed points. On a 2-dimensional plane, the shortest distance between two points tends to be much smaller than the longest distance between two points (Figure 6). However, in high dimensions, the ratio between shortest distance and longest distance is very close to 1.

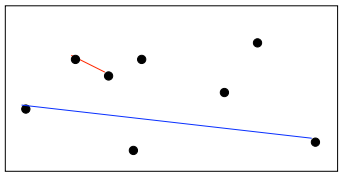
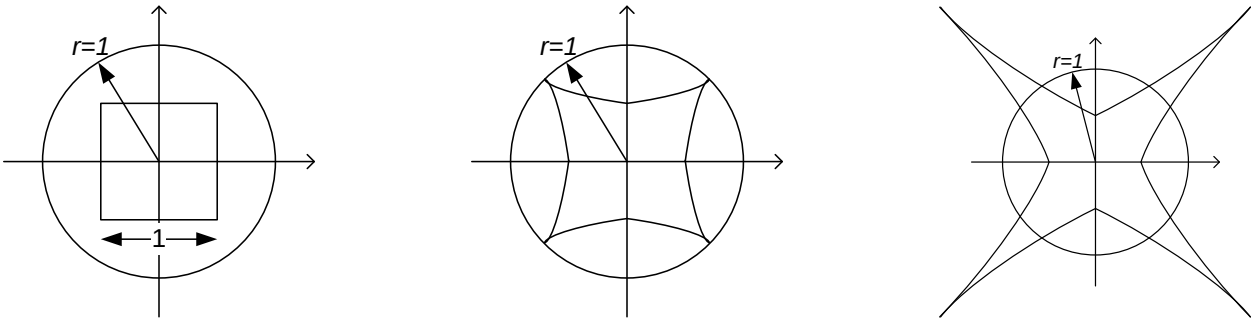


Figure 6: The shortest distance between two points (in red) is much smaller than the longest distance between two points (in blue).

The distance between points x, y is defined as $\sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$. Given that the coordinates are generated uniformly at random, we can show that the distance converges to expected value of distance between two points as d goes to infinity (by applying law of large numbers).



(a) 2-Dimension (b) 4-Dimension (c) 40,000-Dimension

Figure 7: Relation between an origin-centered unit hypercube and unit sphere in d-dimension space

Unit Hypercube vs. Unit Sphere

Consider a unit hypercube and a unit sphere both centered at the origin. In 2 dimensional space, the cube is completely contained within the sphere (Figure 7a); in 4 dimensional space, the vertices of the hypercube lie on the surface of the sphere, because the distance between a vertex and the origin is $\sqrt{\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}}=1$; in higher dimensions, the vertices are outside the sphere, as shown by Figure 7a. Actually, vertices of the hypercube are far beyond the sphere when d is large. For example, when $d=40,000$ (Figure 7c), the distance between a vertex and the origin is $\sqrt{\frac{40000}{4}}=100$.

Volume of a Sphere

Next, we will find an analytical expression for the volume $V(d)$ of a sphere in d-dimensional space.

$$\begin{aligned} V(d) &= \int_{x_1=-1}^{1}\int_{x_2=-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}}\int\cdots\int \mathrm{d}x_d\cdots\mathrm{d}x_1 \\ &= \int_{\Omega}\int_{r=0}^{1} r^{d-1}\mathrm{d}r\mathrm{d}\Omega \\ &= \int_{\Omega}\mathrm{d}\Omega\int_{r=0}^{1} r^{d-1}\mathrm{d}r \\ &= \frac{r^d}{d}|_{r=0}^{1}\int_{\Omega}\mathrm{d}\Omega \;=\; \frac{1}{d}\int_{\Omega}\mathrm{d}\Omega \end{aligned}$$

Since we know that

$$\begin{aligned} I(d) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-(x_1^2+x_2^2+\cdots+x_d^2)}\mathrm{d}x_d\cdots\mathrm{d}x_1 \\ &= \left[\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x\right]^d \\ &= (\sqrt{\pi})^d \;=\; \pi^{\frac{d}{2}} \end{aligned}$$

And in polar coordinate system, $I(d)$ can be written as

$$\begin{aligned} I(d) &= \int_{\Omega} \mathrm{d}\Omega \int_{r=0}^{\infty} e^{-r^2} r^{d-1} \mathrm{d}r \quad (\text{let } x = r^2) \\ &= \int_{\Omega} \mathrm{d}\Omega \int_{x=0}^{\infty} e^{-x} x^{\frac{d-1}{2}} \frac{\mathrm{d}x}{2\sqrt{(}x)} \\ &= \int_{\Omega} \mathrm{d}\Omega \frac{1}{2} \int_{x=0}^{\infty} e^{-x} x^{\frac{d}{2}-1} \mathrm{d}x \quad (\text{let } A(d) = \int_{\Omega} \mathrm{d}\Omega) \\ &= \frac{A(d)}{2} \Gamma(\frac{d}{2}) \end{aligned}$$

$\Gamma(\frac{d}{2})$ can be calculated using equations $\Gamma(x) = (x-1)\Gamma(x-1), \Gamma(\frac{1}{2}) = \sqrt{(}\pi)$. Therefore,

$$A(d) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$$

And

$$V(d) = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$$

From the above expression we know that $\lim_{d\to\infty} V(d) = 0$, which means when dimensions go to infinity, the volume of a unit sphere goes to zero.

Exercise What is the trend of the value of $V(d)$ as d grows?

Exercise Instead of a unit sphere, provide a radius of r such that the value of the sphere stays constant for each dimension d.

Generating Points on a Sphere

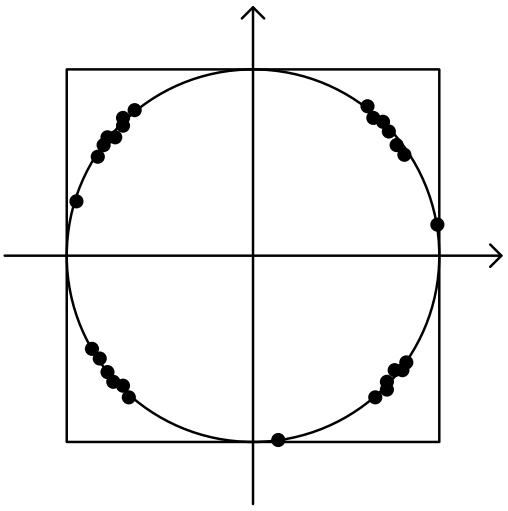


Figure 8: Projecting uniformly distributed points in a hypercube onto sphere surface

Consider the task of generating points uniformly distributed on the surface of a high-dimensional sphere centered at the origin. One attempt might be to generate each coordinate independently and then normalize the vector. This would be equivalent with generating points inside a hypercube, and then project them onto the sphere. However, this method would not work because this way the points would form clusters concentrated in the direction of the vertices of the hypercube, as illustrated in Figure 8.

To avoid this, one might try to generate coordinates using the first method, and keep only the points inside the sphere. However, there would be almost no such points, because we know that the volume of the sphere is close to 0.

A way to generates such points efficiently is to sample each coordinate from a standard Gaussian distribution $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and normalize the vector. This way the probability of a given vector is $p(x_1, x_2, \cdots, x_d) = \alpha e^{-\frac{x_1^2+x_2^2+\cdots+x_d^2}{2}} = \alpha e^{-\frac{1}{2}r^2}$ (where α is a normalizing factor), and therefore spherically symmetric.

Distance Between Two Random Points on a Sphere

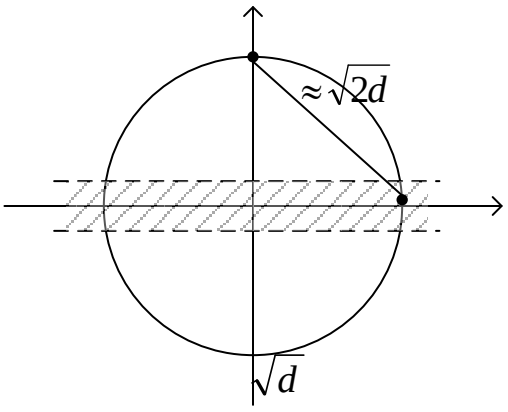


Figure 9: Distance between two random points on sphere surface

Next, we find the distance between two points uniformly distributed on the surface of a radius $\sqrt{d}$ sphere. We position the first point x along the axis 1 by rotating the coordinate system, $x = (\sqrt{d}, 0, \cdots, x_d)$. Let the second point be $y = (y_1, y_2, \cdots, y_d)$. The distance between x and y is:

$$\begin{aligned} d(x,y) &= \sqrt{(\sqrt{d}-y_1)^2 + y_2^2 + \cdots + y_d^2} = \sqrt{d - 2y_1\sqrt{d} + y_1^2 + y_2^2 + \cdots + y_d^2} \\ &= \sqrt{d - 2y_1\sqrt{d} + d} = \sqrt{2d - 2y_1\sqrt{d}} = \sqrt{2d}\sqrt{1 - \frac{2y_1}{\sqrt{2d}}} \end{aligned}$$

Given that y_1 is sampled from a standard Gaussian, it almost always falls in the interval $(-3, 3)$. Therefore, as shown in Figure 9 $\frac{2y_1}{\sqrt{2d}} \approx 0$ and $d(x,y) \approx \sqrt{2d}$.

References

[1] A. Altman, M. Tennenholtz. Ranking Systems: The PageRank Axioms. In Proceedings of the 6th ACM Conference on Electronic Commerce (EC '05).