CS4850 - Mathematical Foundations For The Information Age
Spring 2010
Lecture 1 (1/25/10)

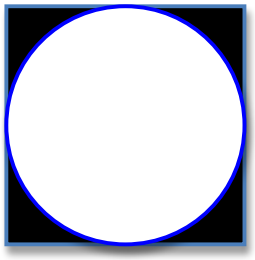
In the last 30 years, the field of theoretical computer science has been devoted to making computers useful through the creation of programming languages, databases, compilers, and various efficient algorithms.
Moving forward, there will be a shift from the discrete mathematics required in the past to probability and statistics that will allow computers to process datasets that are too large for algorithms that run as fast as linear-time.

Common problem of the information age: High dimensional data

We can represent an arbitrary text document in 25000 dimensions (one dimension for each word in the English language).
Given two documents D_1 and D_2, we can represent them as vectors in 25000-space v_1 and v_2 respectively. To find the similarity between the two documents, we can perform the calculation $\text{norm}(v_1) \bullet \text{norm}(v_2)$ and compare the result to 1.

As we explore high dimensions, we will find that they differ fundamentally from the lower dimensions for which we have formed our intuitions.
For example, given a set of randomly generated points in 2d, we can compare these points pairwise and find a set of distances which will range from points that are practically touching to points that are very far from each other.
In high dimensions (such as d=100), these randomly generated points will tend to have pairwise distances that are essentially all the same because the law of large numbers dictates that $\sum_{i=1}^{d}(x_i - y_i)^2$ will be concentrated about an expected value for all points x and y since x and y had their components generated uniformly at random.

Consider a unit sphere in high dimensions.
Imagine we want to generate random points on a unit sphere in d dimensions.
Let's first consider the case where d=2:
To generate the points, we could use the naïve approach of generating a random point in the bounding cube and projecting it onto the sphere. This isn't a good approach because there will be higher concentrations of points on the surface of the sphere that is the boundary of the shaded regions:



A better approach in low dimensions is to generate points in the bounding cube discarding any points that fall outside. This ceases to work for high dimensions because we would

end up throwing out all of the points as the shaded regions become larger relative to the volume of the sphere. This is because the volume of the sphere shrinks as d increases:

Define V(d)= volume of a unit sphere in d dimensions
This is a counterintuitive phenomenon: $\lim_{d\to\infty} V(d) = 0$

In Cartesian coordinates,

$$V(d) = \int_{-1}^{1}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \ldots \int dx_1 \ldots dx_d$$

The bounds here are very difficult to calculate…let's take a look at Polar coordinates:

$$V(d) = \int_{\theta_1}\int_{\theta_2}\ldots\int_{\theta_d}\int_{r=0}^{1} r^{d-1} dr d\theta_1 \ldots d\theta_d = \int_{s^d}\int_{r=0}^{1} r^{d-1} dr d\Omega$$

Note that we have abbreviated the integral over the solid angle. These variables are separable:

$$V(d) = \int_{s^d} d\Omega \int_0^1 r^{d-1} dr = \frac{1}{d}\int_{s^d} d\Omega$$

For now, we don't have an easy way to solve for the area of the solid angle, so we should save it for later:

$$A(d) = \int_{s^d} d\Omega$$

We need to use some creativity to solve for A(d).
Consider this more easily solved integral:

$$I(d) = \int_{x_1=-\infty}^{\infty}\ldots\int_{x_d=-\infty}^{\infty} e^{\frac{-x_1^2+x_2^2+\ldots+x_d^2}{2}} dx_1 \ldots dx_d = \left[\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx\right] = \left(\sqrt{\pi}\right)^d = \pi^{d/2}$$

In polar coordinates:

$$I(d) = \int_{s^d} d\Omega \int_{r=0}^{\infty} e^{\frac{-r^2}{2}} r^{d-1} dr = A(d)\int_{r=0}^{\infty} e^{\frac{-t}{2}} t^{\frac{d-1}{2}} \frac{1}{2\sqrt{t}} dt = \frac{1}{2} A(d)\Gamma\left(\frac{d}{2}\right)$$

Note that $\Gamma(x) = (x-1)\Gamma(x-1), \Gamma(2) = \Gamma(1) = 1, \Gamma(1/2) = \sqrt{\pi}$
(The gamma function is essentially an extension of factorial for real numbers)
Thus,

$$A(d) = \frac{2\pi^{d/2}}{\Gamma\left(d/2\right)} \text{ and } V(d) = \frac{2\pi^{d/2}}{d\Gamma\left(d/2\right)}$$

Now, let's examine our claim:

$$\lim_{d\to\infty} V(d) = \lim_{d\to\infty} \frac{2\pi^{d/2}}{d\Gamma\left(d/2\right)} = 0$$

This is true because the denominator grows factorially and the numerator grows exponentially.

To sum up what we've learned:

- High dimensions are important to the future of computing.
- Our intuition regarding these high dimensional spaces is probably wrong.

- We need a better way to generate random points on a high-dimensional sphere.
- To calculate the volume of a hypersphere, we used both Cartesian and polar coordinate systems since there were aspects of the equations that were easier to integrate in both.