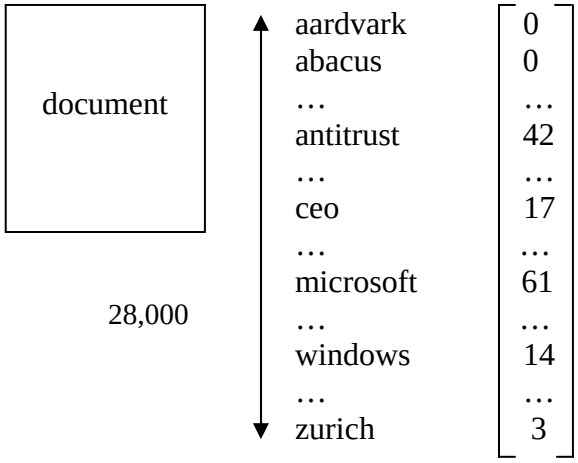
Jeff Davidson (jpd236) and Liaoruo Wang (lw335)
CS 4850 Scribe Notes
01/19/2009

CS 4850 – Math Foundations for Information Age
Professor John Hopcroft – jeh@cs.cornell.edu

High-dimensional data

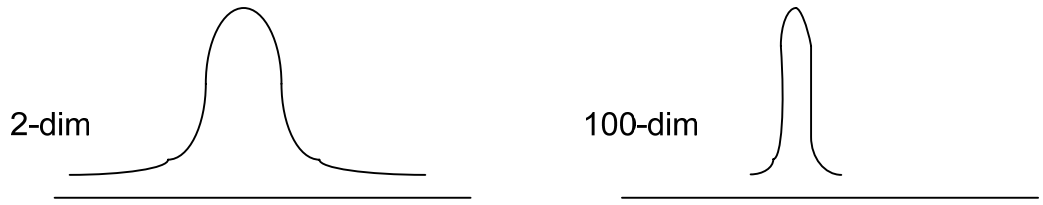


Vector w/ number of occurrences – can normalize and take dot product for two documents. Product is a number between 0 and 1:

- Closer to 0 – orthogonal
- Closer to 1 – similar

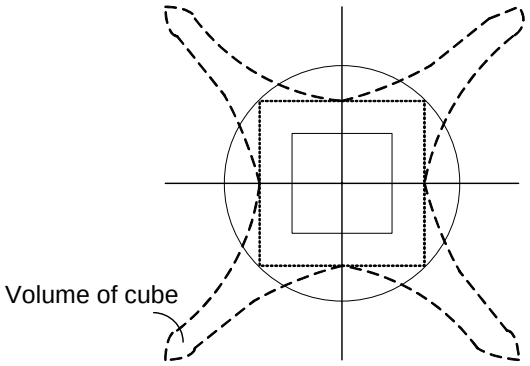
Distance between two points

- In a unit sphere of higher dimensional space *d*, the average distance between two random points is $\sqrt{d}$.
- Dist(x_1,x_2) = $\sqrt{\sum_{i=1}^{d}(x_{1i} - x_{2i})^2}$
  – since x_1 and x_2 are random, what's inside the square root is random.
- Law of large numbers – expect cluster @ expected value



- Volume of higher dimensional objects:

Object	Volume	Max Distance Between Points
Unit "Cube"	Always 1	$\sqrt{d}$
Unit "Sphere"	Approaches 0	2



$$\sqrt{\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2+\cdots+\left(\frac{1}{2}\right)^2}=\sqrt{d}/2$$

- Calculating the volume of a sphere of unit radius: difficult with Cartesian coordinates:

$$V(d)=\int_{-1}^{1}\int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}}\ldots\ldots\int_{-\sqrt{1-x_1^2-x_2^2-\cdots-x_{d-1}^2}}^{\sqrt{1-x_1^2-x_2^2-\cdots-x_{d-1}^2}}dx_1dx_2\ldots dx_d$$

– convert to polar:

$V(d)=\int_{S_d}\int_{r=0}^{1}r^{d-1}drd\Omega=\int_{S_d}d\Omega\int_{r=0}^{1}r^{d-1}dr=\frac{A(d)}{d}$ where $A(d)$ = surface area

- Need $A(d)$ – can calculate by integrating Gaussian in Cartesian and polar:

$$I(d)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}e^{-\left({x_1}^2+{x_2}^2+\cdots+{x_d}^2\right)}dx_1dx_2\ldots dx_d=\left[\int_{-\infty}^{\infty}e^{-x^2}dx\right]^d=\left(\sqrt{\pi}\right)^d=\pi^{d/2}$$

$$I(d)=\int_{S_d}\int_{r=0}^{\infty}r^{d-1}e^{-r^2}drd\Omega=A(d)\int_{r=0}^{\infty}r^{d-1}e^{-r^2}dr$$

$$=\frac{A(d)}{2}\int_{t=0}^{\infty}t^{d/2-1}e^{-t}dt=\frac{A(d)}{2}\Gamma\left(\frac{d}{2}\right)\qquad\text{let } t=r^2$$

- $\Gamma(d)=(d-1)!$
  $\Gamma(d)=(d-1)\Gamma(d-1)$
  $\Gamma(1)=\Gamma(2)=1$
  $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$

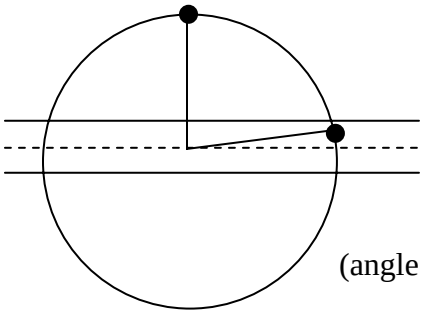
- Comparing two equations above:

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

- $\lim_{d\to\infty}V(d)\sim\lim_{d\to\infty}\frac{3^d}{d!}=0$ – volume goes to 0 in higher dimensions!

Generate points uniformly at random on surface of sphere

- Two dimensions
  - For a box, can generate points by picking each coordinate randomly.
  - If you put a circle inside the sphere, could try normalizing the coordinates by dividing by the radius of the circle – however, this doesn't work, because probability is not uniform (points are more likely to be in corners)
  - So, throw away any random coordinates that are not located within the circle – works in two dimensions.
- Higher dimensions
  - The above approach doesn't work – since a sphere doesn't have any volume, you would throw away all points.
  - So, generate coordinates using a Gaussian distribution, then normalize
  - Pick a point at random on the sphere, then rotate so that this point is located at the pole. The next generated point is likely to be on the "equator" since this is where most of the surface area is. This makes the likely distance between the points $\sqrt{d}$:



(angle is approximately right)