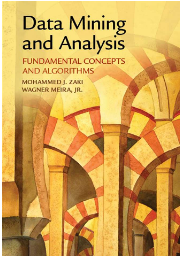
DSTA class 3: High-dimensional data

Slides adapted from from Ch. 6 of M. J. Zaki and W. Meira, CUP, 2012.

http://www.dataminingbook.info/



Download the text from the DSTA class page.

High-dimensional Space

Let **D** be a $n \times d$ data matrix. In data mining typically the data is very high dimensional. Understanding the nature of high-dimensional space, or *hyperspace*, is very important, especially because it does not behave like the more familiar geometry in two or three dimensions.

Hyper-rectangle: The data space is a d-dimensional *hyper-rectangle*

$$R_d = \prod_{j=1}^{d} \Big[\min(X_j), \max(X_j)\Big]$$

where $\min(X_j)$ and $max(X_j)$ specify the range of X_j.

Hypercube: Assume the data is centered, and let m denote the maximum attribute value

$$m = \max_{j=1}^{d} \max_{i=1}^{n} \Big\{ |x_{ij}| \Big\}$$

The data hyperspace can be represented as a *hypercube*, centered at **0**, with all sides of length $l = 2m$, given as

$$H_d(l) = \Big\{ \mathbf{x} = (x_1, x_2, \ldots, x_d)^T \;\Big|\; \forall i,\; x_i \in [-l/2, l/2] \Big\}$$

The *unit hypercube* has all sides of length $l = 1$, and is denoted as $H_d(1)$.

Hypersphere

Assume that the data has been centered, so that $\boldsymbol{\mu} = \mathbf{0}$. Let r denote the largest magnitude among all points:

$$r = \max_i \left\{ \|\mathbf{x}_i\| \right\}$$

The data hyperspace can be represented as a d-dimensional *hyperball* centered at $\mathbf{0}$ with radius r, defined as

$$B_d(r) = \left\{ \mathbf{x} \mid \ \|\mathbf{x}\| \leq r \right\}$$

$$\text{or } B_d(r) = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_d) \;\middle|\; \sum_{j=1}^{d} x_j^2 \leq r^2 \right\}$$

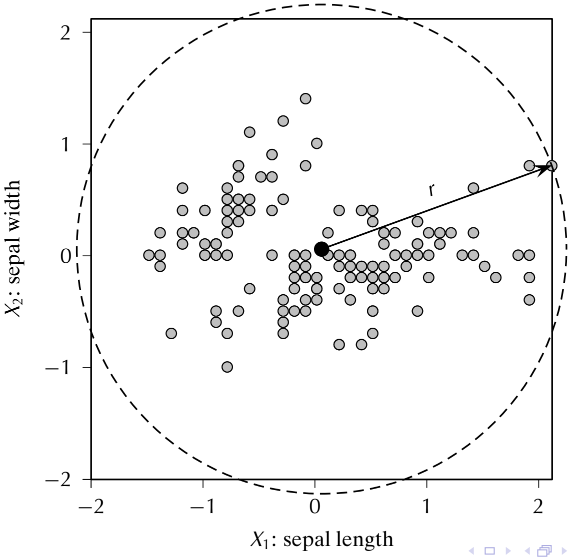
The surface of the hyperball is called a *hypersphere*, and it consists of all the points exactly at distance r from the center of the hyperball

$$S_d(r) = \left\{ \mathbf{x} \mid \ \|\mathbf{x}\| = r \right\}$$

$$\text{or } S_d(r) = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_d) \;\middle|\; \sum_{j=1}^{d} (x_j)^2 = r^2 \right\}$$

Iris Data Hyperspace: Hypercube and Hypersphere

$l = 4.12$ and $r = 2.19$



High-dimensional Volumes

Hypercube: The volume of a hypercube with edge length l is given as

$$\text{vol}(H_d(l)) = l^d$$

HypersphereThe volume of a hyperball and its corresponding hypersphere is identical The volume of a hypersphere is given as

$$\begin{aligned}
\text{In 1 dimension:} \quad & \text{vol}(S_1(r)) = 2r \\
\text{In 2 dimensions:} \quad & \text{vol}(S_2(r)) = \pi r^2 \\
\text{In 3 dimensions:} \quad & \text{vol}(S_3(r)) = \frac{4}{3}\pi r^3 \\
\text{In } d\text{-dimensions:} \quad & \text{vol}(S_d(r)) = K_d r^d = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \right) r^d
\end{aligned}$$

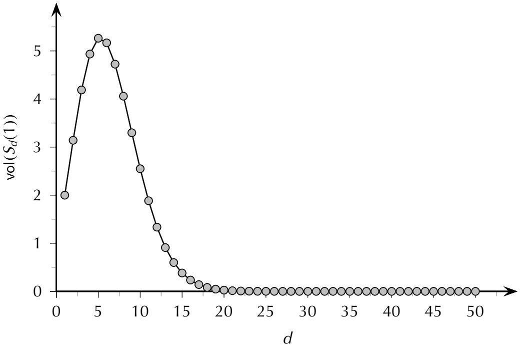
where

$$\Gamma\left(\frac{d}{2}+1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi}\left(\frac{d!!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases}$$

Volume of Unit Hypersphere

With increasing dimensionality the hypersphere volume first increases up to a point, and then starts to decrease, and ultimately vanishes. In particular, for the unit hypersphere with $r = 1$,

$$\lim_{d\to\infty} \mathrm{vol}(S_d(1)) = \lim_{d\to\infty} \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \to 0$$



Hypersphere Inscribed within Hypercube

Consider the space enclosed within the largest hypersphere that can be accommodated within a hypercube (which represents the dataspace).

The ratio of the volume of the hypersphere of radius r to the hypercube with side length $l = 2r$ is given as

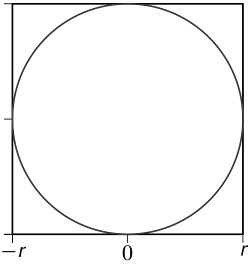
$$\text{In 2 dimensions: } \frac{\text{vol}(S_2(r))}{\text{vol}(H_2(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} = 78.5\%$$

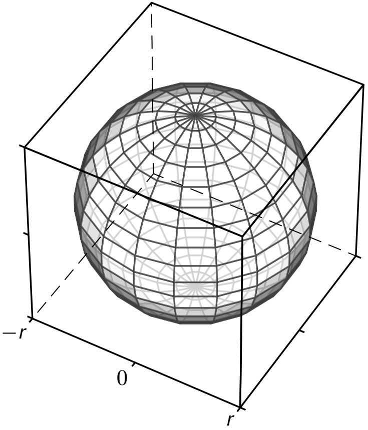
$$\text{In 3 dimensions: } \frac{\text{vol}(S_3(r))}{\text{vol}(H_3(2r))} = \frac{\frac{4}{3}\pi r^3}{8r^3} = \frac{\pi}{6} = 52.4\%$$

$$\text{In } d \text{ dimensions: } \lim_{d\to\infty} \frac{\text{vol}(S_d(r))}{\text{vol}(H_d(2r))} = \lim_{d\to\infty} \frac{\pi^{d/2}}{2^d \Gamma(\frac{d}{2}+1)} \to 0$$

As the dimensionality increases, most of the volume of the hypercube is in the "corners," whereas the center is essentially empty.

Hypersphere Inscribed inside a Hypercube



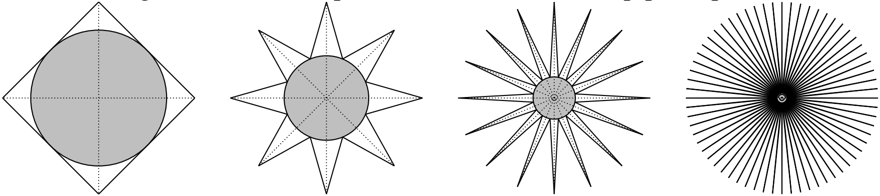


Conceptual View of High-dimensional Space

Two, three, four, and higher dimensions

All the volume of the hyperspace is in the corners, with the center being essentially empty.

High-dimensional space looks like a rolled-up porcupine!

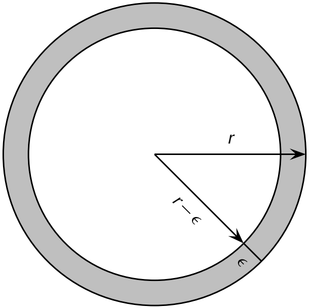


(a) 2D (b) 3D (c) 4D (d) dD

Volume of a Thin Shell

The volume of a thin hypershell of width ϵ is given as

$$\begin{aligned} \text{vol}(S_d(r,\epsilon)) &= \text{vol}(S_d(r)) - \text{vol}(S_d(r-\epsilon)) \\ &= K_d r^d - K_d (r-\epsilon)^d. \end{aligned}$$



The ratio of volume of the thin shell to the volume of the outer sphere:

$$\frac{\text{vol}(S_d(r,\epsilon))}{\text{vol}(S_d(r))} = \frac{K_d r^d - K_d (r-\epsilon)^d}{K_d r^d} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d$$

As d increases, we have

$$\lim_{d\to\infty} \frac{\text{vol}(S_d(r,\epsilon))}{\text{vol}(S_d(r))} = \lim_{d\to\infty} 1 - \left(1 - \frac{\epsilon}{r}\right)^d \to 1$$

Diagonals in Hyperspace

Consider a d-dimensional hypercube, with origin $\mathbf{0}_d = (0_1, 0_2, \dots, 0_d)$, and bounded in each dimension in the range $[-1, 1]$. Each "corner" of the hyperspace is a d-dimensional vector of the form $(\pm 1_1, \pm 1_2, \dots, \pm 1_d)^T$.

Let $\mathbf{e}_i = (0_1, \dots, 1_i, \dots, 0_d)^T$ denote the d-dimensional canonical unit vector in dimension i, and let $\mathbf{1}$ denote the d-dimensional diagonal vector $(1_1, 1_2, \dots, 1_d)^T$.

Consider the angle θ_d between the diagonal vector $\mathbf{1}$ and the first axis $\mathbf{e}_1$, in d dimensions:

$$\cos\theta_d = \frac{\mathbf{e}_1^T \mathbf{1}}{\|\mathbf{e}_1\| \, \|\mathbf{1}\|} = \frac{\mathbf{e}_1^T \mathbf{1}}{\sqrt{\mathbf{e}_1^T \mathbf{e}_1} \sqrt{\mathbf{1}^T \mathbf{1}}} = \frac{1}{\sqrt{1}\sqrt{d}} = \frac{1}{\sqrt{d}}$$

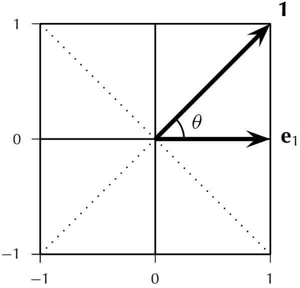
As d increases, we have

$$\lim_{d \to \infty} \cos\theta_d = \lim_{d \to \infty} \frac{1}{\sqrt{d}} \to 0$$

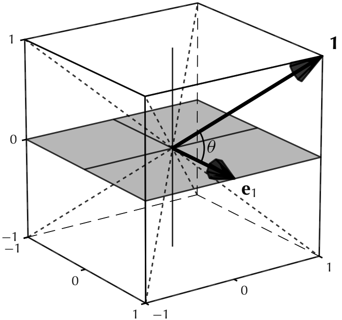
which implies that

$$\lim_{d \to \infty} \theta_d \to \frac{\pi}{2} = 90^\circ$$

Angle between Diagonal Vector $\mathbf{1}$ and $\mathbf{e}_1$



(a) In 2D



(b) In 3D

In high dimensions all of the diagonal vectors are perpendicular (or orthogonal) to all the coordinates axes! Each of the 2^{d-1} new axes connecting pairs of 2^d corners are essentially orthogonal to all of the d principal coordinate axes! Thus, in effect, high-dimensional space has an exponential number of orthogonal "axes."