

# Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki<sup>1</sup>    Wagner Meira Jr.<sup>2</sup>

<sup>1</sup>Department of Computer Science  
Rensselaer Polytechnic Institute, Troy, NY, USA

<sup>2</sup>Department of Computer Science  
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

## Chapter 6: High-dimensional Data

# 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*

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

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

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

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

$$H_d(l) = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_d)^T \mid \forall i, x_i \in [-l/2, l/2] \right\}$$

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:

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

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

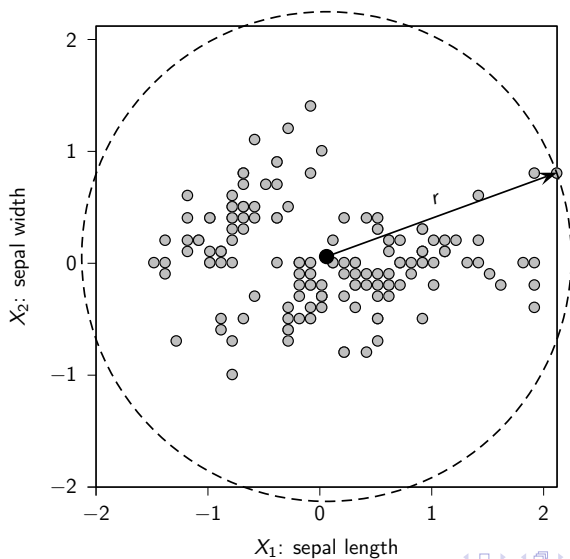
$$B_d(r) = \{ \mathbf{x} \mid \|\mathbf{x}\| \leq r \} \text{ or } B_d(r) = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_d) \mid \sum_{j=1}^d x_j^2 \leq r^2 \right\}$$

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

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

# 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

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

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

$$\text{In 1D: } \text{vol}(S_1(r)) = 2r \quad \text{In 2D: } \text{vol}(S_2(r)) = \pi r^2 \quad \text{In 3D: } \text{vol}(S_3(r)) = \frac{4}{3}\pi r^3$$

In  $d$ -dimensions:

$$\text{vol}(S_d(r)) = K_d r^d = \left( \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} \right) r^d$$

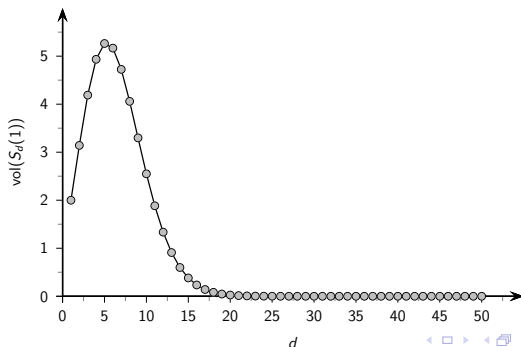
where

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

# 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  $r = 1$ ,

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



# 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

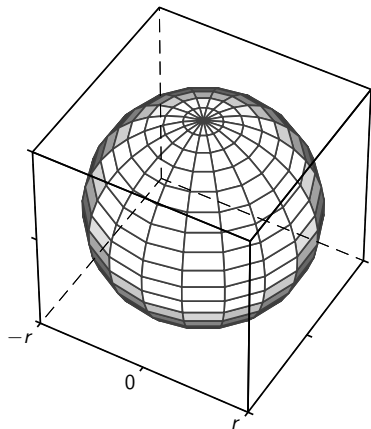
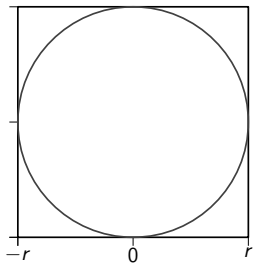
$$\text{In 2 dimensions: } \frac{\text{vol}(S_2(r))}{\text{vol}(H_2(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} = 78.5\%$$

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

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

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



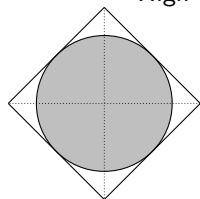


# Conceptual View of High-dimensional Space

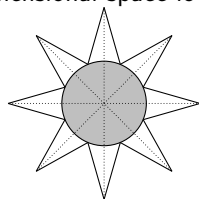
Two, three, four, and higher dimensions

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

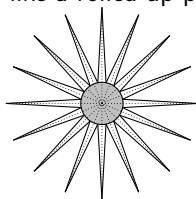
High-dimensional space looks like a rolled-up porcupine!



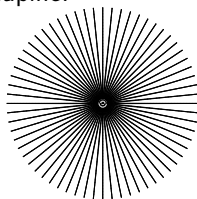
(a) 2D



(b) 3D



(c) 4D



(d)  $dD$

# Volume of a Thin Shell

The volume of a thin hypershell of width  $\epsilon$  is given as

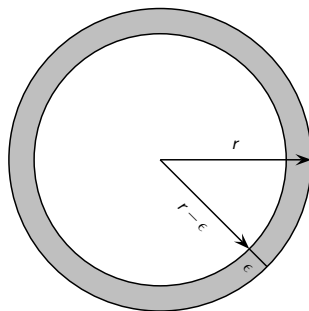
$$\begin{aligned}\text{vol}(S_d(r, \epsilon)) &= \text{vol}(S_d(r)) - \text{vol}(S_d(r - \epsilon)) \\ &= K_d r^d - K_d (r - \epsilon)^d.\end{aligned}$$

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

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



# Diagonals in Hyperspace

Consider a  $d$ -dimensional hypercube, with origin  $0_d = (0_1, 0_2, \dots, 0_d)$ , 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  $\theta_d$  between the diagonal vector  $\mathbf{1}$  and the first axis  $\mathbf{e}_1$ , in  $d$  dimensions:

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

# Diagonals in Hyperspace

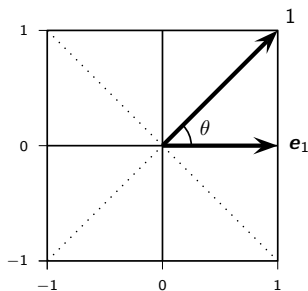
As  $d$  increases, we have

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

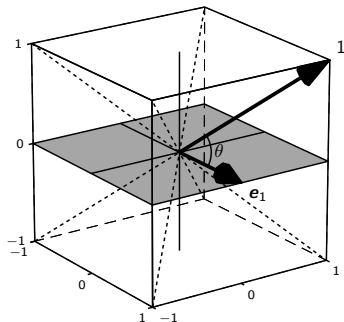
which implies that

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

# Angle between Diagonal Vector 1 and $e_1$



(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.”

# Density of the Multivariate Normal

Consider the standard multivariate normal distribution with  $\mu = 0$ , and  $\Sigma = I$

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^d} \exp\left\{-\frac{\mathbf{x}^T \mathbf{x}}{2}\right\}$$

The peak of the density is at the mean. Consider the set of points  $\mathbf{x}$  with density at least  $\alpha$  fraction of the density at the mean

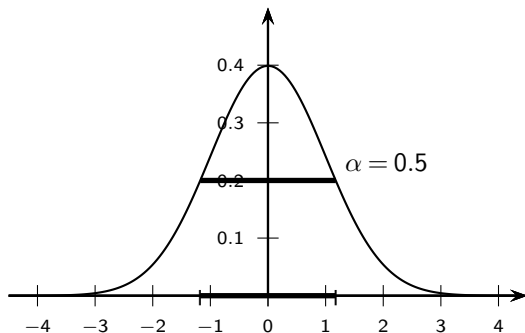
$$\begin{aligned}\frac{f(\mathbf{x})}{f(0)} &\geq \alpha \\ \exp\left\{-\frac{\mathbf{x}^T \mathbf{x}}{2}\right\} &\geq \alpha \\ \mathbf{x}^T \mathbf{x} &\leq -2\ln(\alpha) \\ \sum_{i=1}^d (x_i)^2 &\leq -2\ln(\alpha)\end{aligned}$$

The sum of squared IID random variables follows a chi-squared distribution  $\chi_d^2$ . Thus,

$$P\left(\frac{f(\mathbf{x})}{f(0)} \geq \alpha\right) = F_{\chi_d^2}(-2\ln(\alpha))$$

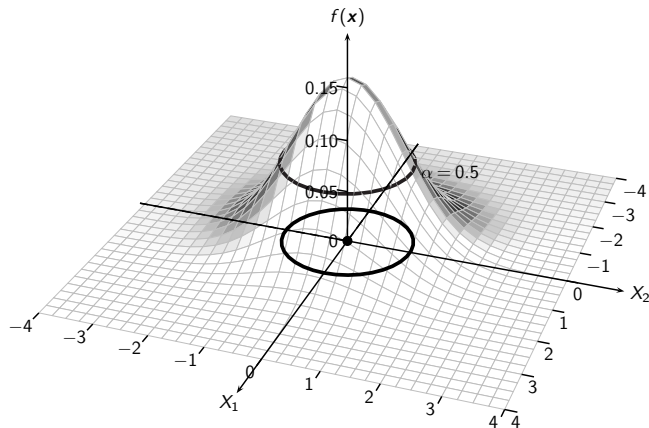
# Density Contour for $\alpha$ Fraction of the Density at the Mean: One Dimension

Let  $\alpha = 0.5$ , then  $-2\ln(0.5) = 1.386$  and  $F_{\chi_1^2}(1.386) = 0.76$ . Thus, 24% of the density is in the tail regions.



# Density Contour for $\alpha$ Fraction of the Density at the Mean: Two Dimensions

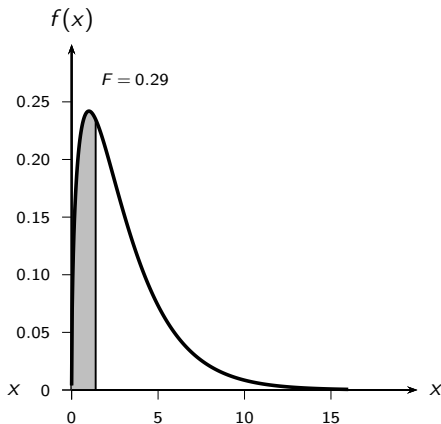
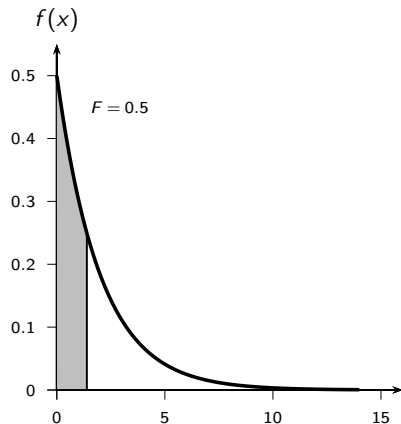
Let  $\alpha = 0.5$ , then  $-2\ln(0.5) = 1.386$  and  $F_{\chi_2^2}(1.386) = 0.50$ . Thus, 50% of the density is in the tail regions.



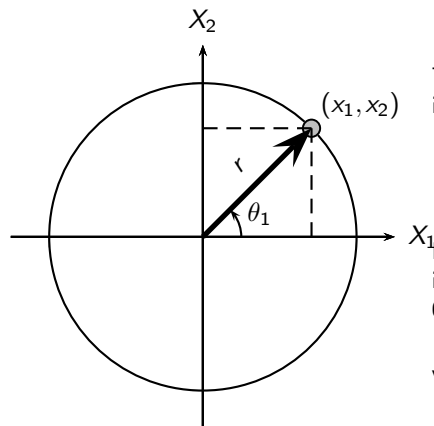


# Chi-Squared Distribution: $P(f(x)/f(0) \geq \alpha)$

This probability decreases rapidly with dimensionality. For 2D, it is 0.5. For 3D it is 0.29, ie., 71% of the density is in the tails. By  $d = 10$ , it decreases to 0.075%, that is, 99.925% of the points lie in the extreme or tail regions.



# Hypersphere Volume: Polar Coordinates in 2D



The point  $\mathbf{x} = (x_1, x_2)$  in polar coordinates

$$x_1 = r \cos \theta_1 = r c_1$$

$$x_2 = r \sin \theta_1 = r s_1$$

where  $r = \|\mathbf{x}\|$ , and  $\cos \theta_1 = c_1$  and  $\sin \theta_1 = s_1$ .

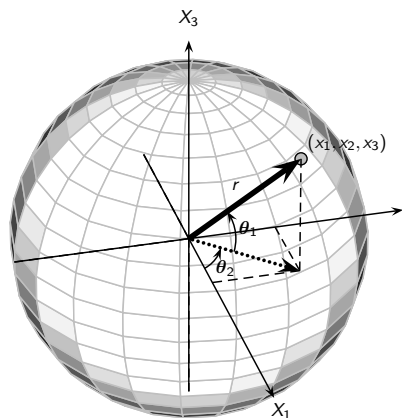
The *Jacobian matrix* for this transformation is given as

$$J(\theta_1) = \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta_1} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta_1} \end{pmatrix} = \begin{pmatrix} c_1 & -r s_1 \\ s_1 & r c_1 \end{pmatrix}$$

Hypersphere volume is obtained by integration over  $r$  and  $\theta_1$  (with  $r > 0$ , and  $0 \leq \theta_1 \leq 2\pi$ ):

$$\begin{aligned} \text{vol}(S_2(r)) &= \int_r \int_{\theta_1} |\det(J(\theta_1))| dr d\theta_1 \\ &= \int_0^r \int_0^{2\pi} r dr d\theta_1 = \int_0^r r dr \int_0^{2\pi} d\theta_1 \\ &= \frac{r^2}{2} \Big|_0^r \cdot \theta_1 \Big|_0^{2\pi} = \pi r^2 \end{aligned}$$

# Hypersphere Volume: Polar Coordinates in 3D



$\mathbf{x} = (x_1, x_2, x_3)$  in polar coordinates

$$x_1 = r \cos \theta_1 \cos \theta_2 = r c_1 c_2$$

$$x_2 = r \cos \theta_1 \sin \theta_2 = r c_1 s_2$$

$$x_3 = r \sin \theta_1 = r s_1$$

The Jacobian matrix is given as

$$J(\theta_1, \theta_2) = \begin{pmatrix} c_1 c_2 & -r s_1 c_2 & -r c_1 s_2 \\ c_1 s_2 & -r s_1 s_2 & r c_1 c_2 \\ s_1 & r c_1 & 0 \end{pmatrix}$$

The volume of the hypersphere for  $d = 3$  is obtained via a triple integral with  $r > 0$ ,  $-\pi/2 \leq \theta_1 \leq \pi/2$ , and  $0 \leq \theta_2 \leq 2\pi$

$$\begin{aligned} \text{vol}(S_3(r)) &= \int_r \int_{\theta_1} \int_{\theta_2} \left| \det(J(\theta_1, \theta_2)) \right| dr d\theta_1 d\theta_2 \\ &= \frac{4}{3} \pi r^3 \end{aligned}$$

# Hypersphere Volume in $d$ Dimensions

The determinant of the  $d$ -dimensional Jacobian matrix is

$$\det(J(\theta_1, \theta_2, \dots, \theta_{d-1})) = (-1)^d r^{d-1} c_1^{d-2} c_2^{d-3} \dots c_{d-2}$$

The volume of the hypersphere is given by the  $d$ -dimensional integral with  $r > 0$ ,  $-\pi/2 \leq \theta_i \leq \pi/2$  for all  $i = 1, \dots, d-2$ , and  $0 \leq \theta_{d-1} \leq 2\pi$ :

$$\begin{aligned} \text{vol}(S_d(r)) &= \int_r \int_{\theta_1} \int_{\theta_2} \dots \int_{\theta_{d-1}} \left| \det(J(\theta_1, \theta_2, \dots, \theta_{d-1})) \right| dr d\theta_1 d\theta_2 \dots d\theta_{d-1} \\ &= \int_0^r r^{d-1} dr \int_{-\pi/2}^{\pi/2} c_1^{d-2} d\theta_1 \dots \int_{-\pi/2}^{\pi/2} c_{d-2} d\theta_{d-2} \int_0^{2\pi} d\theta_{d-1} \\ &= \frac{r^d}{d} \frac{\Gamma(\frac{d-1}{2}) \Gamma(\frac{1}{2})}{\Gamma(\frac{d}{2})} \frac{\Gamma(\frac{d-2}{2}) \Gamma(\frac{1}{2})}{\Gamma(\frac{d-1}{2})} \dots \frac{\Gamma(1) \Gamma(\frac{1}{2})}{\Gamma(\frac{3}{2})} 2\pi \\ &= \frac{\pi \Gamma(\frac{1}{2})^{d/2-1} r^d}{\frac{d}{2} \Gamma(\frac{d}{2})} \\ &= \left( \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} \right) r^d \end{aligned}$$

# Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki<sup>1</sup>    Wagner Meira Jr.<sup>2</sup>

<sup>1</sup>Department of Computer Science  
Rensselaer Polytechnic Institute, Troy, NY, USA

<sup>2</sup>Department of Computer Science  
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

## Chapter 6: High-dimensional Data