数据科学基础

Foundations of Data Science

5.2 高维几何性质

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

距离(度量)

集合 A 上的度量$d : X \times X \rightarrow R$称之为“距离函数”或简称“距离”，如果对于A内任意的 x、y、z，均满足如下条件:

- $d(x, y) \geq 0$ (非负性)，$d(x, y) = 0$ 当且仅当 $x = y$
- $d(x, y) = d(y, x)$ (对称性)
- $d(x, z) \leq d(x, y) + d(y, z)$ (三角不等式)

范数

映射 $\|\cdot\|: R^n \to R^+ \cup \{0\}$ 称之为范数:

①非负性 $\|X\| \geq 0$, 且 $\|X\| = 0 \Leftrightarrow X = 0$

②齐次性 $\forall a \in R, \|aX\| = |a| \cdot \|X\|$

③三角不等式 $\|X + Y\| \leq \|X\| + \|Y\|$

向量空间中，我们通过范数诱导度量定义两点的距离: $\|X - Y\|$

p范数

常用p范数定义如下:$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$

- 1范数定义如下:$\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- 2范数定义如下:$\|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2}$
- ∞范数定义如下:$\|x\|_\infty = \max_i(|x_i|)$

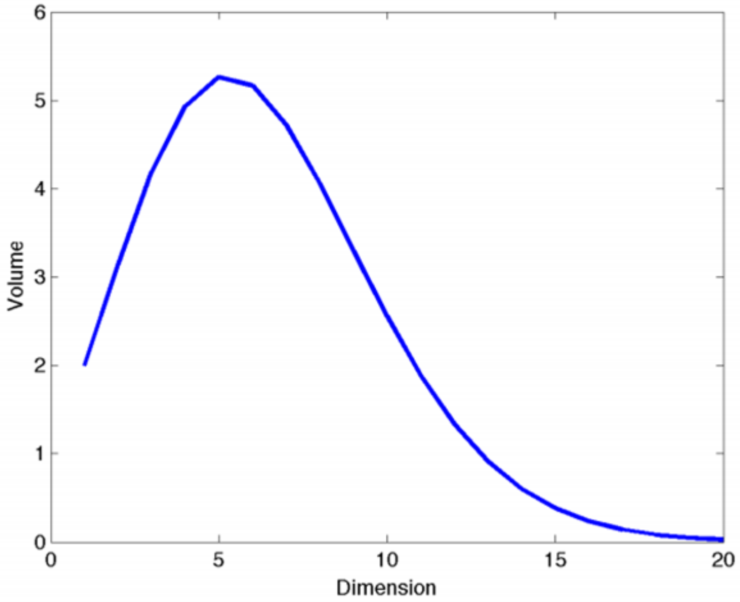
单位球体

【定理】d维单位球体的体积为：$V(d) = \dfrac{\pi^{\frac{d}{2}}}{\frac{d}{2}\,\Gamma\left(\frac{d}{2}\right)}$

其中Γ是Gamma函数，使用斯特林公式，

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}}\left(\frac{n}{e}\right)^n$$

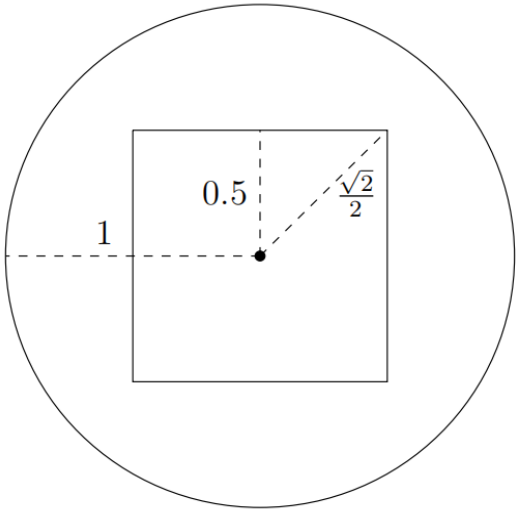
$$V(d) \to 0 \quad \text{as} \quad d \to \infty.$$

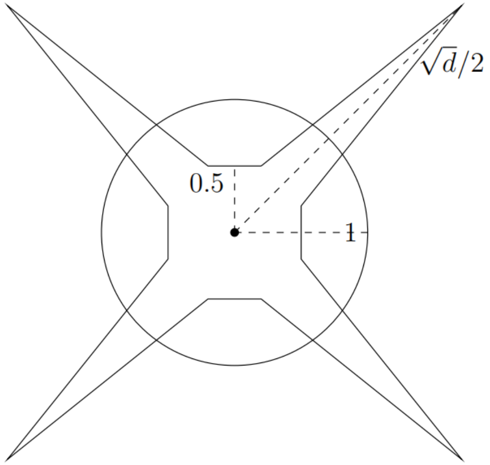


Dimension	Volume of a ball of radius R
0	1
1	$2R$
2	$\pi R^2 \approx 3.142 \times R^2$
3	$\frac{4\pi}{3} R^3 \approx 4.189 \times R^3$
4	$\frac{\pi^2}{2} R^4 \approx 4.935 \times R^4$
5	$\frac{8\pi^2}{15} R^5 \approx 5.264 \times R^5$
6	$\frac{\pi^3}{6} R^6 \approx 5.168 \times R^6$
7	$\frac{16\pi^3}{105} R^7 \approx 4.725 \times R^7$
8	$\frac{\pi^4}{24} R^8 \approx 4.059 \times R^8$
9	$\frac{32\pi^4}{945} R^9 \approx 3.299 \times R^9$
10	$\frac{\pi^5}{120} R^{10} \approx 2.550 \times R^{10}$
11	$\frac{64\pi^5}{10395} R^{11} \approx 1.884 \times R^{11}$
12	$\frac{\pi^6}{720} R^{12} \approx 1.335 \times R^{12}$

单位球体与立方体

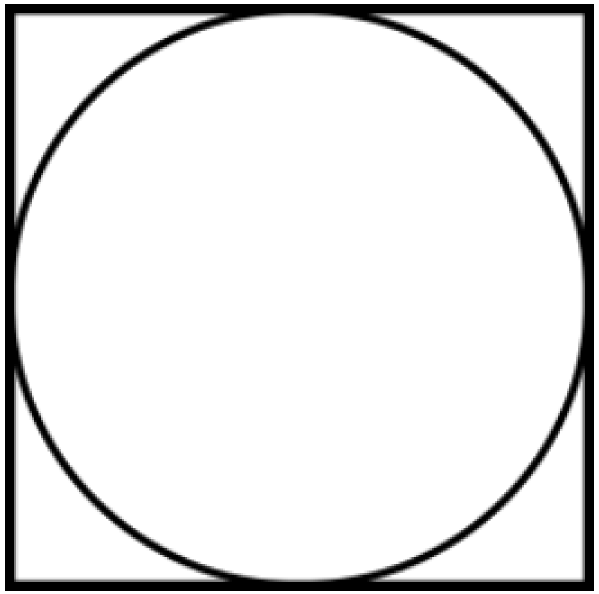
【定理】d维单位立方体的顶点距离为：$\frac{\sqrt{d}}{2}$





高维球体与立方体

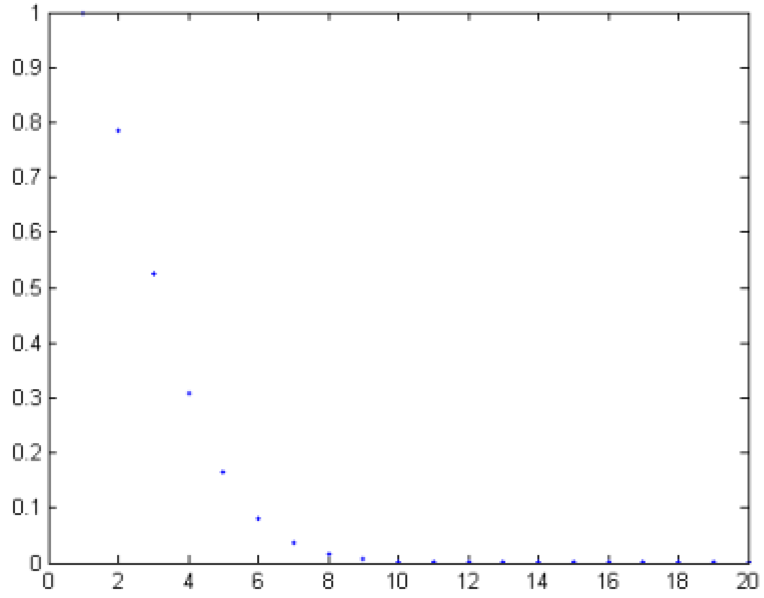
高维立方体的大部分体积位于其角落。



$$\mathrm{area}(B^2/C^2) = \frac{\pi}{4}$$

$$\mathrm{vol}(B^3/C^3) = \frac{\pi}{6}$$

$$\mathrm{vol}(B^{20})/\mathrm{vol}(C^{20}) \approx 2.5 \cdot 10^{-8}$$



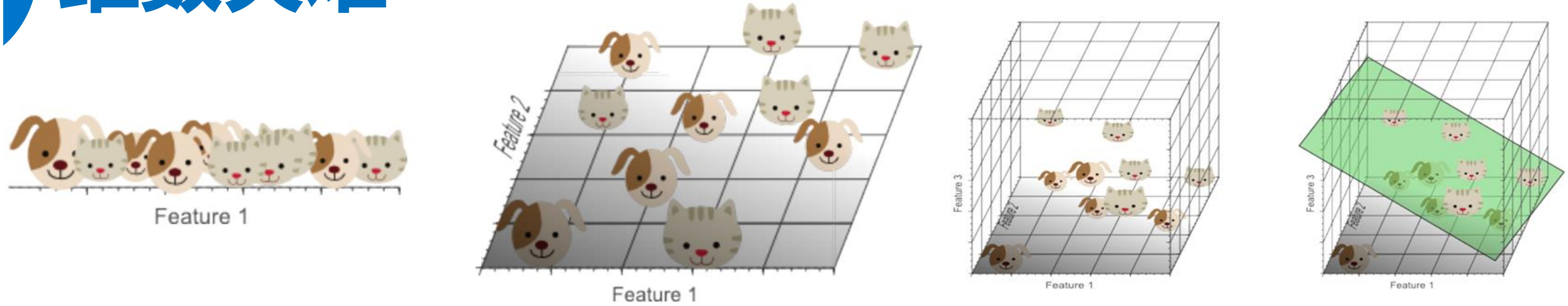
高维环

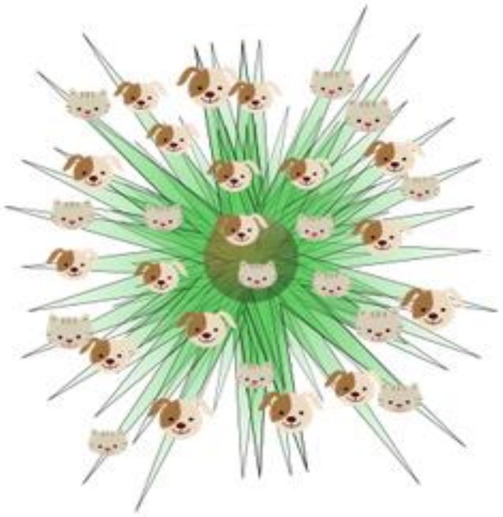
$$\{x \in \mathbb{R}^n : (1-r)^2 \leq x_1^2 + \cdots + x_n^2 \leq 1\}$$

$$\frac{\text{vol Shell}(r)}{\text{vol}(B^n)} = \frac{\text{vol}(B^n) - (1-r)^n \text{vol}(B^n)}{\text{vol}(B^n)} = 1 - (1-r)^n.$$

如果r=0.01 (1/100) 而 n=500, 则超过99% 的体积在百分之一的环上。

维数灾难





- 特征维度增加，更易找到分类的超平面
- 给定训练样本，维度增加容易带来过拟合
- 维度增加，需要的样本数量呈指数增长

Johnson–Lindenstrauss定理

【定理】对给定的 $\varepsilon \in (0,1)$ 以及N维欧氏空间的 m个点 $\{x_1, \ldots, x_m\}$，对于任意满足条件 $n > (\log m)/(\varepsilon^2/2 - \varepsilon^3/3)$的正整数 n，存在一个线性映射 $f: R^N \to R^n$，将这 m 个点，从 R^N (高维空间)中映射到R^n (低维空间)中，同时“基本上”保持了点集成员两两之间的距离，即：对于任意两个点 x_i, x_j: $1 \leq i < j \leq m$，都有

$$(1-\epsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2$$

更进一步地，这个线性映射 f还可以在随机多项式时间内求出。