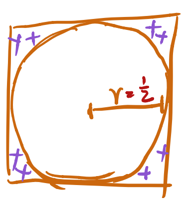
降维

解决:

- 过拟合
  - 增加 data
  - 正则化
  - 降维 →
    - 直接降维 → 特征选择
    - 线性降维 →
      - PCA
      - MDS
    - 非线性降维 → 流形
      - Isomap
      - LLE

- 维度灾难

$\boxed{\text{几何角度}}$



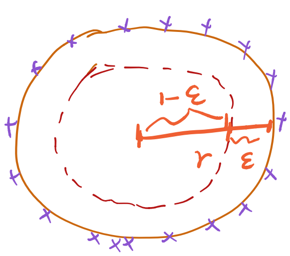
D: 维度

$V_{超立方体} = 1$

$V_{超球体} = K \cdot R^D = K \cdot (0.5)^D \quad D \to \infty$

$\to 0$ 说明里面为空

即 D↑, 数据稀疏化



$V_{外} = K \cdot 1^D = K$

$V_{环形带} = V_{外} - V_{内}$

$= K - K \cdot (1-\varepsilon)^D$

$$\frac{V_{环形带}}{V_{外}} = \frac{K - K \cdot (1-\varepsilon)^D}{K} = 1-(1-\varepsilon)^D$$

$\forall\ 0<\varepsilon<1 \qquad \lim_{D\to\infty}(1-\varepsilon)^D=0 \quad \Rightarrow \lim_{D\to\infty}\frac{V_{环}}{V_{外}}=1$ 一个空壳

维度太高，距离都很大，样本间距离就没有实际意义了

[样本均值，样本方差的矩阵表示]：

Data： $X=(x_1, x_2 \cdots x_N)^T_{N\times p}=\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N\times p}$

$x_i\in \mathbb{R}^{\overset{p维}{p}} \quad i=1,2,3\cdots N \qquad 1_N=\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}_{N\times 1} \qquad H_N=I_N-\frac{1}{N}1_N1_N^T$

[p维]

Sample mean： $\bar{X}_{p\times 1}=\frac{1}{N}\sum_{i=1}^{N}x_i=\frac{1}{N}X^T1_N$

Sample covariance： $S_{p\times p}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T=\frac{1}{N}X^THX$

① $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i=\frac{1}{N}\underbrace{(x_1\ x_2\ \cdots\ x_N)}_{X^T}\underline{\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}_{N\times 1}}=\frac{1}{N}X^T1_N$

② $S=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T$

$=\frac{1}{N}\underbrace{(x_1-\bar{x}\quad x_2-\bar{x}\quad x_3-\bar{x}\ \cdots\ x_N-\bar{x})}\begin{pmatrix}(x_1-\bar{x})^T\\ (x_2-\bar{x})^T\\ \vdots \\ (x_N-\bar{x})^T\end{pmatrix}$

$$(x_1 \ x_2 \ \cdots \ x_N) - (\bar{x} \ \bar{x} \ \cdots \ \bar{x})$$

$$X^T - \bar{x}\underbrace{(1 \ 1 \ 1 \ \cdots \ 1)}_{1_N^T}$$

$$X^T - \frac{1}{N} X^T 1_N \bar{X} 1_N^T$$

$$X^T\left(I_N - \frac{1}{N} 1_N 1_N^T\right)$$

$$= \frac{1}{N} X^T \underbrace{\left(I_N - \frac{1}{N} 1_N 1_N^T\right)}_{H_{N\times N}\text{ 中心矩阵}} \cdot \left(I_N - \frac{1}{N} 1_N 1_N^T\right)^T X$$

$$= \frac{1}{N} X^T H \cdot H^T X$$

$$= \frac{1}{N} X^T H X$$

$$\begin{cases} H = I_N - \frac{1}{N} 1_N 1_N^T \\ H^T = \left(I_N - \frac{1}{N} 1_N 1_N^T\right)^T = I_N - \frac{1}{N} 1_N 1_N^T = H \end{cases}$$

$$H^2 = H \cdot H = \left(I_N - \frac{1}{N} 1_N 1_N^T\right)\left(I_N - \frac{1}{N} 1_N 1_N^T\right)$$

$$= I_N - \frac{2}{N} 1_N 1_N^T + \frac{1}{N^2} 1_N \underbrace{1_N^T 1_N}_{N} 1_N^T$$

$$\frac{1}{N}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}_{N\times N}$$

$$= I_N - \frac{1}{N} 1_N 1_N^T = H$$

$$\Rightarrow H^n = H$$

[主成分分析(PCA) - 最大投影方差角度]:

data, Mean, Covariance ... 同上

一个中心: 原始特征空间的重构
使相关 → 无关 (正交基)

两个基本点: 最大投影方差
最小重构距离

// Todo