MACHINE LEARNING

机器学习

第十章

降维与度量学习

大纲

- k近邻学习
- 降维：低维嵌入 — 多维缩放
- 主成分分析
- 流形学习
- 度量学习

降维

- 样本空间缩小，维度不变
  - k近邻学习
- 样本空间维度降低
  - 低维嵌入（降维）缓解 维数灾难
  - 低维子空间 获取方法
    - 线性降维
      - 主要方法：主成分分析 PCA

降维

- 样本空间缩小，维度不变
  - k近邻学习
- 样本空间维度降低
  - 低维嵌入（降维）缓解 维数灾难
  - 低维子空间 获取方法
    - 线性降维
      - 主要方法：主成分分析 PCA

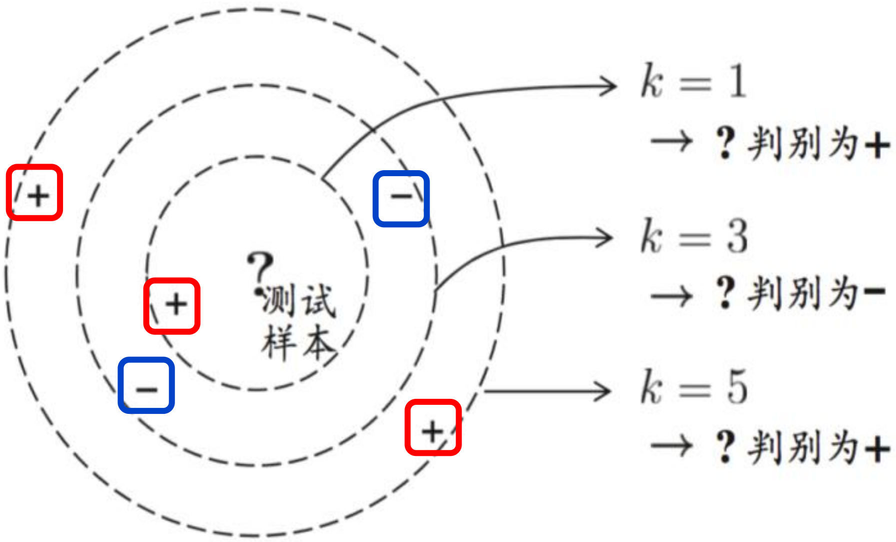
“懒惰学习”与“急切学习”

- **急切学习 (eager learning)**
  - 训练阶段，就对训练样本，进行学习处理。
- **懒惰学习 (lazy learning)**
  - 训练阶段，仅仅是把样本保存起来，训练时间开销为零
  - 待收到测试样本后，再进行处理。
- ***k*-近邻学习 *k*-Nearest Neighbor, *k*-NN**
  - 没有显式的训练过程，懒惰学习著名代表
  - 一种常用的监督学习方法

k 近邻学习 —— 监督学习

k 近邻分类器

- 一种常用的监督学习方法
- k 是一个重要参数：k 取值不同，分类结果 有显著不同。



虚线 ： 等距线
测试样本

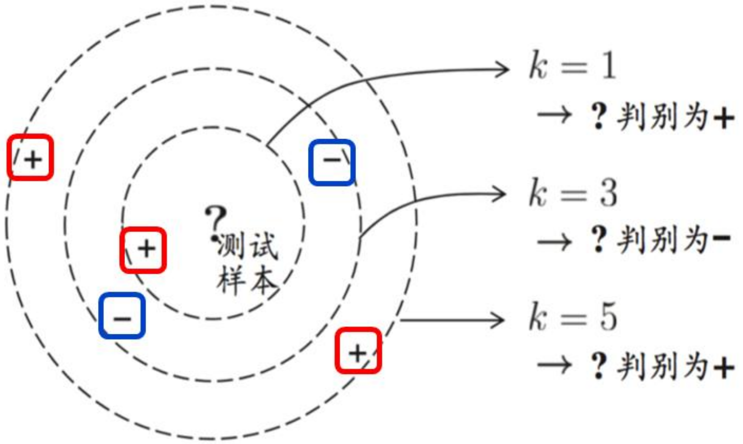
- $k=1$ 或 5 时，被判别为正例
- $k=3$ 时，被判别为反例

- 不同的距离计算方式，找出的“近邻”可能有显著差别
- 从而，导致分类结果 有显著不同。

k 近邻学习 — 监督学习

工作机制

1. 首先，确定训练样本、某种 距离度量。
2. 然后，对于某个给定的测试样本，找到训练集中，距离最近的 k 个样本。



分类 问题

✓ 基于距离获得分类结果

对距离进行加权平均，然后，进行分类

✓ 用“投票法”获得分类结果

选择 k 个样本中出现最多的类别，标记作为预测结果

k 近邻学习 —— 监督学习

工作机制

1. 首先，确定训练样本、某种 距离度量。
2. 然后，对于某个给定的测试样本，找到训练集中，距离最近的 k 个样本。

- **回归** 问题
  - ✓ 加权投票法 获得 预测结果

    用 k 个样本的预测结果，进行加权投票，距离越近的样本，投票权重越大。
  - ✓ 平均法 获得 预测结果

    将 k 个样本的实值输出标记的平均值，作为预测结果。

k 近邻学习 —— 监督学习

工作机制

1. 首先，确定训练样本、某种 距离度量。
2. 然后，对于某个给定的测试样本，找到训练集中，距离最近的 k 个样本。

◆ 回归 问题

✓ 加权投票法 获得 预测结果

✓ 平均法 获得 预测结果

波士顿房价数据集，每条数据包含：房屋以及房屋周围的详细信息。其中，包括城镇犯罪率、一氧化氮浓度、住宅平均房间数、到中心区域的加距离以及自住房平均房价等。

因此，波士顿房价数据集能够应用到回归问题上。

k 近邻学习 — 监督学习

- k 近邻分类器中，k 是一个重要参数

- $k = 1$，最近邻分类器（1-NN）
  - 2分类错误率 $P(err)$
    - 给定测试样本 x，若其最近邻样本为 z，则 出错的概率，即 x 与 z 类别标记不同的概率，即

$$P(err) = 1-\sum P(c|x) \cdot P(c|z)$$

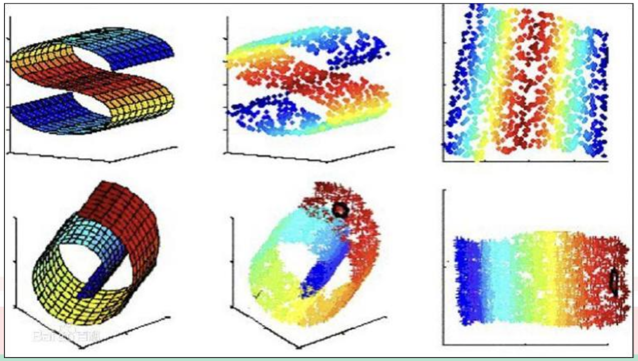
  - 优点
    - 简单
    - 泛化错误率低：不超过 贝叶斯最优分类器 错误率的两倍！

降维

- 样本空间缩小，维度不变
  - k近邻学习
- 样本空间维度降低
  - 低维嵌入（降维）缓解 维数灾难
  - 低维子空间 获取方法
    - 线性降维
      - 主要方法：主成分分析 PCA

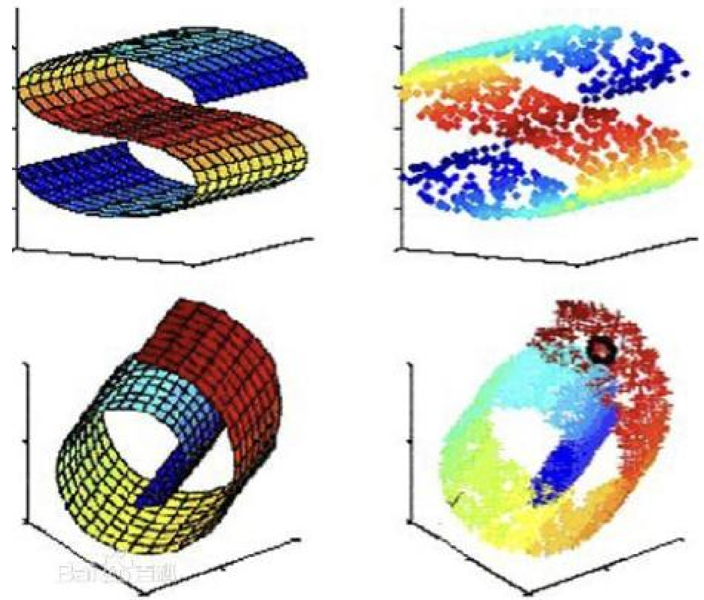
维数灾难

- 上述讨论，基于一个重要的假设：
  - 任意测试样本 x 附近，任意小的 δ 距离范围内，总能找到一个训练样本，即训练样本的采样密度足够大，或称为“**密采样**”
  - **然而**，在现实任务中，**数据样本稀疏**，很难满足密采样
    - 若**属性维数**为1
      - 单位距离内，当**距离** $\delta=10^{-3}$，将 10^3 个**样本**点，平均分布 在 归一化后的属性取值（单位距离）范围内，才可使得：任意测试样本，在其附近 $\delta=10^{-3}$ 距离范围内，总能找到一个训练样本。



维数灾难

- 现实应用中，属性维数经常成千上万，数据样本稀疏，很难满足密采样条件所需的样本数目



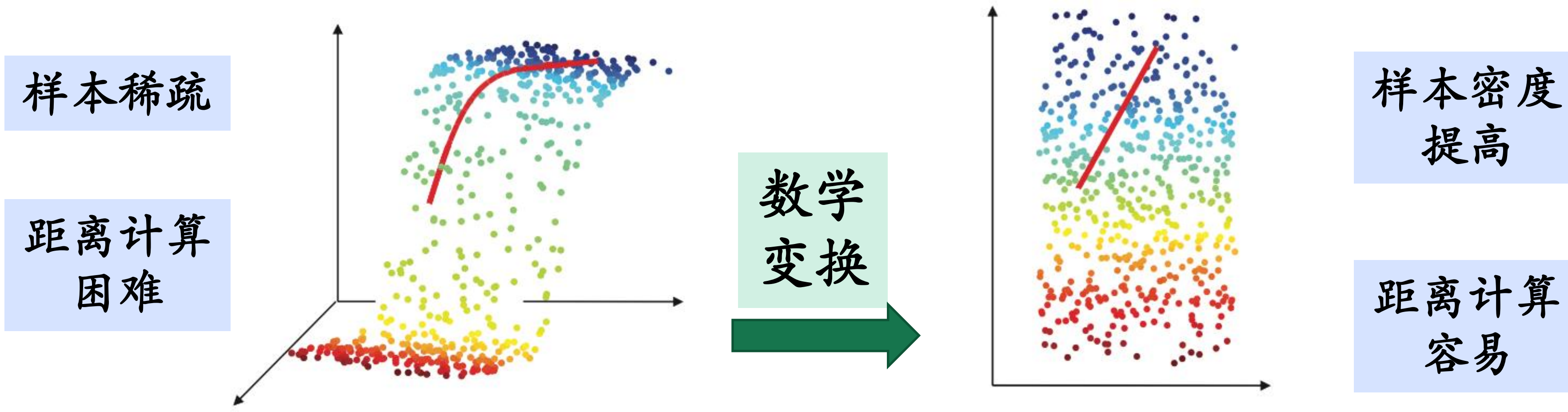
- 高维空间中，距离计算 困难
  - 许多学习方法，都涉及距离计算
  - 高维空间，会给距离计算，带来很大的麻烦。
    例如，当维数很高时，甚至连计算内积，都不再容易。

- 高维情形下，数据样本稀疏、距离计算困难等问题，称为“维数灾难”。
  - 所有机器学习方法，共同面临的严重障碍。

低维嵌入 — 降维

缓解 维数灾难 的一个重要途径是 降维

- 通过数学变换，将原始高维属性空间转变为 一个低维“子空间”



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

低维子空间：样本密度，大幅提高；距离计算，更为容易。

原始高维空间中的样本点，在低维嵌入子空间中，更容易学习

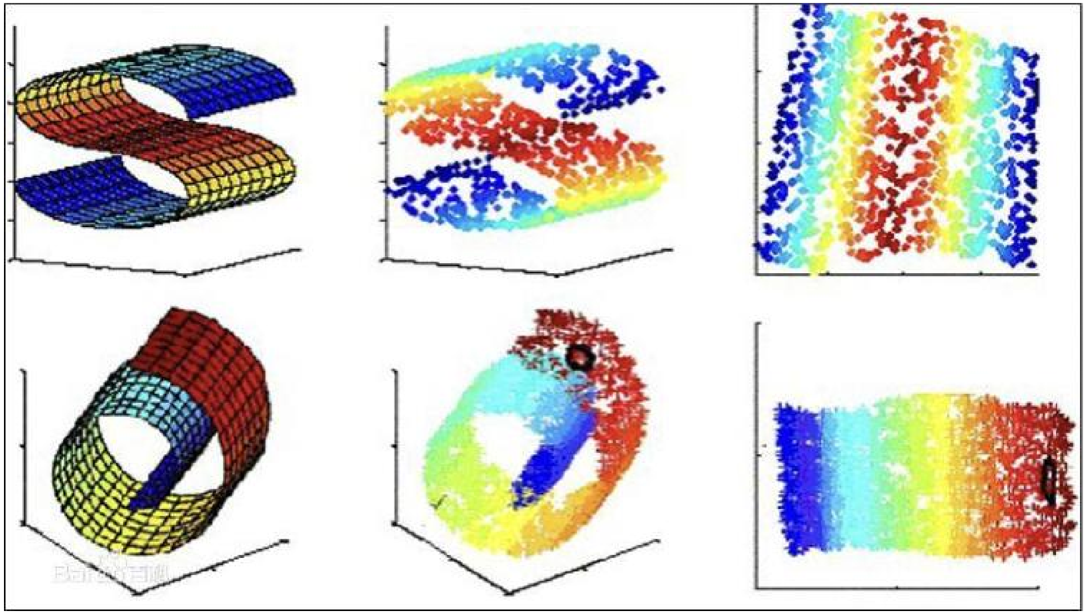
低维嵌入 — 降维

缓解 维数灾难 的一个重要途径是 降维

- 通过数学变换，将原始高维属性空间转变为 一个低维“子空间”

样本稀疏

距离计算困难



样本密度提高

距离计算容易

低维子空间：样本密度，大幅提高；距离计算，更为容易。

原始高维空间中的样本点，在低维嵌入子空间中，更容易学习

低维嵌入 — 降维

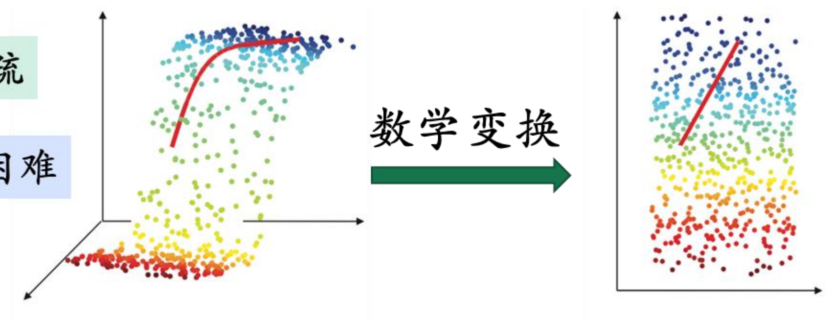
低维嵌入：高维空间通过数学变换转变为 一个低维子空间

- 为什么降维能使样本密度 提高，距离计算 简单?
  - 数据样本，虽然是高维的
  - 但是，与学习任务密切相关的，也许仅是 某个低维分布，即，高维空间中的一个低维“嵌入”
  - 因而，可以对高维数据 进行有效的降维

人脸图像
通过降维，获得仅包含眼睛、鼻子、嘴巴的数据样本

样本稀疏

距离计算:困难



(a) 三维空间中观察到的样本点

(b) 二维空间中的曲面

样本密度提高

距离计算:容易

原始高维空间中的样本点，在低维嵌入子空间中，更容易学习

低维嵌入 — 降维

降维 原则

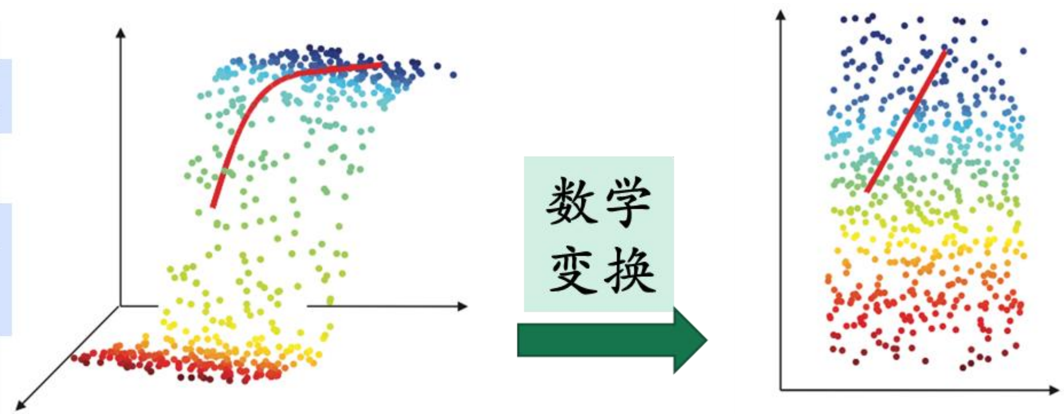
- 低维子空间，能够保持 原始高维空间 数据分布

降维 实现方式

- 原始空间中样本之间的距离，在低维子空间中得以保持，即“多维缩放”(MDS)

样本稀疏

距离计算困难



(a) 三维空间中观察到的样本点

距离不变

(b) 二维空间中的曲面

样本密度提高

距离计算容易

$$dist_{ij} = ||\boldsymbol{x}_i-\boldsymbol{x}_j|| = ||\boldsymbol{z}_i-\boldsymbol{z}_j||$$

低维嵌入 — 降维

原始空间中样本之间的距离，在低维子空间中得以保持不变

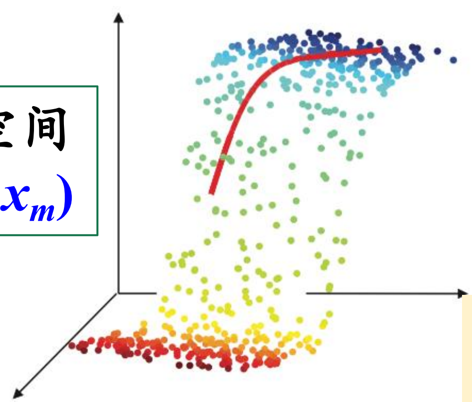
- 样本在d'维低维空间 Z 中的欧氏距离 等于 原始空间X中的距离

✓ d 维原始空间和d' 维低维空间，具有相同的距离矩阵D

数学变换 $\mathbf{Z} = \mathbf{B} \cdot \mathbf{X}$

目标：求内积矩阵B，使得:

d 维原始空间 $\mathbf{X}=(x_1,\ldots, x_m)$



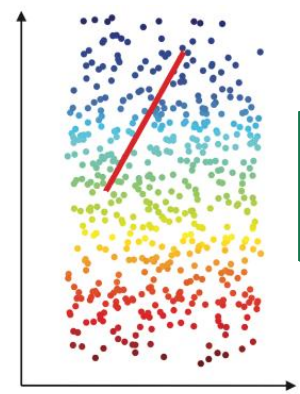
(a) 三维空间中观察到的样本点

数学变换

$\mathbf{Z} = \mathbf{B} \cdot \mathbf{X}$

距离不变

$$D=\begin{pmatrix} dist_{11} & dist_{21} & \cdots & dist_{m1} \\ dist_{12} & dist_{22} & \vdots & dist_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ dist_{1m} & dist_{2m} & \cdots & dist_{mm} \end{pmatrix}$$



(b) 二维空间中的曲面

d' 维低维空间 $\mathbf{Z}=(z_1,\ldots, z_m)$

$$dist_{ij} = ||x_i - x_j|| = ||z_i - z_j||$$

低维嵌入 —— 降维

原始空间中样本之间的距离，在低维子空间中得以保持不变

- 数学变换 $\mathbf{Z} = \mathbf{B} \cdot \mathbf{X}$
- 目标：求内积矩阵B，使得：

令 $\mathbf{B} = \mathbf{Z}^{T} \cdot \mathbf{Z}$

- 对矩阵B 进行特征值分解 $\mathbf{B} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{T}$
  - $\mathbf{V}$：特征向量矩阵
  - $\boldsymbol{\Lambda}$：$\mathbf{V} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$为特征值构成的对角矩阵，

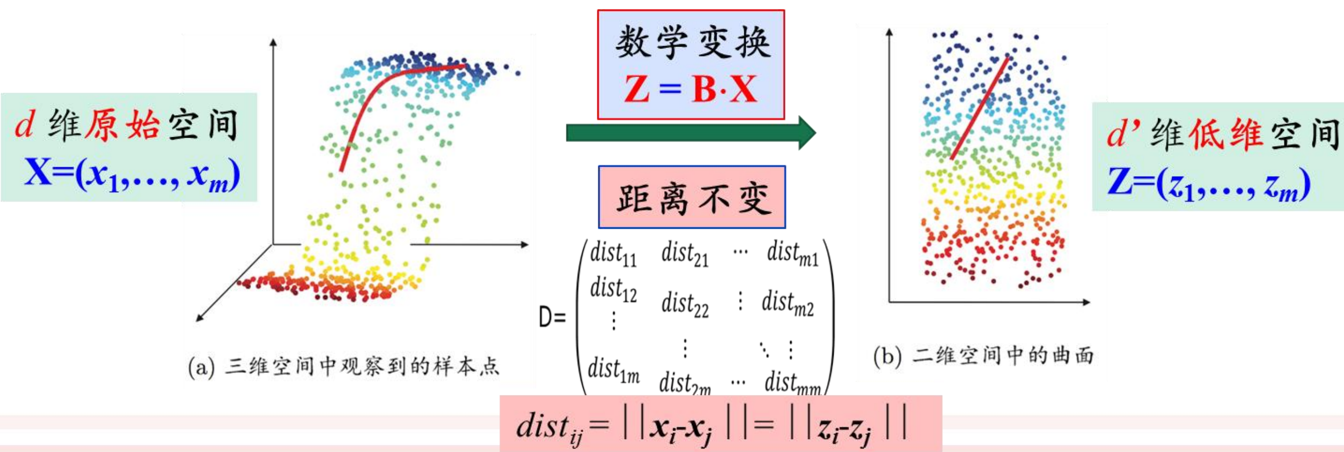
    $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$
- 假定，有d^*个非零特征值，对应的
  - 对角矩阵 $\boldsymbol{\Lambda}^* = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d^*})$，相应的特征矩阵 为$\mathbf{V}^*$
- $\mathbf{Z}$表示为：$\mathbf{Z} = \lambda_{d^*}^{1/2}\, \mathbf{V}_*^{T} \in \mathbf{R}^{d^* \times m}$， 近似解 $\mathbf{Z} = \bar{\boldsymbol{\Lambda}}^{1/2}\, \bar{\mathbf{V}}^{T}$

低维嵌入 — 降维

原始空间中样本之间的距离，在低维子空间中得以保持不变

- 数学变换 $\mathbf{Z} = \mathbf{B}\cdot\mathbf{X}$
- 目标：求内积矩阵B，使得：

- 样本在d'维低维空间中的欧氏距离 等于 原始空间中的距离
  即，d 维原始空间和d' 维低维空间，具有相同的距离矩阵D
- 在现实应用中，为了有效降维，往往仅需降维后的距离与原始空间中的距离，尽可能接近，而不必严格相等。



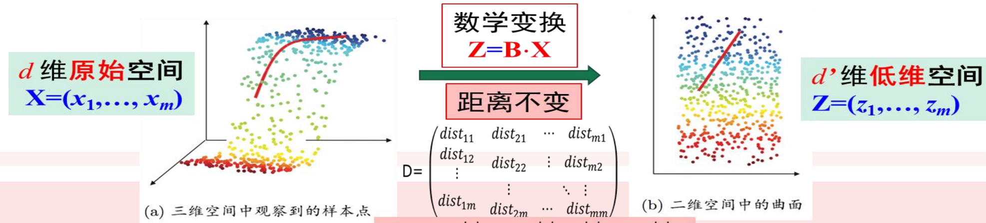
10.2 低维嵌入 -- 降维

- **降维 原则：低维子空间，能够保持 原始高维空间 数据分布**
  - 原始空间中样本之间的距离，在低维子空间中得以保持不变

- 数学变换 $\mathbf{Z} = \mathbf{B}\cdot\mathbf{X}$
  - 可通过降维前后 保持不变的距离矩阵D，求取内积矩阵B

- 令 $\mathbf{B}=\mathbf{Z}^{T}\cdot \mathbf{Z}$　则Z精确解：$\mathbf{Z} = \lambda_{d^*}^{1/2}\, \mathbf{V}_{*}^{T} \in \mathbf{R}^{d^*\times m}$
- 在现实应用中，为了有效降维，往往仅需降维后的距离与原始空间中的距离，尽可能接近，而不必严格相等。
  - 可取 $d' \ll d$ 个最大特征值构成对角矩阵 $\bar{\Lambda}=\mathrm{diag}(\lambda_1,\dots,\lambda_{d'})$，$\bar{\mathbf{V}}$表示相应的特征向量矩阵

- **Z 近似解**　$\mathbf{Z} = \bar{\Lambda}^{1/2}\, \bar{\mathbf{V}}^{T} \in \mathbf{R}^{d'\times m}$



$dist_{ij} = ||\mathbf{x}_i-\mathbf{x}_j|| = ||z_i-z_j||$

10.2 多维缩放

MDS算法的描述

输入: 距离矩阵 $\mathbf{D} \in \mathbb{R}^{m\times m}$, 其元素 $dist_{ij}$ 为样本 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 的距离;
低维空间维数 d'.

过程:

1: 根据式(10.7)–(10.9)计算 $dist^2_{i\cdot}$, $dist^2_{\cdot j}$, $dist^2_{\cdot\cdot}$;

2: 根据式(10.10)计算矩阵 $\mathbf{B}$;

3: 对矩阵 $\mathbf{B}$ 做特征值分解;

4: 取 $\tilde{\mathbf{\Lambda}}$ 为 d' 个最大特征值所构成的对角矩阵, $\tilde{\mathbf{V}}$ 为相应的特征向量矩阵.

输出: 矩阵 $\tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}^{1/2} \in \mathbb{R}^{m\times d'}$, 每行是一个样本的低维坐标

图 10.3 MDS 算法

求 距离矩阵 D

求 内积矩阵 B

内积矩阵 B特征值分解$\mathbf{B}=\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$

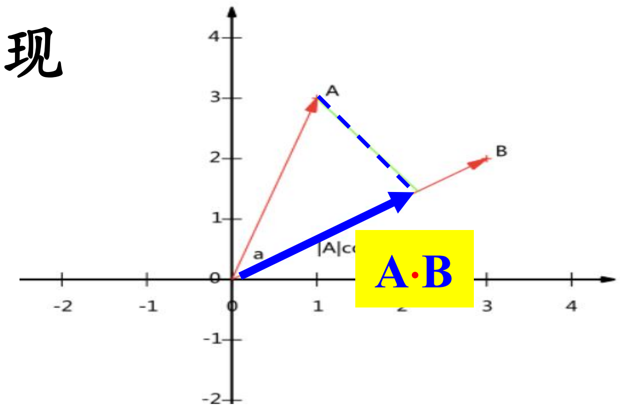
$\mathbf{Z} = \bar{\mathbf{\Lambda}}^{1/2}\bar{\mathbf{V}}^T$ $\bar{\mathbf{\Lambda}}=\mathrm{diag}(\lambda_1,\ldots,\lambda_{d'})$

降维

- 样本空间缩小，维度不变
  - k近邻学习
- 样本空间**维度降低**
  - 低维嵌入（降维）缓解 维数灾难
  - 低维子空间 获取方法
    - 线性降维
      - 主要方法：主成分分析 PCA

线性降维 — 2维→1维

- 获得低维子空间，最简单：对原始高维空间 进行 线性变换
- n 维向量：为n维空间中的一条从原点发射的有向线段。
  - 二维平面上，A(x_1, y_1,)、B(x_2, y_2,)
  - A和B均为二维向量，可以用两条发自原点的有向线段表示
- 2维向量降为1维实数，可通过内积运算实现



- 内积运算
  - 从A点向B所在直线引一条垂线
    垂线与B的交点，叫做 A在B上的投影
  - A与B的夹角是a，则投影的矢量长度为$|A|\cos(a)$，即

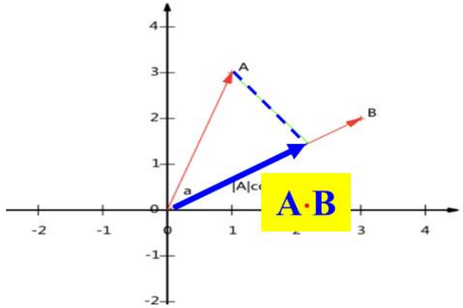
$$A \cdot B = |A|\cos(a) \quad \text{s.t.} \quad |B| = 1$$

  - A·B：A向B所在直线投影的矢量长度

A·B：A投影到B上

线性降维

- 2 维 → 1 维 内积运算:
  - A·B，即A投影到B上 $A \cdot B = |A|\cos(a)$



A·B：A投影到B上

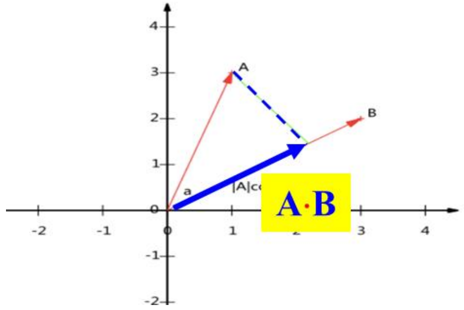
$Z = W^{T} \cdot X$，$Z^{T} = X^{T} \cdot W$

X^{T}投影到W上，得到Z^{T}

- 对原始d维高维空间进行线性变换，获得d'维低维子空间
  - d维空间中的样本 $X = (x_1, \ldots, x_m) \in R^{d \times m}$
  - d'维低维子空间中的样本 $Z = W^{T} \cdot X$， $Z \in R^{d' \times m}$
  - Z是样本X在新空间中的表达；$W \in R^{d \times d'}$正交变换矩阵

线性降维

- **2 维 → 1 维 内积运算:**
  - A·B，即A投影到B上 $A \cdot B = |A|\cos(a)$



A·B：A投影到B上

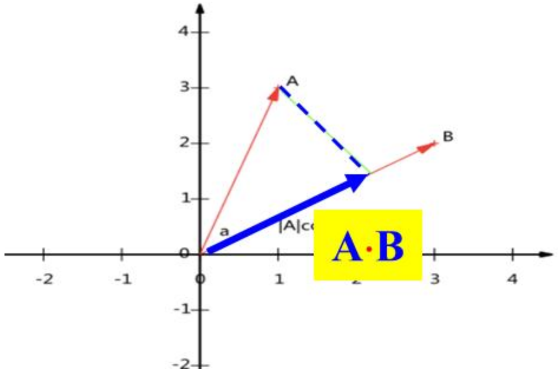
$Z = W^{T} \cdot X$，$Z^{T} = X^{T} \cdot W$

X^{T}投影到W上，得到Z^{T}

- d 维高维空间的样本 $X = (x_1, \ldots, x_m) \in R^{d \times m}$
- d' 维低维子空间的样本 $Z = W^{T} \cdot X$，$W \in R^{d \times d'}$正交变换矩阵
- 矩阵相乘 $Z = W^{T} \cdot X$
  - 将矩阵X中的每一列向量 变换到 矩阵W为基所表示的空间中去，得到新投影矩阵Z
  - 样本x_i在新坐标系$W = \{\omega_1, \ldots, \omega_{d'}\}$中的坐标向量(投影) $z_i = W^{T} \cdot x_i$
  - 新空间中的属性，是原空间中的属性的线性组合

线性降维

- 2 维 → 1 维 内积运算：
  - A·B，即A投影到B上 A·B = |A|cos(a)



A·B：A投影到B上

$Z = W^T \cdot X$，$Z^T = X^T \cdot W$

X^T 投影到W上，得到Z^T

- d 维高维空间的样本 $X = (x_1, \ldots, x_m) \in R^{d \times m}$
- d' 维低维子空间的样本 $Z = W^T \cdot X$，$W \in R^{d \times d'}$正交变换矩阵

- 对低维子空间性质的要求，可通过对 W 施加约束来实现。
- 若要求 低维子空间 对样本具有最大可分性，则得到一种极为常用的线性降维方法 主成分分析 PCA

主成分分析 PCA

- 对于正交属性空间中的样本点，如何用一个超平面对所有样本进行恰当的表达?

  - 原样本点 x_i，新空间中的投影是 $z_i = \mathbf{W}^{\mathrm{T}} x_i$
  - 基于投影 z_i 重构的样本点 $\hat{x}_i$

1. 最近重构性

✓ 样本点到超平面的距离足够近

2. 最大可分性

✓ 样本点在超平面上的投影能尽可能分开

主成分分析 PCA

寻找一个超平面，应具有这样的性质：

1. 最近重构性

- ✓ 样本点 到这个超平面的距离足够近
- ✓ 样本点 x_i 与基于投影 z_i 重构的样本点$\hat{x}_i$之间，距离应最小

整个训练集（m个样本），原样本点 x_i 与 基于投影 z_i（d'维）重构的样本点$\hat{x}_i$之间的距离

$$\sum_{i=1}^{m}\left\|\hat{x}_i - x_i\right\|_2^2=\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'}\left(z_{ij}\cdot w_j - x_i\right)\right\|_2^2 = \sum_{i=1}^{m} z_i^{\mathrm{T}}\cdot z_i - 2\sum_{i=1}^{m} z_i^{\mathrm{T}}\cdot \mathrm{W}^{\mathrm{T}}\cdot x_i + const$$

$$\approx -\sum \mathrm{W}^{\mathrm{T}}\left(\sum x_i x_i^{\mathrm{T}}\right)\mathrm{W} = -\mathrm{W}^{\mathrm{T}}\cdot \mathrm{XX}^{\mathrm{T}}\cdot \mathrm{W}$$

10.3 主成分分析 PCA

- 对于正交属性空间中的样本点，如何用一个超平面对所有样本进行恰当的表达？
  - 原样本点 x_i，新空间中的投影是 $z_i=\mathrm{W}^\mathrm{T}x_i$
  - 基于投影 z_i 重构的样本点 $\hat{x}_i$

整个训练集，原样本点 x_i 与 基于投影 z_i 重构的样本点 $\hat{x}_i$ 之间的**距离**：

$$\sum_{i=1}^{m}\left\|\hat{x}_i - x_i\right\|_2^2=\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'}\left(z_{ij}\cdot w_j - x_i\right)\right\|_2^2=\sum_{i=1}^{m}z_i^\mathrm{T}\cdot z_i-2\sum_{i=1}^{m}z_i^\mathrm{T}\cdot\mathrm{W}^\mathrm{T}\cdot x_i+const$$

$$\approx -tr\left(\sum \mathrm{W}^\mathrm{T}\left(\sum x_i x_i^\mathrm{T}\right)\mathrm{W}\right)$$

- **tr (A)**：矩阵A的迹trace
  - 主对角线元素的总和
  - 所有特征值的和

主成分分析 PCA

寻找一个超平面，应具有这样的性质：

1. 最近重构性

- ✓ 样本点 到这个超平面的**距离**足够**近**
- ✓ 样本点 x_i 与基于投影 z_i 重构的样本点$\hat{x}_i$之间，**距离**应**最小**

$$\min \sum_{i=1}^{m} \left\| \hat{x}_i - x_i \right\|_2^2 = \min - \mathrm{W}^{\mathrm{T}} \cdot \mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} \quad s.t. \quad \mathrm{W}^{\mathrm{T}} \mathrm{W} = I$$

拉格朗日函数 $L(\mathrm{W}) = -\mathrm{W}^{\mathrm{T}} \cdot \mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} + \lambda \left(\mathrm{W}^{\mathrm{T}} \mathrm{W} - I \right)$

$$\boxed{\mathrm{W} = \arg\min_{w} - \mathrm{W}^{\mathrm{T}} \cdot \mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} + \lambda \left(\mathrm{W}^{\mathrm{T}} \mathrm{W} - I \right)}$$

$$\frac{\partial L(\mathrm{W})}{\partial \mathrm{W}} = 0 \quad \Rightarrow \quad \mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} = \lambda \mathrm{W}$$

$$\mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} = \lambda \mathrm{W}$$

W为矩阵XX^T对应的特征向量矩阵，λ为对应的特征值

主成分分析 PCA

寻找一个超平面，应具有这样的性质：

1. 最近重构性

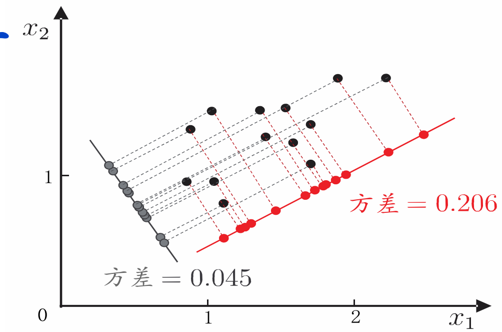
✓ 样本点到超平面Z的距离足够近，即 x_i 和 $\hat{x}_i$ 足够近

$$\min_{\mathbf{W}} \ -\mathbf{W}^{T}\mathbf{X}\mathbf{X}^{T}\mathbf{W} \quad s.t.\ \mathbf{W}^{T}\mathbf{W}=\mathbf{I}$$

$$\mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} = \lambda\, \mathrm{W}$$

2. 最大可分性

✓ 样本点在这个超平面上的投影能尽可能分开



分散程度可用数学上的方差来表述，

即，应使得投影后样本点 $\{z_i = \mathbf{W}^{T} \cdot x_i\}$ 的

方差 $\sum \mathbf{W}^{T}(\sum x_i x_i^{T})\,\mathbf{W}$ 最大化，即

$$\max_{\mathbf{W}} \ \mathbf{W}^{T}\mathbf{X}\mathbf{X}^{T}\mathbf{W} \quad s.t.\ \mathbf{W}^{T}\mathbf{W}=\mathbf{I}$$

$$\mathrm{XX}^{\mathrm{T}} \cdot \mathrm{W} = \lambda\, \mathrm{W}$$

主成分分析 PCA

寻找一个超平面，应具有这样的性质：

1. 最近重构性

✓ 样本点 到这个超平面的距离足够近

$$\min_W \ -W^T X X^T W \quad s.t.\ W^T W = I$$

2. 最大可分性

✓ 样本点 在这个超平面上的投影能尽可能分开

$$\max_W \ W^T X X^T W \quad s.t.\ W^T W = I$$

PCA优化目标：计算投影矩阵W

$$\min_W \ W^T X X^T W \quad s.t.\ W^T W = I$$

$$XX^T \cdot W = \lambda W$$

主成分分析 PCA

PCA优化目标： 计算投影矩阵W

$$\min_W \ -\mathrm{tr}(W^T X X^T W) \quad \text{s.t. } W^T W = I$$

PCA求解 $XX^T W = \lambda W$：拉格朗日乘子法

① 对协方差矩阵 XX^T 进行特征值分解 $XX^T = V\Lambda V^T$

V：特征向量矩阵

Λ：$\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$：特征值矩阵，$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

协方差矩阵 $C = XX^T$

$c_{ij} = Cov(x_i, x_j) = E\{[x_i - E(x_i)][x_j - E(x_j)]\}$

$Cov(x_i, x_j) = 0$，x_i与x_j 不存在线性相关

C：对称矩阵，对角元素c_{ii}为x_i的方差

$$z_i = \Lambda^{1/2} \cdot V^T$$

主成分分析 PCA

PCA优化目标： 计算投影矩阵**W**

$$\min_W \ -tr(W^T X X^T W) \quad s.t.\ W^T W = I$$

PCA求解 $XX^T W = \lambda W$ ：拉格朗日乘子法

① 对协方差矩阵 XX^T 进行特征值分解 $XX^T = V\Lambda V^T$

② 求得的特征值排序： $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

③ 取前 d' 个特征值对应的特征向量构成 $W = \{\omega_1, \ldots, \omega_{d'}\}$

✓ 即 主成分分析**PCA**的解

主成分分析 PCA

PCA优化目标： 计算投影矩阵**W**

$$\min_{\mathbf{W}} \ -\mathrm{tr}\,(\mathbf{W}^{\mathrm{T}}\mathbf{X}\,\mathbf{X}^{\mathrm{T}}\mathbf{W}) \quad \text{s.t.}\ \mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}$$

PCA求解 $\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}=\lambda\mathbf{W}$：拉格朗日乘子法

① 只需对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 进行特征值分解 $\mathbf{X}\mathbf{X}^{\mathrm{T}}=\mathbf{V}\Lambda\mathbf{V}^{T}$

② 并将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

③ 再取前 d' 个特征值对应的特征向量构成$\{\omega_1,\ldots,\omega_{d'}\}$

✓ 这就是主成分分析的解。

输入： 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
低维空间维数 d'.

过程:

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$;

2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;

3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;

4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}$

输出： 投影矩阵 $\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'})$.

PCA仅需保留**W**与样本的均值向量

便可通过简单的向量减法和矩阵-向量乘法将新测试样本 x 投影至低维空间中

主成分分析 PCA

- PCA的**维数** d' 的选择
  - 通常，是由用户事先指定
  - 或，通过在 d' 值不同的低维空间中对 k-近邻分类器（或其它开销较小的学习器）进行交叉验证来选取较好的 d' 值。
  - 或，从重构的角度设置一个重构阈值，例如令能量比 t=95%，然后选取使下式成立的最小 d' 值：

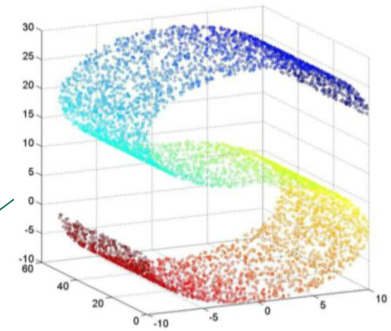
$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq t.$$

主成分分析　PCA

- 降维优点：舍弃部分信息后
  - 能使得样本的采样密度增大
  - 当数据受到噪声影响时，最小的特征值所对应的特征向量往往与噪声有关，舍弃可以起到去噪效果。

- 降维缺点
  - 对应于最小的 $d-d'$ 个特征值的特征向量被舍弃了
  - 信息的损失

核化线性降维

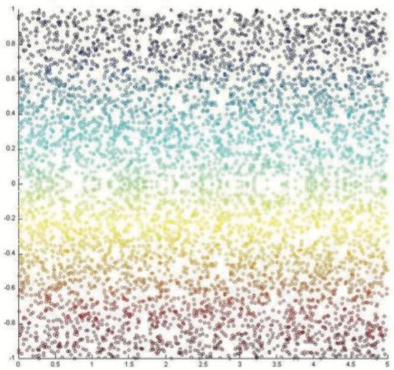


(a) 三维空间中的观察

S 形曲面

$Z=(z_1,\ldots,z_m)$

投影/采样

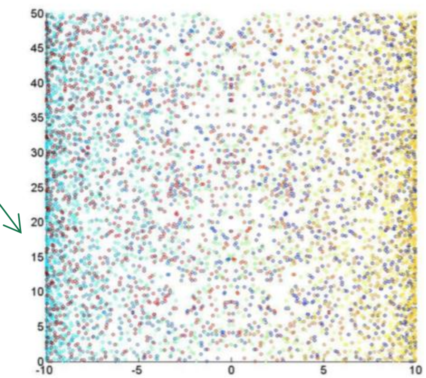


(b) 本真二维结构

数据点的染色显示出低维空间的结构

$X=(x_1,\ldots,x_m)$

PCA



(c) PCA 降维结果

线性降维

丢失低维结构

- **线性降维方法**假设 从高维空间 到 低维空间 的函数映射是**线性**的
- 然而，在不少现实任务中，可能需要**非线性映射**才能找到恰当的低维嵌入。

核化线性降维

核化主成分分析 (Kernelized PCA, 简称KPCA)

- 非线性降维的一种常用方法，基于核技巧对线性降维方法进行“核化”
  - ✓ 使用函数ϕ将原始样本x_i投影/采样 高维特征空间W （超平面）获得样本z_i 在高维特征空间W 上，是线性可分的 $\mathbf{W}=\sum z_i \cdot \boldsymbol{\alpha}_i$

$$z_i = \phi(x_i)$$

  - ✓ PCA求解

- 一般情形下，我们不清楚 ϕ 的具体形式，因此引入核函数

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j)$$

- 新样本 z 投影后的第 j 维坐标:

$$z_j = \boldsymbol{w}_j^{\mathrm{T}} \phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i^j \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i^j \kappa(\boldsymbol{x}_i, \boldsymbol{x})$$

大纲

- k近邻学习
- 低维嵌入--多维缩放
- 主成分分析
- 流形学习
- 度量学习

10.4 流形学习

- 流形学习是一类借鉴了拓扑流形概念的降维方法。
  - “流形”，是在局部与欧氏空间同胚的空间。
  - “流形”，在局部具有欧氏空间的性质，能用欧氏距离来进行距离计算。
- 若低维流形嵌入到高维空间中，则数据样本在高维空间的分布虽然看上去非常复杂，但在局部上仍具有欧氏空间的性质。

  因此，可以容易地在局部建立降维映射关系，然后，再设法将局部映射关系推广到全局。
- 当维数被降至二维或三维时，能对数据进行可视化展示，因此流形学习也可被用于可视化。
- 两种著名的流形学习方法：
  - 等度量映射：试图保持近邻样本之间的距离
  - 局部线性嵌入：试图保持邻域内的线性关系，并使得该线性关系在降维后的低维空间中继续保持。

大纲

- k近邻学习
- 低维嵌入--多维缩放
- 主成分分析
- 流形学习
- 度量学习

10.5 度量学习

研究动机

- 在机器学习中，对高维数据进行降维的主要目的是希望找到一个合适的低维空间，在此空间中进行学习能比原始空间性能更好。
- 事实上，每个空间对应了在样本属性上定义的一个距离度量。寻找合适的空间，实质上就是在寻找一个合适的距离度量。
- 那么，为何不直接尝试“学习”出一个合适的距离度量呢?

- 聚类算法中的距离计算(9.3)给出了多种距离度量的表达式，但它们都是固定的、没有可调节的参数，不能通过对数据样本的学习加以改善.

闵可夫斯基距离

$$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{u=1}^{n}|x_{iu}-x_{ju}|^p\right)^{\frac{1}{p}}$$

欧氏距离

$$\mathrm{dist}_{\mathrm{ed}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=||\boldsymbol{x}_i-\boldsymbol{x}_j||_2=\sqrt{\sum_{u=1}^{n}|x_{iu}-x_{ju}|^2}$$

曼哈顿距离

$$\mathrm{dist}_{\mathrm{man}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=||\boldsymbol{x}_i-\boldsymbol{x}_j||_1=\sum_{u=1}^{n}|x_{iu}-x_{ju}|$$

- 若对距离度量进行学习，就需要构造一个距离的函数，这个函数包含一些变量，即必须有一个便于学习的距离度量表达形式。

降维

- 样本空间缩小，维度不变
  - k近邻学习
- 样本空间维度降低
  - 低维嵌入 (降维) 缓解 维数灾难
  - 低维子空间 获取方法
    - 线性降维
      - 主要方法：主成分分析 PCA