机器学习初步

—主成分分析算法

李爽

助理教授, 特别副研究员

计算机学院 数据科学与知识工程研究所

E-mail: shuangli@bit.edu.cn

Homepage: shuangli.xyz

北京理工大學

BEIJING INSTITUTE OF TECHNOLOGY

本章的主要内容

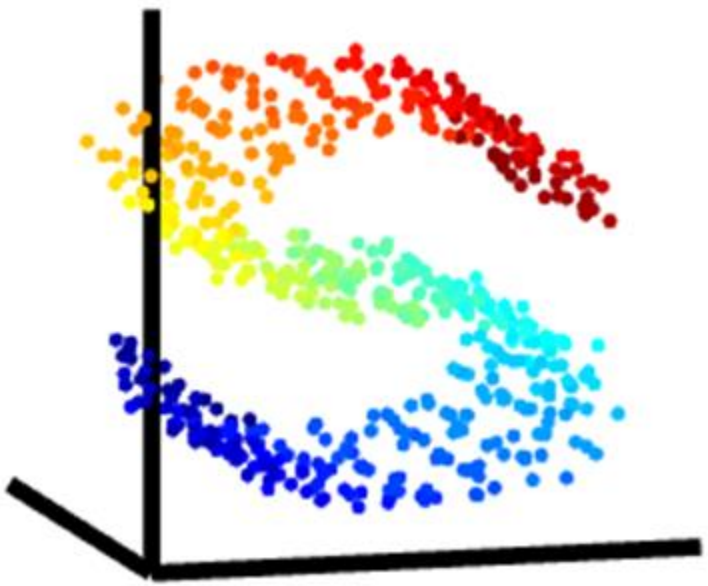
- 低维嵌入
- 主成分分析
- 核化线性降维
- 流形学习
- 度量学习

主成分分析算法

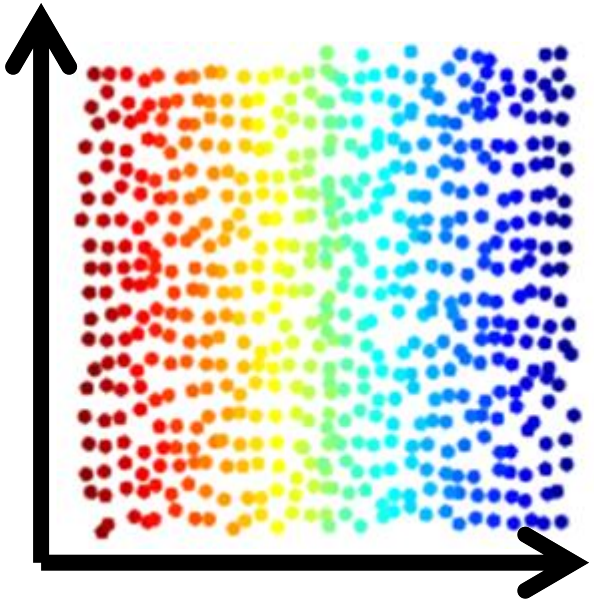
- 低维嵌入
- 主成分分析
- 核化线性降维
- 流形学习
- 度量学习

一、低维嵌入

高维数据 → 低维表示 （直觉上）



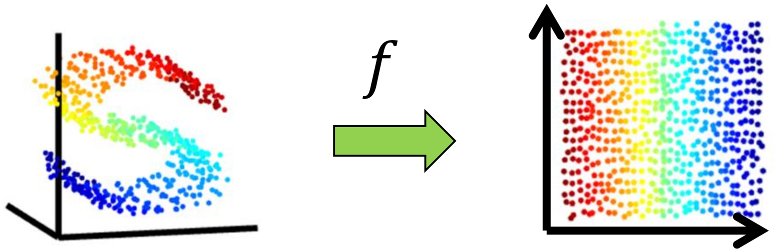
在三维表面上的点
高维数据



曲面点的二维坐标
低维表示

一、低维嵌入

目标：找到映射函数



高维数据 X 低维表示 Z

高维数据 低维表示

一般地：
$$\begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \xrightarrow{\text{降维}} \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}\right)$$

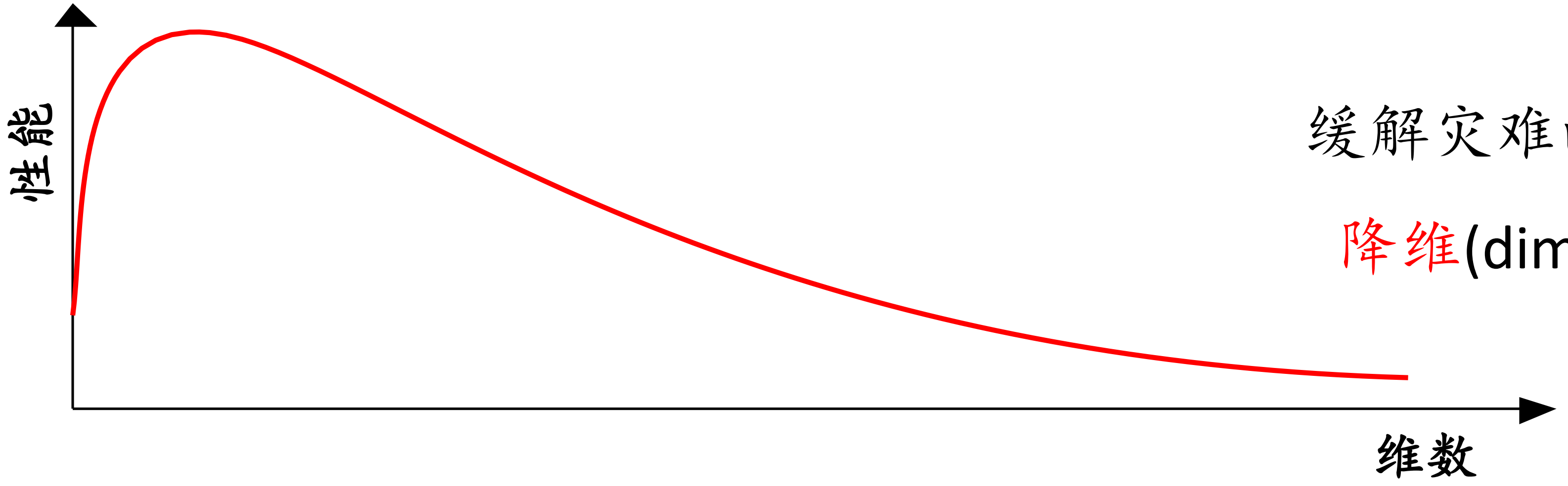
线性降维：
$$\begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mp} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

主成分分析算法

一、低维嵌入

1、维数灾难

- 许多学习方法（如K近邻学习）都涉及距离计算, 而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.
- 事实上, 在高维情形下出现的数据样本稀疏、距离计算困难等问题, 是所有机器学习方法共同面临的严重障碍, 被称为“维数灾难”(curse of dimensionality).



缓解灾难的一个重要途径是降维(dimension reduction)

一、低维嵌入

- 随着维数的增加，目标函数复杂性的指数增长（密度估计）

  "A function defined in high-dimensional space is likely to be much more complex than a function defined in a lower-dimensional space, and those complications are harder to discern." ——弗里德曼

为了更好的学习它，更复杂的目标函数需要更密集的样本点

一、低维嵌入

2、降维

- 降维，亦称“维数约简”，即通过某种数学变换将原始高维属性空间转变为一个低维“子空间”，在这个子空间中样本密度大幅提高，距离计算也变得更为容易。

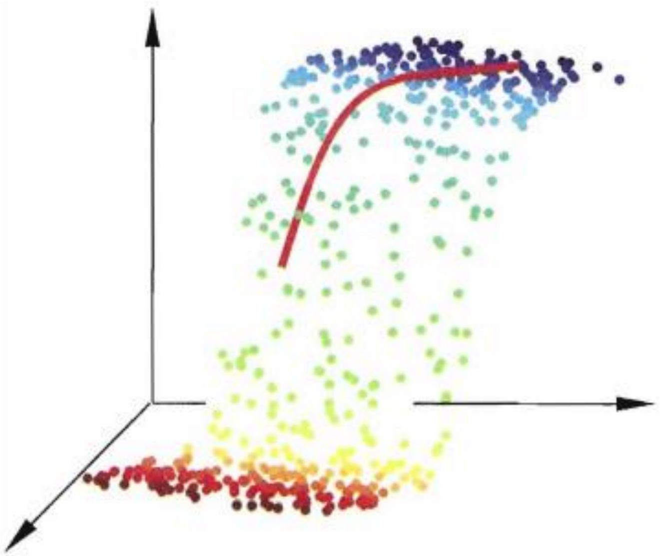
思考：为什么能进行降维?

- 这是因为在很多时候，人们观测或收集到的数据样本虽是高维的，但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”。

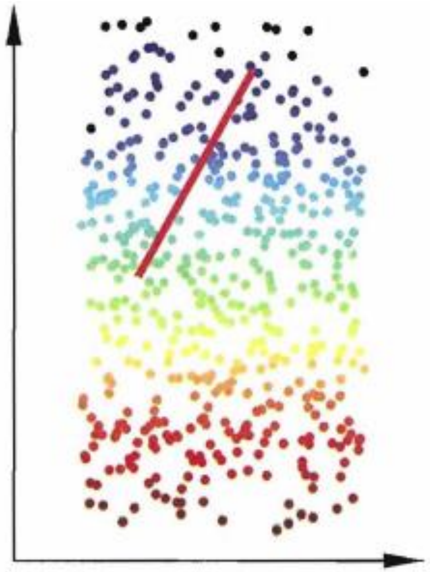
一、低维嵌入

举例：

- 如下图，原始高维空间中的样本点，在这个低维嵌入子空间中更容易进行学习。



三维空间中观察到的样本点



二维空间中的曲面

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

一、低维嵌入

3、经典降维方法：多维缩放

- 若要求原始空间中样本之间的距离在低维空间中得以保持，如上图那样，此方法称为“多维缩放”（MDS）。

方法介绍：

- 假定m个样本在原始空间的距离矩阵为$\mathbf{D} \in \mathbb{R}^{m \times m}$，其第$i$行第$j$列的元素$dist_{ij}$为样本$\boldsymbol{x}_i$到$\boldsymbol{x}_j$的距离。
- 目标：获得样本在d'维空间的表示$\boldsymbol{Z} \in \mathbb{R}^{d' \times m}$，$d' < d$，且任意两个样本在$d'$维空间中的欧式距离等于原始空间中的距离，即

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

一、低维嵌入

- 令$\boldsymbol{B}=\boldsymbol{Z}^T\boldsymbol{Z}\in\mathbb{R}^{m\times m}$其中$\boldsymbol{B}$为降维后样本的内积矩阵，$b_{ij}=\boldsymbol{z}_i^T\boldsymbol{z}_j$，有

$$dist_{ij}^2=\|\boldsymbol{z}_i\|^2+\|\boldsymbol{z}_j\|^2-2\boldsymbol{z}_i^T\boldsymbol{z}_j$$
$$=b_{ii}+b_{jj}-2b_{ij} \quad (1)$$

- 为了便于讨论，令降维后的样本$\boldsymbol{Z}$被中心化，即$\sum_{i=1}^m\boldsymbol{z}_i=0$。显然，矩阵$\boldsymbol{B}$的行与列之和均为0，即$\sum_{i=1}^m b_{ij}=\sum_{j=1}^m b_{ij}=0$。易知

$\sum_{i=1}^m dist_{ij}^2=tr(\boldsymbol{B})+mb_{jj}$ (2)	$\sum_{j=1}^m dist_{ij}^2=tr(\boldsymbol{B})+mb_{ii}$ (3)	$\sum_{i=1}^m\sum_{j=1}^m dist_{ij}^2=2m\,tr(\boldsymbol{B})$ (4)

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

一、低维嵌入

- 其中$tr(\mathbf{B})=\sum_{i=1}^{m}\|\mathbf{z}_i\|^2$，令

$$dist_{i\cdot}^2=\frac{1}{m}\sum_{j=1}^{m}dist_{ij}^2, \qquad (5)$$

$$dist_{\cdot j}^2=\frac{1}{m}\sum_{i=1}^{m}dist_{ij}^2, \qquad (6)$$

$$dist_{\cdot\cdot}^2=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}dist_{ij}^2, \qquad (7)$$

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

一、低维嵌入

- 由式(1)~式(7)可得

$$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2)$$

- 由此即可通过降维前后保持不变的距离矩阵$\boldsymbol{D}$求取内积矩阵$\boldsymbol{B}$。
- 对矩阵$\boldsymbol{B}$做特征值分解，$\boldsymbol{B} = \boldsymbol{V\Lambda V}^T$，其中$\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \dots, \lambda_d)$为特征值构成的对角矩阵，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$，$V$为特征向量矩阵。
- 假定其中由d^*个非零特征值,它们构成对角矩阵$\boldsymbol{\Lambda}_* = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d^*})$,令$\boldsymbol{V}^*$表示相应的特征向量矩阵，则$\boldsymbol{Z}$可表达为

$$\boldsymbol{Z} = \boldsymbol{\Lambda}_*^{1/2}\boldsymbol{V}_*^T \in \mathbb{R}^{d^* \times m}$$

一、低维嵌入

4、线性降维

- 一般来说，欲获得低维子空间，最简单的是对原始高维空间进行线性变换。

- 给定d维空间中的样本$\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)\in\mathbb{R}^{d\times m}$，变换之后得到$d'\leq d$维空间中的样本

$$\boldsymbol{Z}=\boldsymbol{W}^T\boldsymbol{X}$$

- 其中$\boldsymbol{W}\in\mathbb{R}^{d\times d'}$是变换矩阵，$\boldsymbol{Z}\in\mathbb{R}^{d'\times m}$是样本在新空间中的表达

基于线性变换来进行降维的方法称为线性降维方法

主成分分析算法

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

- 低维嵌入
- 主成分分析
- 核化线性降维
- 流形学习
- 度量学习

二、主成分分析

思考:

- 对于正交属性空间中的样本点,如何用一个超平面(直线的高维推广)对所有样本进行恰当的表达?
- 若存在这样的超平面,那么它大概应具有这样的性质:
  - ✓ 最近重构性:样本点到这个超平面的距离都足够近;
  - ✓ 最大可分性:样本点在这个超平面上的投影能尽可能分开。
- 基于最近重构性和最大可分性,能分别得到主成分分析的两种等价推导。

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

二、主成分分析

1、从最近重构性来推导

- 数据样本进行了中心化，即$\sum_i \boldsymbol{x}_i = 0$
- 投影变换后得到的新坐标系为$\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_d\}$，其中$\boldsymbol{\omega}_i$是坐标正交基向量，$\|\boldsymbol{\omega}_i\|_2 = 1, \boldsymbol{\omega}_i^T \boldsymbol{\omega}_j = 0 (i \neq j)$
- 若丢弃新坐标系中的部分坐标，即将维度降低到$d' < d$，则样本点$\boldsymbol{x}_i$在低维坐标系中的投影是$\boldsymbol{z}_i = (z_{i1}, \boldsymbol{z}_{i2}, \dots, z_{id'})$，其中$z_{ij} = \boldsymbol{\omega}_j^T \boldsymbol{x}_i$是$\boldsymbol{x}_i$在低维坐标系下第$j$维的坐标。若基于$\boldsymbol{z}_i$来重构$\boldsymbol{x}_i$，则会得到

$$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij} \boldsymbol{\omega}_j$$

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

二、主成分分析

- 考虑整个训练集，原样本点$\boldsymbol{x}_i$与基于投影重构的样本点$\hat{\boldsymbol{x}}_i$之间的距离为

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{\omega}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\boldsymbol{z}_i^T\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^T\boldsymbol{W}^T\boldsymbol{x}_i+const$$

是个常数

$$\propto -tr\left(\boldsymbol{W}^T\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{W}\right) \qquad (8)$$

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

二、主成分分析

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{\omega}_j - \boldsymbol{x}_i\right\|_2^2 = \sum_{i=1}^{m}\boldsymbol{z}_i^T\boldsymbol{z}_i - 2\sum_{i=1}^{m}\boldsymbol{z}_i^T\mathbf{W}^T\boldsymbol{x}_i + const$$
$$\propto -tr\left(\mathbf{W}^T\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\mathbf{W}\right) \quad (8)$$

- 其中$\mathbf{W} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_d)$.
- 根据最近重构性，式(8)应该被最小化，考虑到$\boldsymbol{\omega}_j$是标准正交基，$\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T$是协方差矩阵，有

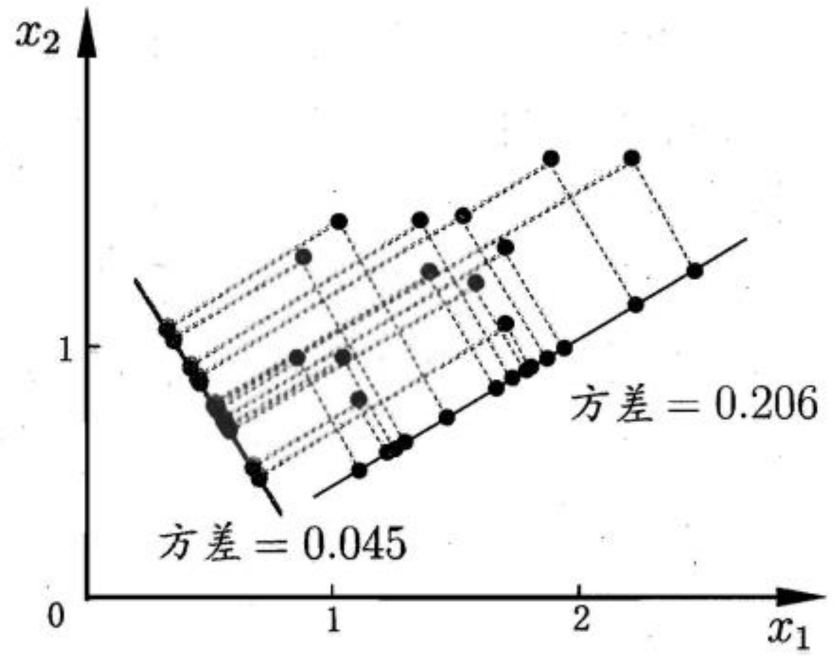
$$\min_{\mathbf{W}} -tr\left(\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}\right)$$
$$s.t.\ \mathbf{W}^T\mathbf{W} = \mathbf{I} \qquad (9)$$

- 这就是主成分分析的优化目标.

二、主成分分析

2、从最大可分性来推导

- 我们知道,样本点在新空间中超平面上的投影是$\boldsymbol{W}^T\boldsymbol{x}_i$,若所有样本点的投影能尽可能分开,则应该使投影后样本点的方差最大化,如下图所示。



二、主成分分析

$$\min_{\boldsymbol{W}} -tr\left(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W}\right)$$
$$s.t.\ \ \boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{I} \qquad (9)$$

- 投影后样本点的方差是$\sum_i \boldsymbol{W}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{W}$
- 于是优化目标可写为

$$\max_{\boldsymbol{W}} \ \ tr\left(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W}\right)$$
$$s.t.\ \ \boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{I} \qquad (10)$$

- 显然，式(9)与(10)等价.

二、主成分分析

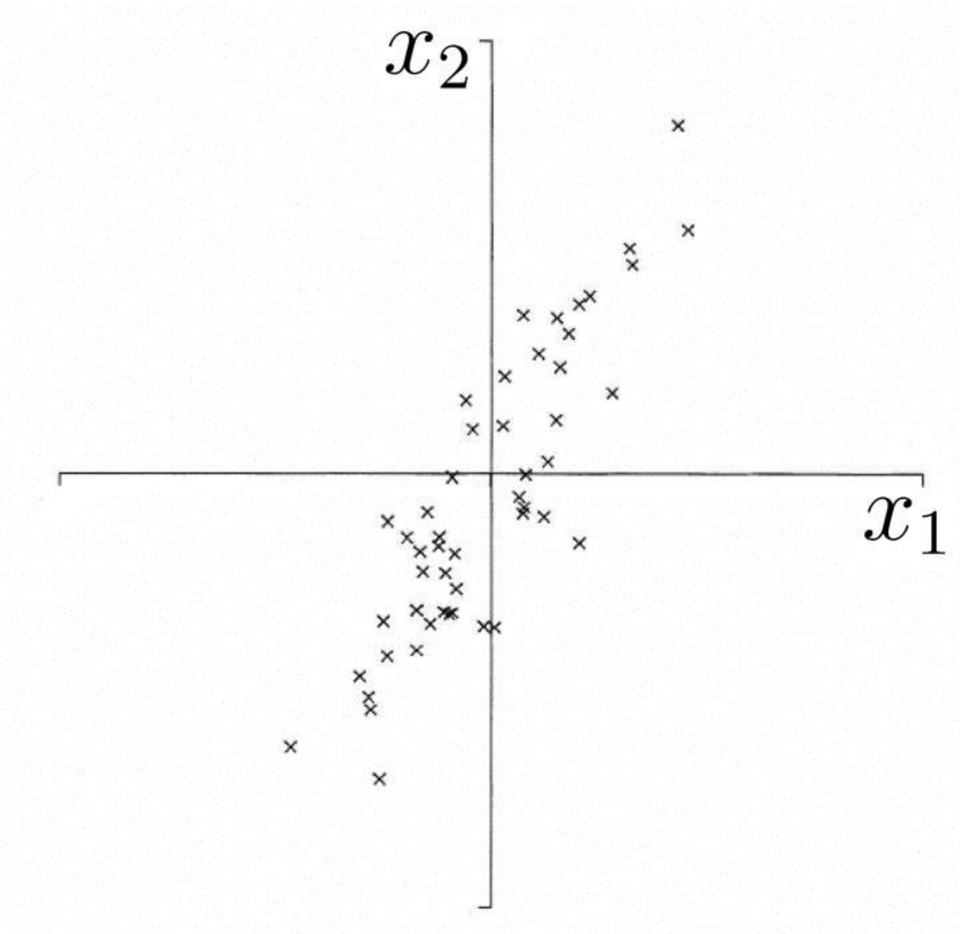
- 对式(9)或式(10)使用拉格朗日乘子法可得

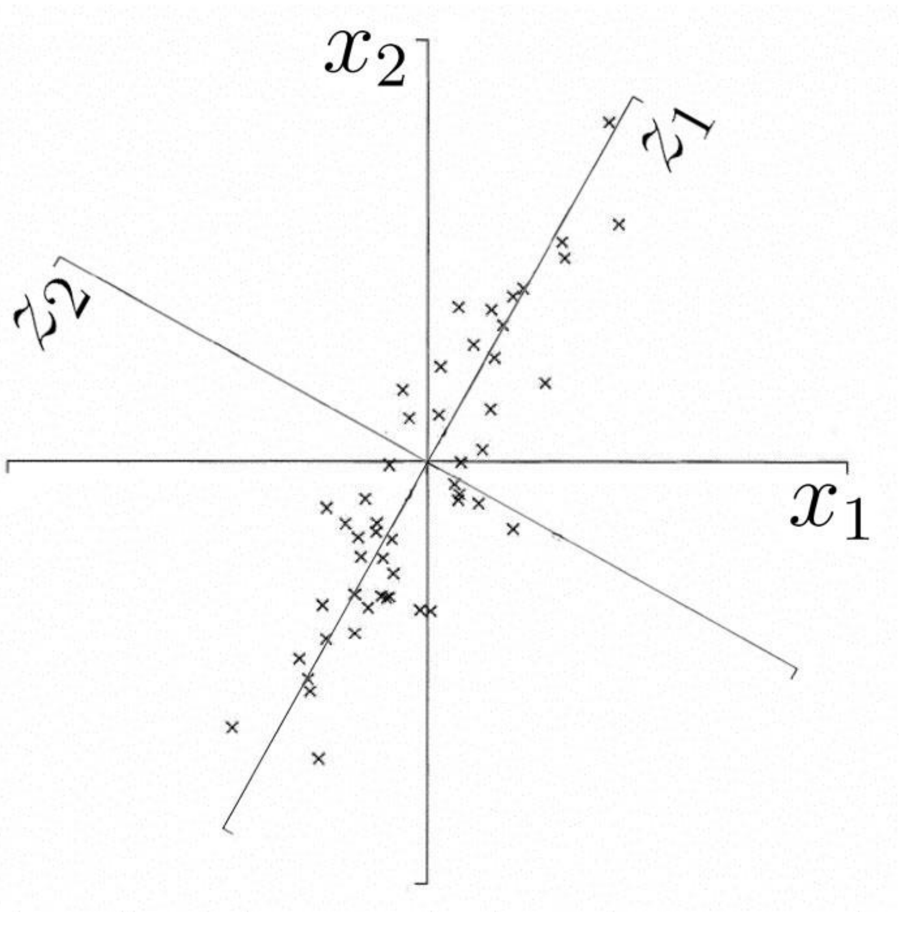
$$\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\omega}_i = \lambda_i\boldsymbol{\omega}_i \qquad (11)$$

- 只需对协方差矩阵$\boldsymbol{X}\boldsymbol{X}^T$进行特征值分解，将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$，再取前$d'$个特征值对应的特征向量构成$\boldsymbol{W}^* = (\omega_1, \omega_2, \ldots, \omega_{d'})$。这就是主成分分析的解.

主成分分析算法

二、主成分分析





X：高维数据的原始坐标　　　　Z：PCA投影坐标

二、主成分分析

- 降维后低维空间的维数d'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（或其他开销较小的学习器）进行交叉验证来选取较好的d'值.
- 对PCA，还可从重构的角度设置一个重构阈值，例如$t = 95\%$，然后选取使下式成立的最小d'值：

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

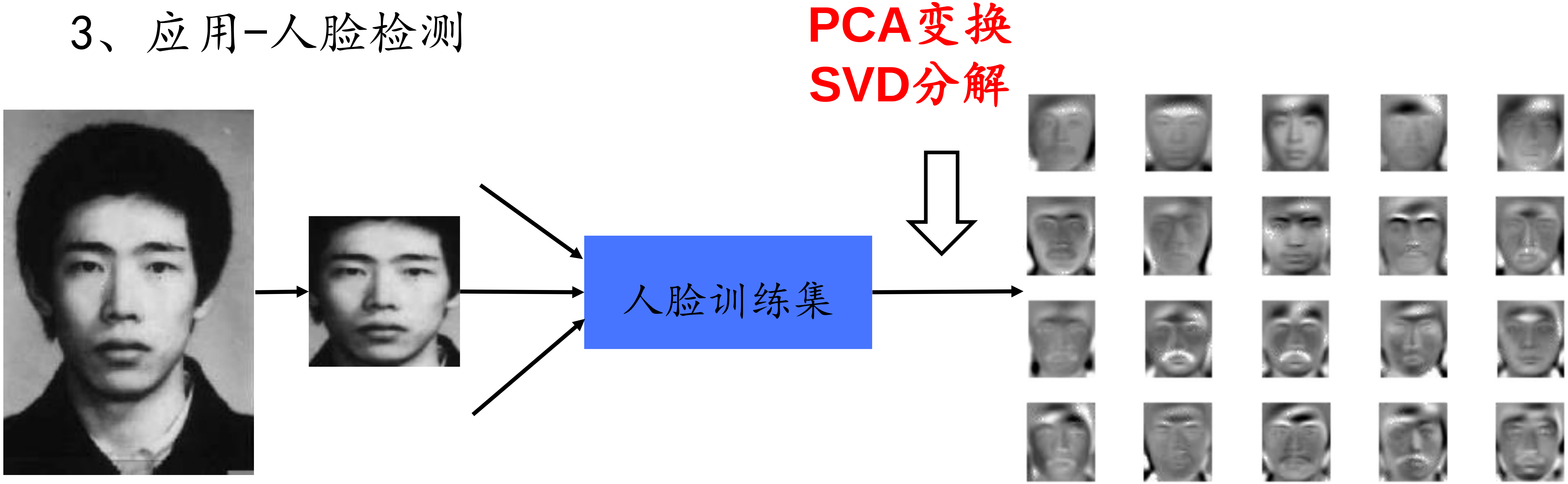
二、主成分分析

- PCA仅需保留$\boldsymbol{W}^*$与样本的均值向量即可通过简单的向量减法和矩阵-向量乘法将新样本投影至低维空间中。
- 显然，降维会导致最小的$d-d'$个特征值的特征向量被舍弃了，但舍弃这部分信息往往是必要的：
  - ✓ 舍弃这部分信息之后能使样本的采样密度增大，这正是降维的重要动机；
  - ✓ 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，将它们舍弃能在一定程度上起到去噪的效果。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

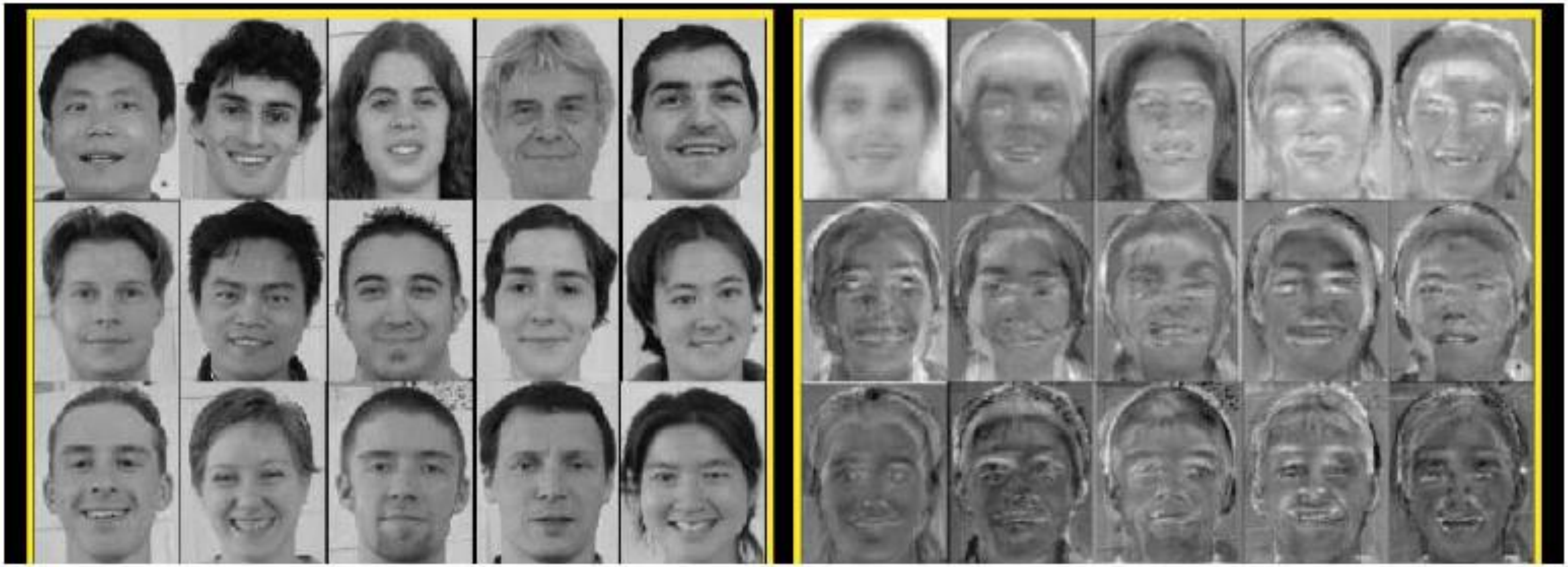
二、主成分分析

3、应用-人脸检测



输入人脸　器官定位归一化　建立人脸训练集　训练出来的特征脸

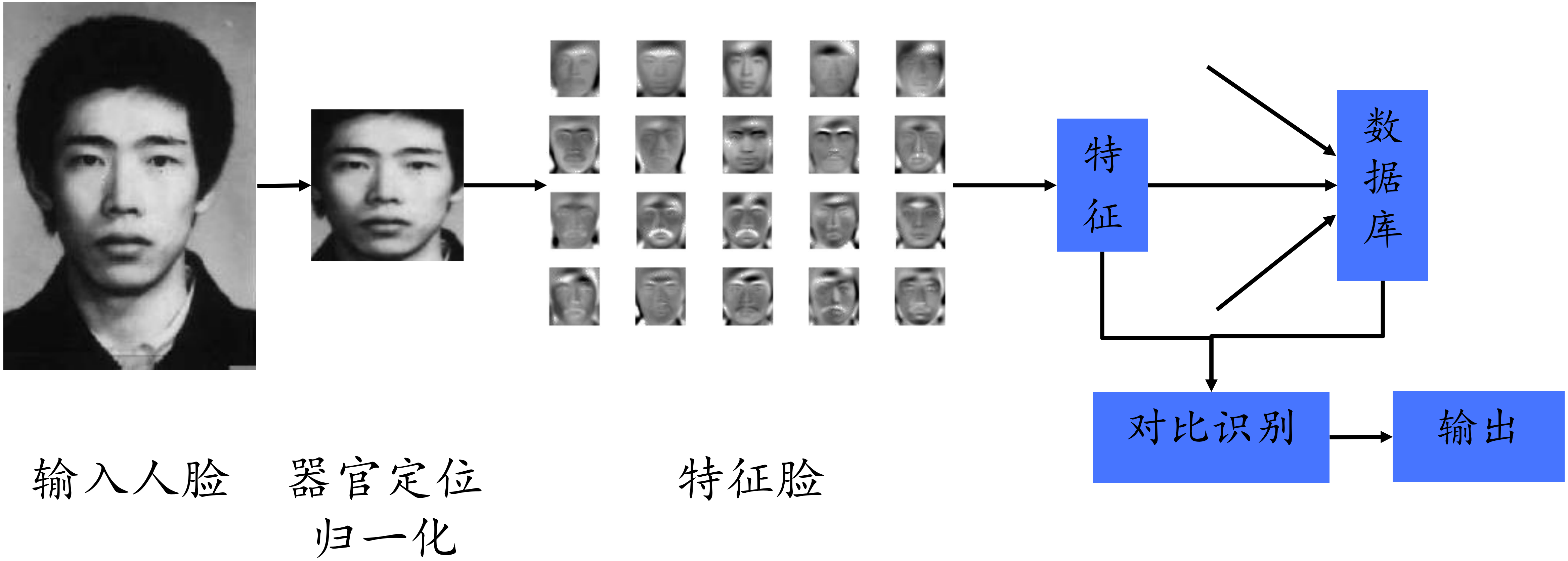
二、主成分分析



二、主成分分析

- 选择大量的人脸图像，PCA变换，得到特征脸
- 选择一个窗口的图像x，向人脸空间投影
- 把投影向量反变换到原始图像空间，得到y
- 计算x和y的差

二、主成分分析



二、主成分分析

应用——图像压缩




d=1 d=2 d=4 d=8

d=16 d=32 d=64 d=100 Original Image

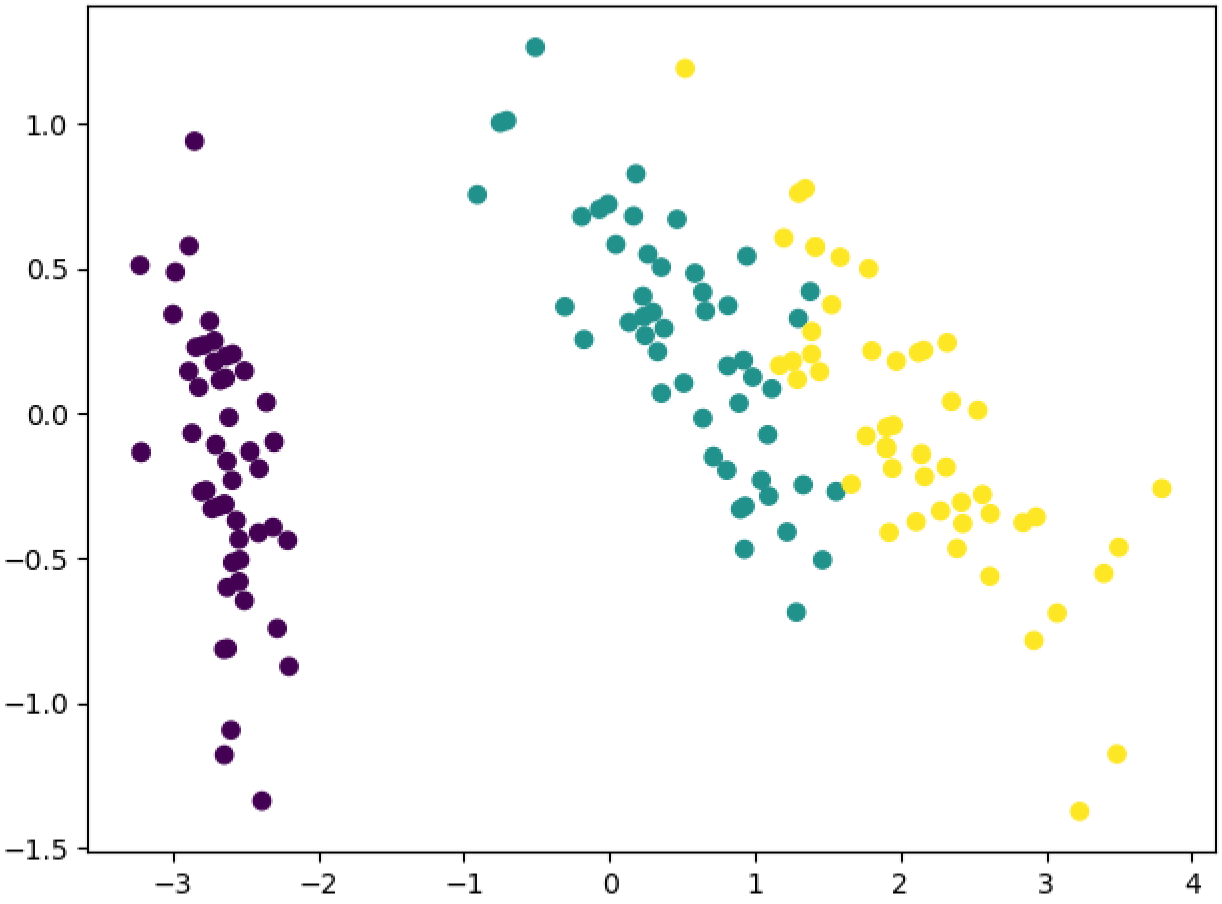




应用——可视化

□ 可视化Iris dataset

- 一种典型的、非常简单的多分类数据集
- 类别：3
- 每类样本数：50
- 维数：4



- 低维嵌入
- 主成分分析
- 核化线性降维
- 流形学习
- 度量学习

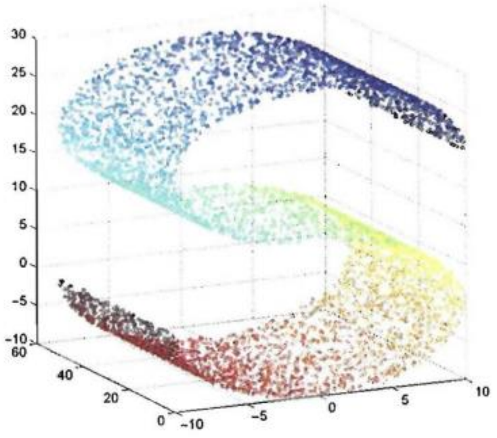
三、核化线性降维

- 线性降维的缺点：

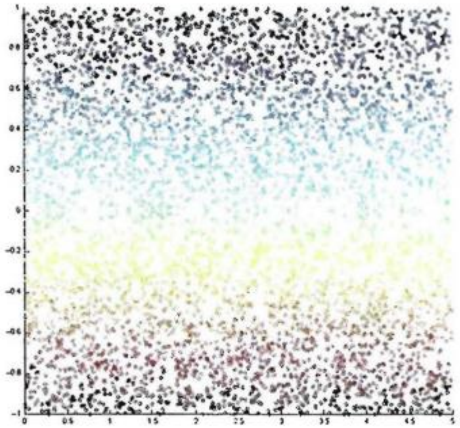
  许多现实任务无法满足线性降维的假设

- 所以，这些任务需要非线性映射才能找到恰当的低维嵌入。

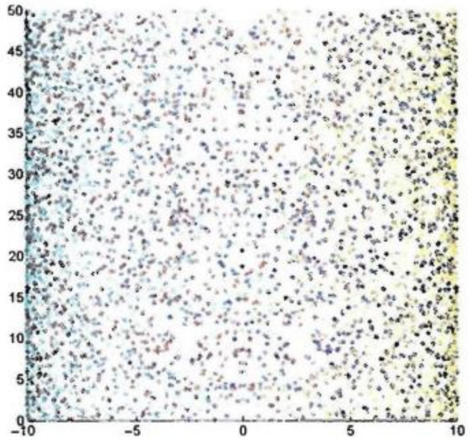
举例：



三维空间中的观察



本真二维结构



PCA降维结果

三、核化线性降维

$$\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\omega}_i = \lambda_i\boldsymbol{\omega}_i \qquad (11)$$

核化线性降维：

- 它是非线性降维的一种常用方法，是基于核技巧对线性降维方法进行“核化”(kernelized)。

举例：核主成分分析(KPCA)

- 假定我们将在高维特征空间中把数据投影到由$\boldsymbol{W} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_d)$确定的超平面上，则对于$\boldsymbol{\omega}_j$，由式(11)有

$$\left(\sum_{i=1}^{m} \boldsymbol{z}_i\boldsymbol{z}_i^T\right)\boldsymbol{\omega}_j = \lambda_j\boldsymbol{\omega}_j, \qquad (12)$$

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

三、核化线性降维

- 其中$\boldsymbol{z}_i$是样本点$\boldsymbol{x}_i$在高维特征空间中的像。易知

$$\boldsymbol{\omega}_j = \frac{1}{\lambda_j}\left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^T\right)\boldsymbol{\omega}_j = \sum_{i=1}^{m} \boldsymbol{z}_i \frac{\boldsymbol{z}_i^T \boldsymbol{\omega}_j}{\lambda_j} = \sum_{i=1}^{m} \boldsymbol{z}_i \alpha_i^j, \qquad (13)$$

- 其中$\alpha_i^j = \frac{1}{\lambda_j}\boldsymbol{z}_i^T \boldsymbol{\omega}_j$是$\boldsymbol{\alpha}_i$的第$j$个分量

三、核化线性降维

$$\left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^T\right) \boldsymbol{\omega}_j = \lambda_j \boldsymbol{\omega}_j, \qquad (12)$$

$$\boldsymbol{\omega}_j = \frac{1}{\lambda_j}\left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^T\right) \boldsymbol{\omega}_j = \sum_{i=1}^{m} \boldsymbol{z}_i \frac{\boldsymbol{z}_i^T \boldsymbol{\omega}_j}{\lambda_j} = \sum_{i=1}^{m} \boldsymbol{z}_i \alpha_i^j, \qquad (13)$$

- 假定$\boldsymbol{z}_i$是由原始属性空间中的样本点$\boldsymbol{x}_i$通过映射ϕ产生，即

$$\boldsymbol{z}_i = \phi(\boldsymbol{x}_i), \qquad i = 1,2,\dots,m$$

- 若ϕ能被显式表达出来，则通过它将样本映射至高维特征空间，再在特征空间中实施PCA即可

式(12)变换为

$$\left(\sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^T\right) \boldsymbol{\omega}_j = \lambda_j \boldsymbol{\omega}_j, \qquad (14)$$

式(13)变换为

$$\boldsymbol{\omega}_j = \sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\, \alpha_i^j, \qquad (15)$$

三、核化线性降维

$$\left(\sum_{i=1}^{m}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^T\right)\boldsymbol{\omega}_j=\lambda_j\boldsymbol{\omega}_j, \qquad (14)$$

$$\boldsymbol{\omega}_j=\sum_{i=1}^{m}\phi(\boldsymbol{x}_i)\,\alpha_i^j, \qquad (15)$$

- 一般情形下，我们不清楚ϕ的具体形式，于是引入核函数

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j), \qquad (16)$$

- 将式(15)和式(16)代入式(14)后化简可得

$$\boldsymbol{K\alpha}^j=\lambda_j\boldsymbol{\alpha}^j, \qquad (17)$$

- 其中$\boldsymbol{K}$为κ对应的核矩阵，$(\boldsymbol{K})_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$，$\boldsymbol{\alpha}^j=(\alpha_1^j,\alpha_2^j,\ldots,\alpha_m^j)$.
- 显然，式(17)是特征值分解问题，取$\boldsymbol{K}$最大的d'个特征值对应的特征向量即可。

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

三、核化线性降维

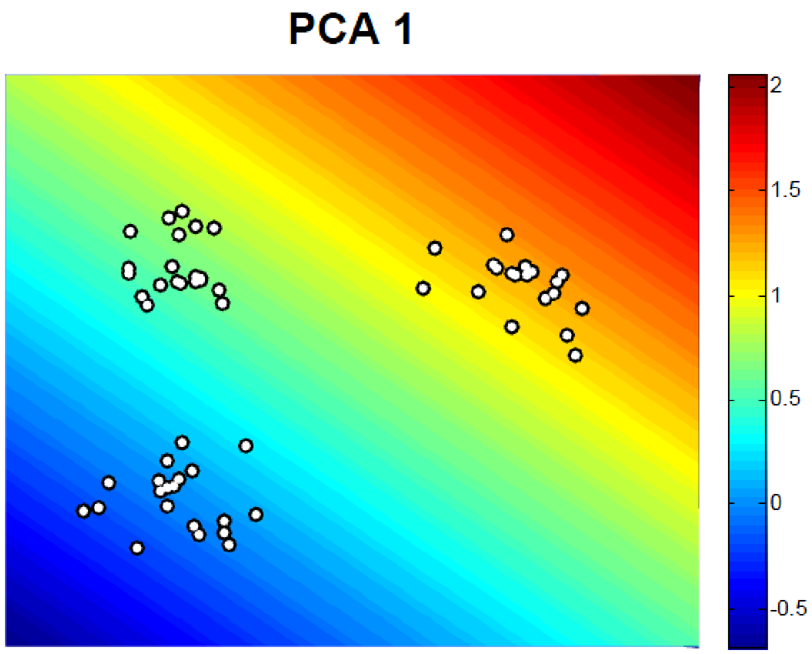
- 对新样本$\boldsymbol{x}$，其投影后的第$j(j=1,2,\ldots,d')$维坐标为

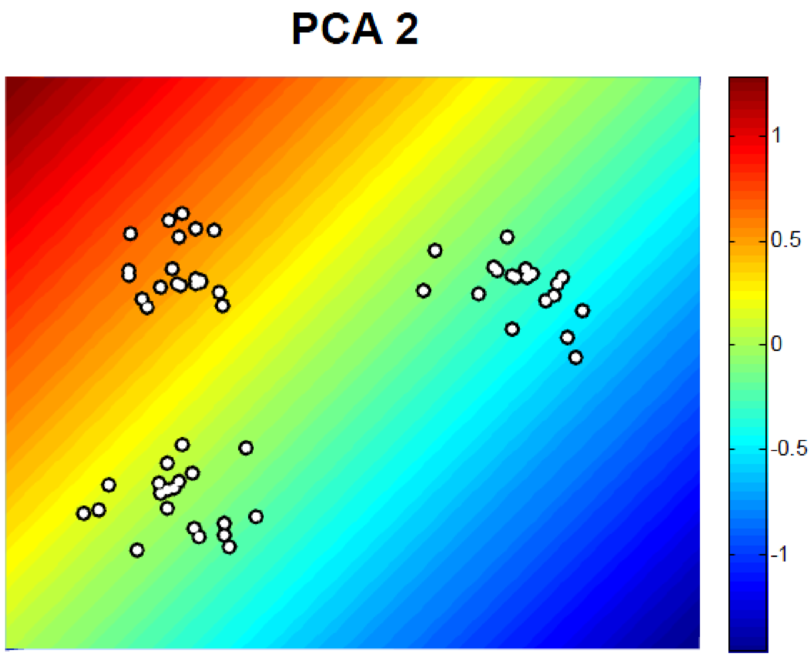
$$z_j=\boldsymbol{\omega}_j^T\phi(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_i^j\phi(\boldsymbol{x})^T\phi(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_i^j\kappa(\boldsymbol{x}_i,\boldsymbol{x}), \qquad (18)$$

- 其中$\boldsymbol{\alpha}_i$已经规范化。
- 式(18)显示出，为获得投影后的坐标，KPCA需对所有样本求和，因此它的计算开销较大

三、核化线性降维

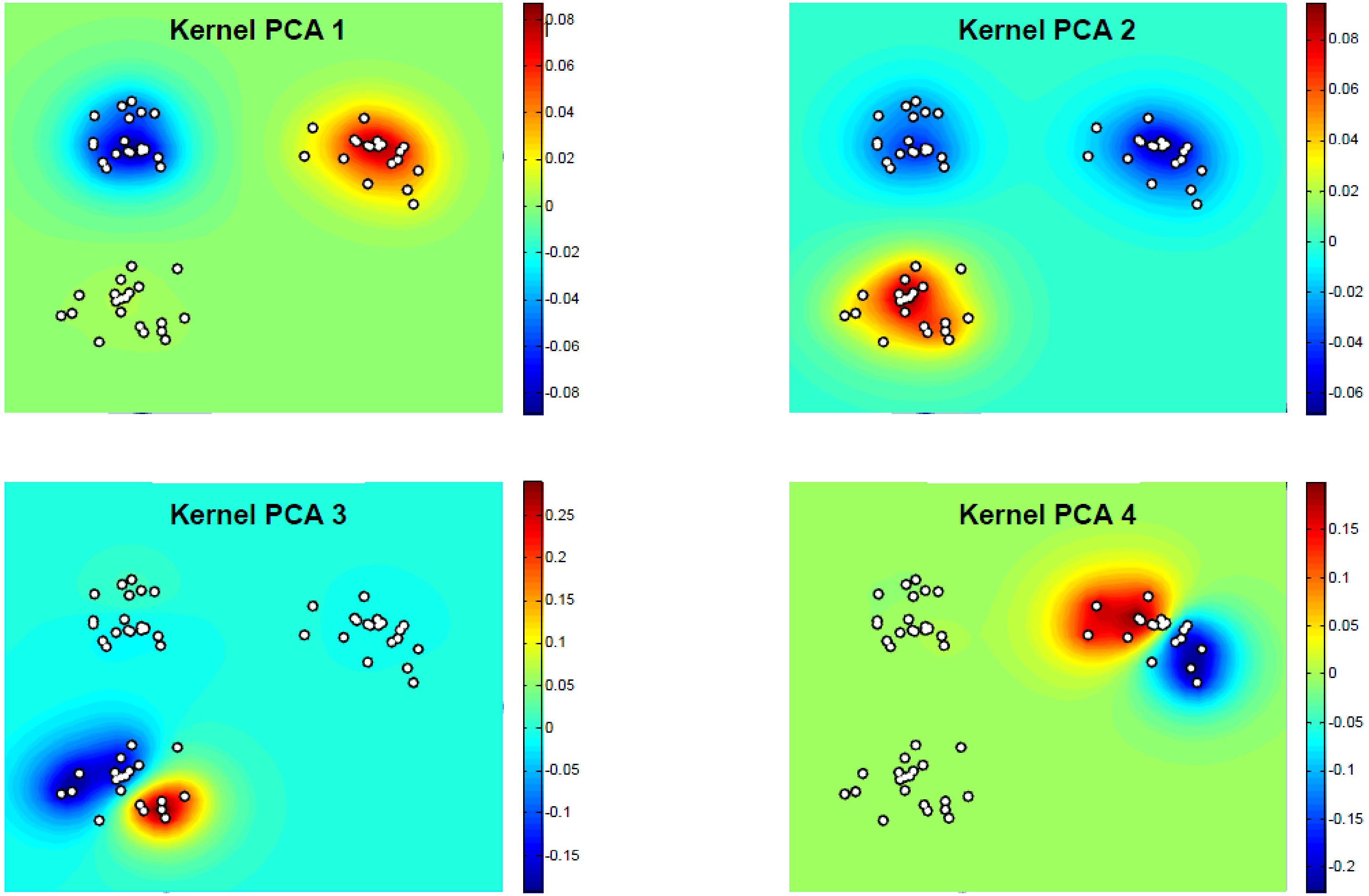
- 线性PCA解决方案





三、核化线性降维

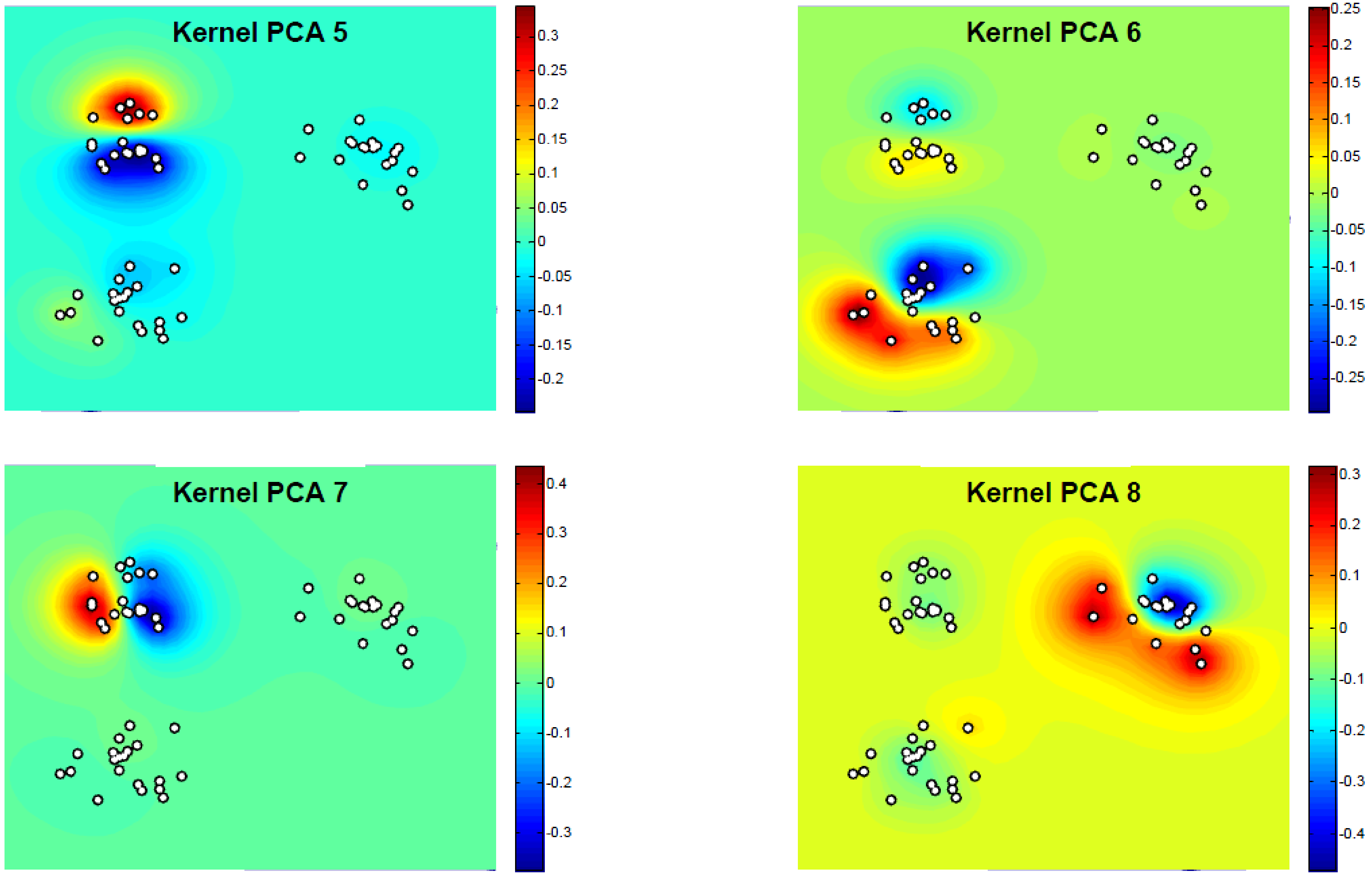
- 核PCA解决方案（高斯核）



主成分分析算法

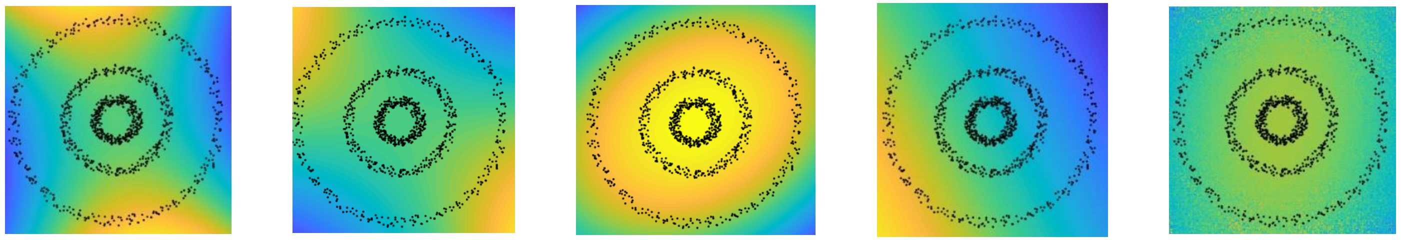
三、核化线性降维

- 更多核PCA投影结果

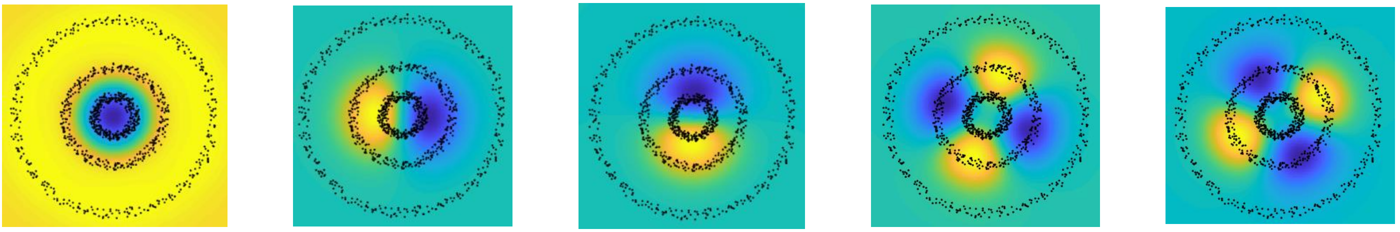


三、核化线性降维

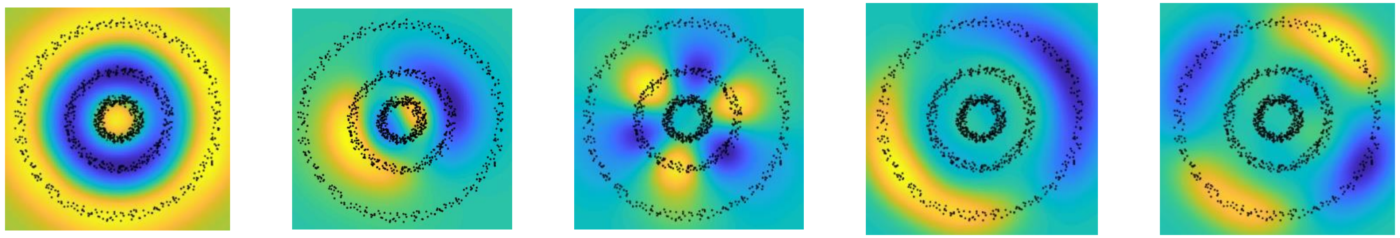
$$k(x,y)=\left(x^{T}y+1\right)^{2}$$



$$k(x,y)=e^{\frac{-\|x-y\|^{2}}{2\sigma^{2}}}$$



$$\sigma=2$$



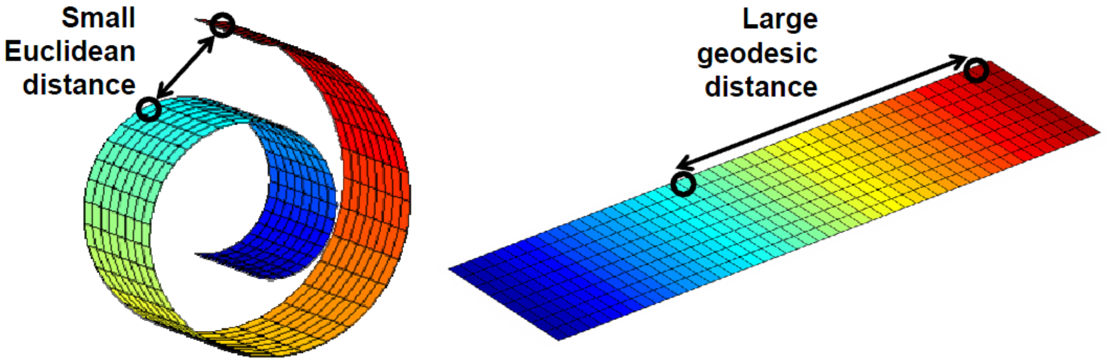
- 低维嵌入
- 主成分分析
- 核化线性降维
- 流形学习
- 度量学习

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

四、流形学习

1、背景

- 传统的机器学习方法中，数据点和数据点之间的距离和映射函数*f*都是定义在欧式空间中的
- 在实际情况中，这些数据点可能不是分布在欧式空间中的，因此传统欧式空间的度量难以用于真实世界的非线性数据，从而需要对数据的分布引入新的假设。



四、流形学习

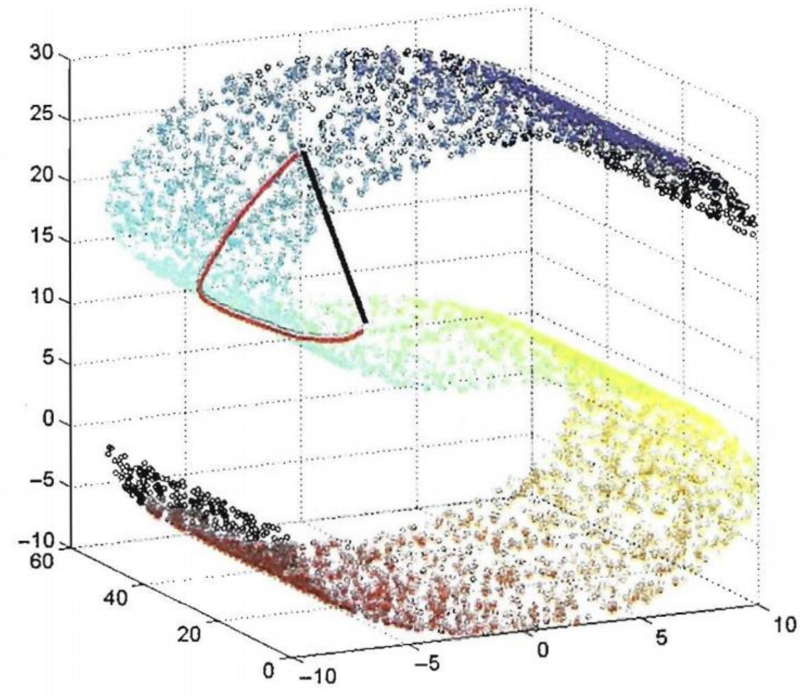
2、概念

- 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，换言之，它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。
- “流形”带来的启发：
  - ✓ 若低维流形嵌入到高维空间中，虽数据样本分布复杂，但在局部上仍具有欧氏空间的性质
  - ✓ 因此，在局部建立降维映射关系，然后设法推广到全局
  - ✓ 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

四、流形学习

3、等度量映射(Isomap)

- 它认为低维流形嵌入到高维空间之后，直接在高维空间中计算直线距离具有误导性，因为高维空间中的直线距离在低维嵌入流形上是不可达的。如图所示



红线 ⇨ 低维嵌入流形上的距离（又叫测地线距离）

黑线 ⇨ 高维空间上的距离

显然，直接在高维空间中计算

直线距离是不恰当的

四、流形学习

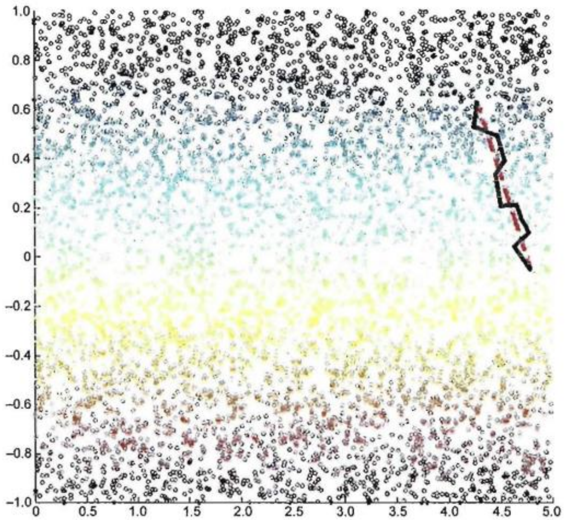
- 问：如何计算测地线距离呢？

利用流形在局部上与欧氏空间同胚这个性质

✓ 对每个点基于欧氏距离找出其近邻点，然后就能建立一个近邻连接图

✓ 图中近邻点之间存在连接，而非近邻点之间不存在连接

✓ 计算两点之间测地线距离的问题，就转变为计算近邻连接图上两点之间的最短路径问题，如图



结论：基于近邻距离逼近能获得低维流形上测地线距离很好的近似.

四、流形学习

Isomap算法描述

输入：样本集$D = \{1,2,...,m\}$；近邻参数k；低维空间维数d'.

过程：

1： **for** $i = 1,2,...,m$ **do**

2： 确定的k近邻；

3： $\boldsymbol{x}_i$与k近邻点之间的距离设置为欧氏距离，与其他点的距离设置为无穷大；

4： **end for**

5： 调用最短路径算法(Dijkstra/Floy)计算任意两样本点之间的距离 dist$(\boldsymbol{x}_i, \boldsymbol{x}_j)$；

6： 将dist$(\boldsymbol{x}_i, \boldsymbol{x}_j)$作为MDS算法的输入；

7： **return** MDS 算法的输出

输出： 样本集D在低维空间的投影$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m\}$.

四、流形学习

- 通过算法，我们可以看到 Isomap 仅是得到了训练样本在低维空间的坐标

问：对于新样本，如何将其映射到低维空间呢？

常用解决方案：

- 将训练样本的高维空间坐标作为输入、低维空间坐标作为输出，训练一个回归学习器来对新样本的低维空间坐标进行预测。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

四、流形学习

- 近邻图的构建方法

✓ 方法一：指定近邻点个数 ⇨ 得到k近邻图

✓ 方法二：指定距离阈值ϵ，距离小于ϵ的点被认为是近邻点 ⇨ 得到ϵ近邻图

两种方法的不足：

- “短路”问题：近邻范围指定得较大，则距离很远的点可能被误认为近邻
- “断路”问题：近邻范围指定得较小，则图中有些区域可能与其他区域不存在连接。

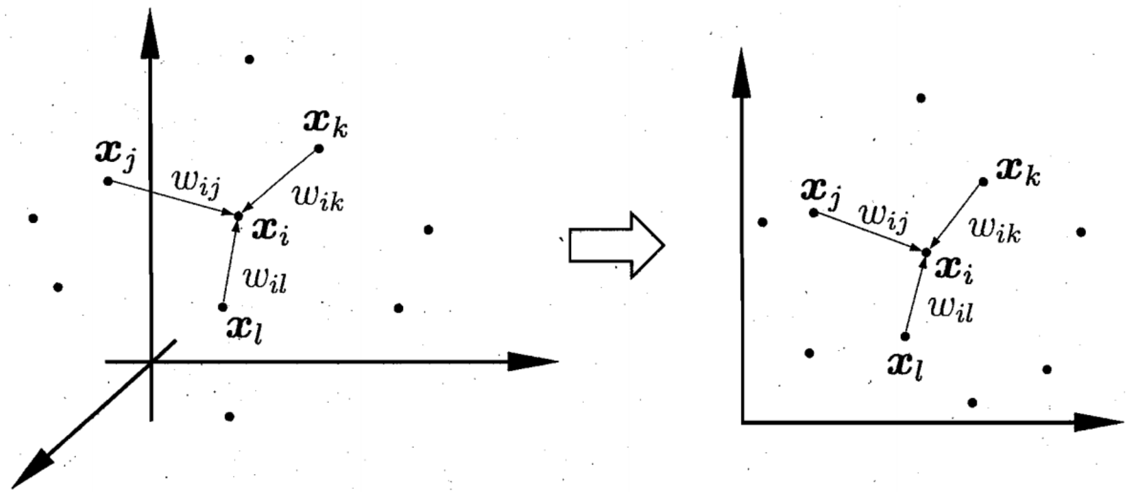
短路与断路都会给后续的最短路径计算造成误导

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

四、流形学习

4、局部线性嵌入

- 局部线性嵌入(LocallyLinear Embedding,简称LLE)与Isomap试图保持近邻样本之间的距离不同，它试图保持邻域内样本之间的线性关系，如下图



- 假定样本点的坐标x_i能通过它的邻域样本x_j , x_k, x_l 的坐标通过线性组合而重构出来，即

$$x_i = w_{ij}x_j + w_{ik}x_k + w_{il}x_l$$

LLE希望上式的关系在低维空间中得以保持.

四、流形学习

- LLE先为每个样本$\boldsymbol{x}_i$找到其近邻下标集合Q_i，然后计算出基于Q_i中的样本点对$\boldsymbol{x}_i$进行线性重构的系数$\boldsymbol{\omega}_i$：

$$\min_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\ldots,\boldsymbol{\omega}_m} \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} \omega_{ij}\, \boldsymbol{x}_j \right\|_2^2 \qquad (19)$$

$$s.t. \sum_j \omega_{ij} = 1$$

其中$\boldsymbol{x}_i$和$\boldsymbol{x}_j$均为已知

四、流形学习

- 令$C_{jk}^{-1}=\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)^{\mathrm{T}}\left(\boldsymbol{x}_i-\boldsymbol{x}_k\right)$，$\omega_{ij}$有闭式解

$$\omega_{ij}=\frac{\sum_{k\in Q_i}C_{jk}^{-1}}{\sum_{l,s\in Q_i}C_{ls}^{-1}}$$

- LLE在低维空间中保持$\boldsymbol{\omega}_i$不变，于是$\boldsymbol{x}_i$对应的低维空间坐标$\boldsymbol{z}_i$可通过下式求解：

$$\min_{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i-\sum_{j\in Q_i}\omega_{ij}\,\boldsymbol{z}_j\right\|_2^2 \qquad (20)$$

四、流形学习

$$\min_{\omega_1,\omega_2,\ldots,\omega_m} \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} \omega_{ij}\, \boldsymbol{x}_j \right\|_2^2 \qquad (19)$$

$$s.t. \sum_{j} \omega_{ij} = 1$$

$$\min_{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m} \sum_{i=1}^{m} \left\| \boldsymbol{z}_i - \sum_{j\in Q_i} \omega_{ij} \boldsymbol{z}_j \right\|_2^2 \qquad (20)$$

- 式(19)与(20)的优化目标同形，唯一的区别是式(19)中需确定的是$\boldsymbol{\omega}_i$，而式(20)中需确定的是$\boldsymbol{x}_i$对应的低维空间坐标$\boldsymbol{z}_i$.

四、流形学习

$$\min_{z_1, z_2, \ldots, z_m} \sum_{i=1}^{m} \left\| z_i - \sum_{j \in Q_i} \omega_{ij} z_j \right\|_2^2, (20)$$

- 令$\mathbf{Z} = (z_1, z_2, \cdots, z_m) \in \mathbb{R}^{d' \times m}$，$(\mathbf{W})_{ij} = \omega_{ij}$

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}} (\mathbf{I} - \mathbf{W}) \qquad (21)$$

- 则式(20)可重写为

$$\min_{\mathbf{Z}} \ \mathrm{tr}\left(\mathbf{Z}\mathbf{M}\mathbf{Z}^{\mathrm{T}}\right) \qquad (22)$$
$$s.t. \ \mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}$$

- 式(22)可通过特征值分解求解:$\mathbf{M}$最小的d'个特征值对应的特征向量组成的矩阵即为$\mathbf{Z}^{\mathrm{T}}$。

四、流形学习

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}}(\mathbf{I} - \mathbf{W})$$

LLE算法描述

输入: 样本集$D = \{1, 2, \ldots, m\}$;近邻参数k;低维空间维数d'.

过程:

1: **for** $i = 1,2,\ldots,m$ **do**

2: 确定$\boldsymbol{x}_i$的k近邻;

3: 从式(19)求得 ω_{ij} , $j \in Q_i$;

4: 对于$j \notin Q_i$, 令$\omega_{ij} = 0$;

5: **end for**

6: 从式(21)得到$\mathbf{M}$;

7: 对$\mathbf{M}$进行特征值分解;

8: **return** $\mathbf{M}$的最小d'个特征值对应的特征向量

输出: 样本集D在低维空间的投影$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_m\}$.

- 低维嵌入
- 主成分分析
- 核化线性降维
- 流形学习
- 度量学习

五、度量学习

1、基本动机

- 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,而寻找合适的空间,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?这就是度量学习(metric learning)的基本动机.

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

五、度量学习

- 欲对距离度量进行学习，必须有一个便于学习的距离度量表达形式
- 做法：可以对平方欧式距离进行推广，得到我们想要的表达形式
- 对两个d维样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$，它们之间的平方欧氏距离可写为

$$dist_{\mathrm{ed}}^2(\boldsymbol{x}_i,\boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = dist_{ij,1}^2 + dist_{ij,2}^2 + \cdots + dist_{ij,d}^2$$

- 其中$dist_{ij,k}$表示$\boldsymbol{x}_i$和$\boldsymbol{x}_j$在第k维上的距离．若假定不同属性的重要性不同，则可引入属性权重$\boldsymbol{\omega}$，得到

$\omega_i \geq 0, \mathbf{W} = diag(\boldsymbol{\omega})$是个对角阵，$(\mathbf{W})_{ii} = \omega_i$

$$dist_{\mathrm{wed}}^2(\boldsymbol{x}_i,\boldsymbol{x}_j) = ||x_i - x_j||_2^2 = \omega_1.dist_{ij,1}^2 + \omega_2.dist_{ij,2}^2 + \cdots + \omega_d.dist_{ij,d}^2$$

$$= (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}\mathbf{W}(\boldsymbol{x}_i - \boldsymbol{x}_j) \qquad (23)$$

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

五、度量学习

- 式(23)中的**W**可通过学习确定.
- 我们已知**W**是对角阵, 它的非对角元素均为零, 这意味着坐标轴是正交的, 即属性之间无关;
- 但现实问题中往往不是这样, 例如考虑西瓜的"重量"和"体积"这两个属性, 它们显然是正相关的, 其对应的坐标轴不再正交.
- 为此, 将式(23)中的**W**替换为一个普通的半正定对称矩阵**M**, 于是就得到了马氏距离(Mahalanobis distance)

$$dist_{mah}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \mathbf{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_{\mathbf{M}}^2$$

五、度量学习

2、以近邻成分分析(NCA)为例求 M

- 假定**M**学习目标是提高近邻分类器的性能，则可将**M**直接嵌入到近邻分类器的评价指标中去，通过优化该性能指标相应地求得**M**.

- 近邻分类器在进行判别时通常使用多数投票法，邻域中的每个样本投1票，邻域外的样本投0票. 不妨将其替换为概率投票法. 对于任意样本$\boldsymbol{x}_j$，它对$\boldsymbol{x}_i$分类结果影响的概率为

$$p_{ij} = \frac{exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\mathbf{M}}^2\right)}{\sum_l exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\mathbf{M}}^2\right)} \qquad (24)$$

五、度量学习

$$p_{ij} = \frac{exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\mathbf{M}}^2\right)}{\sum_l exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\mathbf{M}}^2\right)} \qquad (24)$$

- 当$i = j$时，p_{ij}最大. 显然，$\boldsymbol{x}_j$对$\boldsymbol{x}_i$的影响随着它们之间距离的增大而减小. 若以留一法正确率的最大化为目标，则可计算$\boldsymbol{x}_i$的留一法正确率，即它被自身之外的所有样本正确分类的概率为

$$p_i = \sum_{j \in \Omega_i} p_{ij}$$

- 其中Ω_i表示与$\boldsymbol{x}_i$属于相同类别的样本的下标集合.

五、度量学习

$$p_{ij} = \frac{\exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\mathbf{M}}^2\right)}{\sum_l \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_l\|_{\mathbf{M}}^2\right)} \qquad (24)$$

- 于是, 整个样本集上的留一法正确率为

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \sum_{j \in \Omega_i} p_{ij} \qquad (25)$$

- 将式(24)代入(25), 再考虑到$\mathbf{M} = \mathbf{P}\mathbf{P}^{\mathrm{T}}$, 则NCA的优化目标为

$$\min_{\mathbf{P}} \ 1 - \sum_{i=1}^{m} \sum_{j \in \Omega_i} \frac{\exp\left(-\|\mathbf{P}^T \boldsymbol{x}_i - \mathbf{P}^T \boldsymbol{x}_j\|_2^2\right)}{\sum_l \exp\left(-\|\mathbf{P}^T \boldsymbol{x}_i - \mathbf{P}^T \boldsymbol{x}_l\|_2^2\right)} \qquad (26)$$

- 求解式(26)即可得到最大化近邻分类器留一法正确率的距离度量矩阵$\mathbf{M}$.

五、度量学习

3、扩展

- 实际上，我们不仅能把错误率这样的监督学习目标作为度量学习的优化目标，还能在度量学习中引入领域知识.
- 例如，若已知某些样本相似、某些样本不相似，则可定义“必连”(must-link)约束集合$\mathcal{M}$与“勿连”(cannot-link)约束集合$\mathcal{C}$，$(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{M}$表示$\boldsymbol{x}_i$与$\boldsymbol{x}_j$相似，$(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{C}$表示$\boldsymbol{x}_i$与$\boldsymbol{x}_j$不相似. 显然，我们希望相似的样本之间距离较小，不相似的样本之间距离较大.

五、度量学习

- 于是可通过求解下面这个凸优化问题获得适当的度量矩阵**M**

$$\min_{\mathbf{M}} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_{\mathbf{M}}^2$$

$$s.t. \sum_{(x_i,x_j)\in\mathcal{C}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_k\right\|_{\mathbf{M}} \geq 1, \qquad \mathbf{M} \succcurlyeq 0$$

此公式要求在不相似样本间的距离不小于1的前提下，使相似样本间的距离尽可能小.

- 其中约束 $\mathbf{M} \succcurlyeq 0$ 表明 $\mathbf{M}$ 必须是半正定的.

五、度量学习

- 不同的度量学习方法针对不同目标获得“好”的半正定对称距离度量矩阵 $\mathbf{M}$，若 $\mathbf{M}$ 是一个低秩矩阵，则通过对 $\mathbf{M}$ 进行特征值分解，总能找到一组正交基，其正交基数目为矩阵 $\mathbf{M}$ 的秩rank($\mathbf{M}$)，小于原属性数d.
- 于是，度量学习学得的结果可衍生出一个降维矩阵$\mathbf{P} \in \mathbb{R}^{d \times rank(\mathbf{M})}$，能用于降维之目的.

课后作业

【必做题】PCA算法的详细数学推导

【选做题】谈谈你对核技巧（kernel trick）的理解与应用

谢谢！

李爽

E-mail: shuangli@bit.edu.cn

Homepage: shuangli.xyz

北京理工大學

BEIJING INSTITUTE OF TECHNOLOGY