第十一章 数据降维

§11.1 问题引入

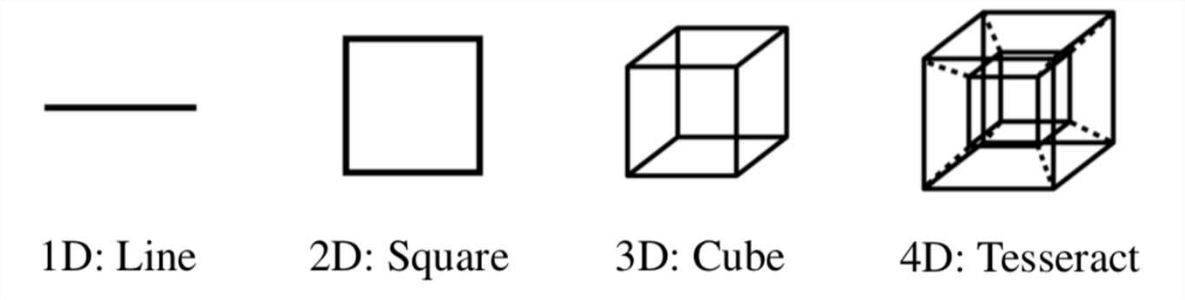
一、数据可视化

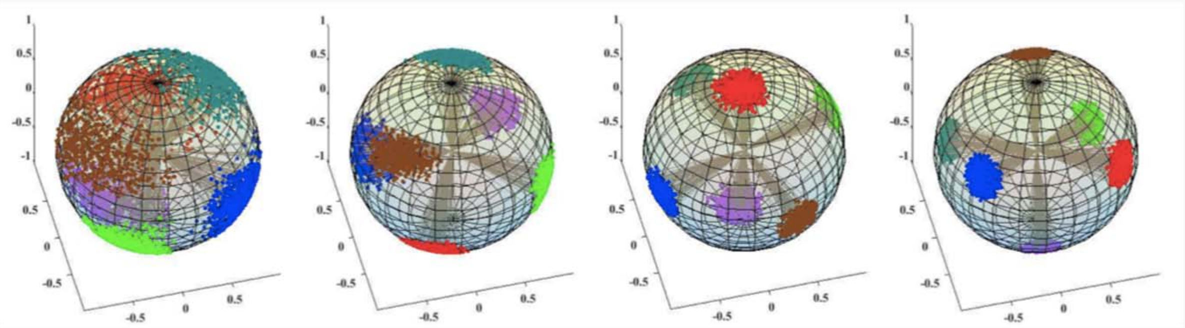
二、小样本学习

三、数据的压缩

问题引入

- 高维数据的低维嵌入可视化
  - 低维数据可直接在空间中作图进行表示，便于对分布进行直观的表现、对比和分析
  - 高维数据的可视化无法直接对原始数据进行展示

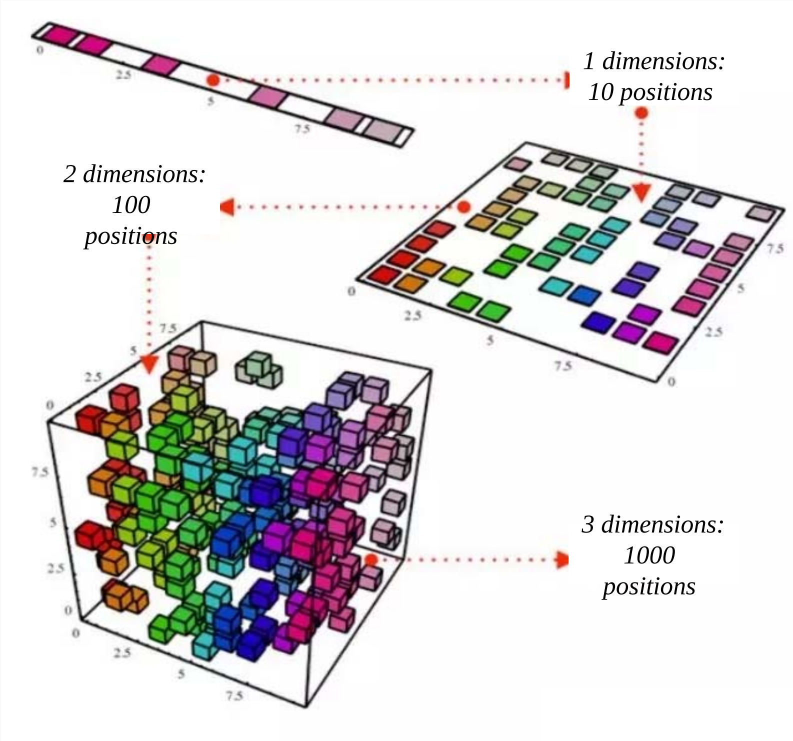




问题引入

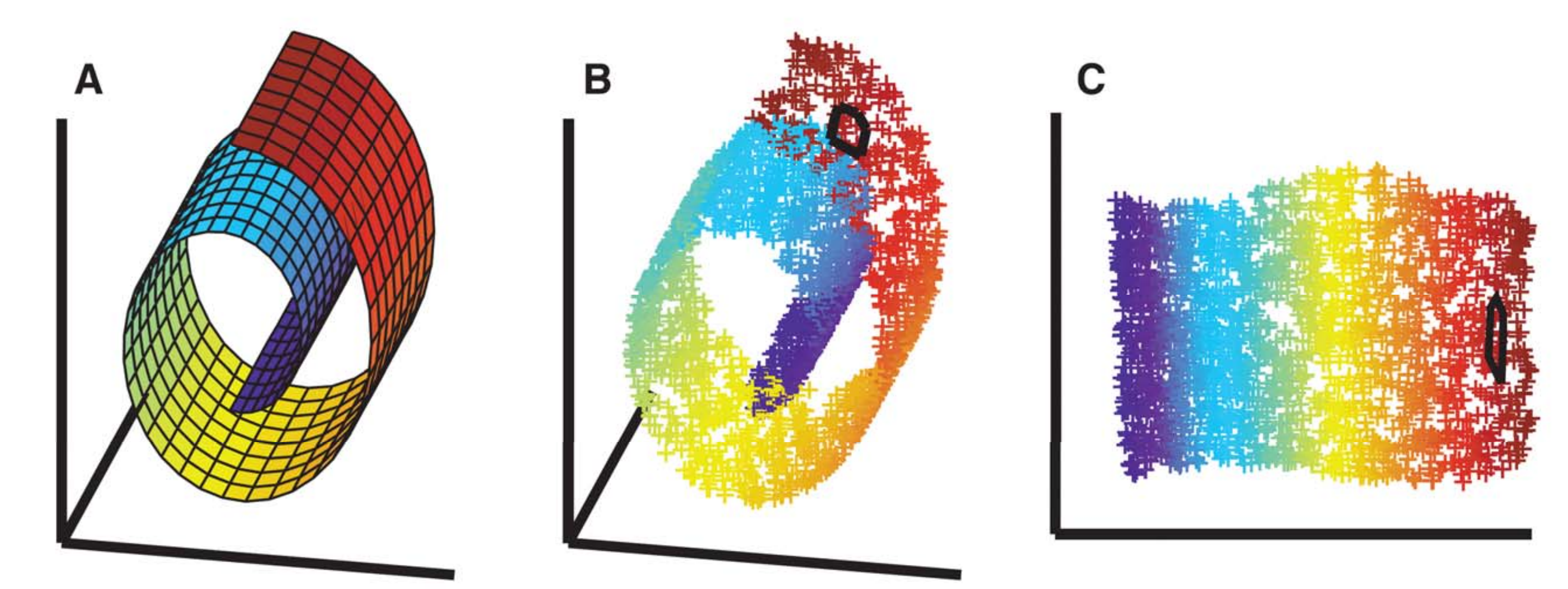
☐ 小样本学习

- 训练集容量小于数据的维度时，通常认为样本数非常小
- 小样本导致特征空间中特征的密度过于低，机器学习算法难以进行有效的学习
- 如何能够将高维数据映射到低维空间，并尽可能保留原始数据的特征？



问题引入：数据压缩

- 数据压缩
  - 受存储条件影响，或在原始的高维空间中包含冗余信息和噪声信息时，需要对数据进行压缩，以更小的维度（容量）更高效的表示数据，同时避免冗余和噪声的影响



§11.2 子空间降维

一、主成分分析（PCA）

二、线性判别分析（LDA）

三、局部保持投影（LPP）

清華大学

低维嵌入

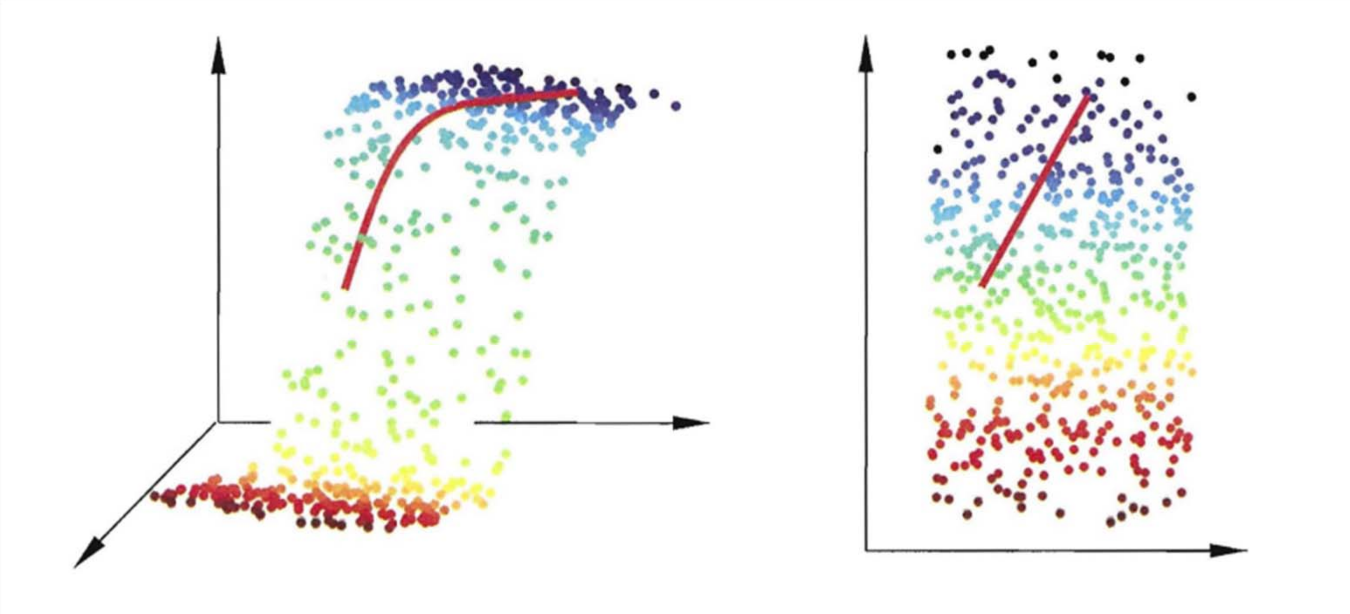
- 思考1：在高维情形，训练样本“密采样”的总样本数目是否可得？
  - 当$\delta=0.001$，样本维数为20 → 需要$(10^3)^{20}=10^{60}$个样本 ！！！
- 思考2：SVM使用核函数数“低维计算，高维表现”的原因

- 如何分析高维数据？
  - 降维(dimension reduction) ：通过某种数学变换将原始高维属性空间转变为低维“子空间”（subspace）→子空间中样本密度大幅提高，距离计算变得更为容易
- 为什么能数据降维？
  - 与学习任务密切相关的也许仅是某个低维分布

低维嵌入

□ “低维嵌入”举例

- 下图三维空间中的样本点，在低维嵌入子空间（为其空间中的曲面）中更容易进行学习



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

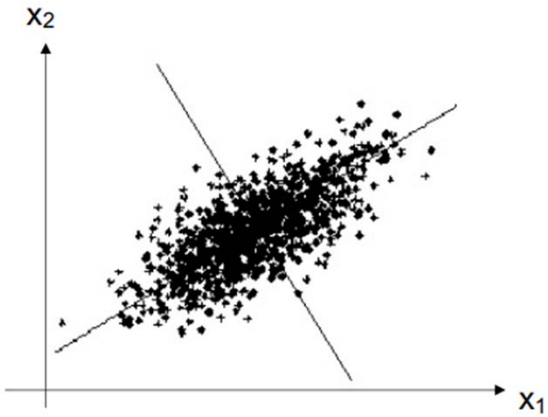
主成分分析

- 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?
  - 超平面需要具有的属性

1. 最近重构性：样本点到这个超平面的距离都足够近
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

（Principal Component Analysis，PCA）

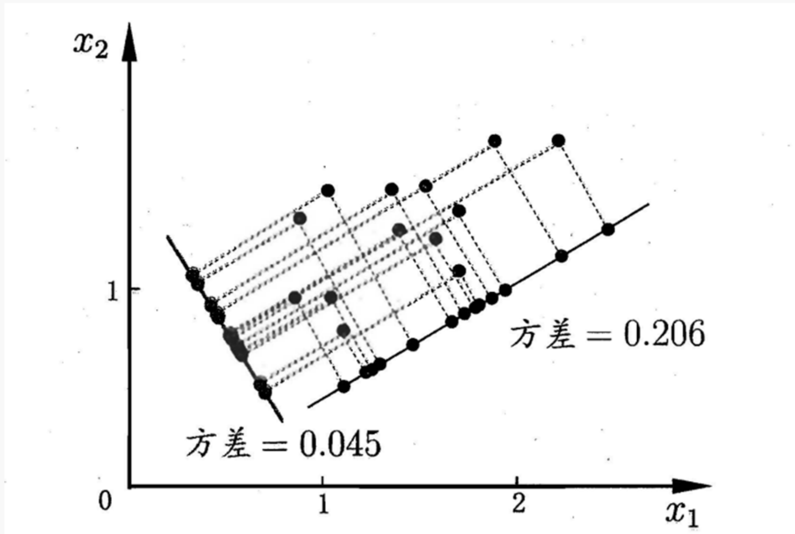
Karl Pearson (1901)



主成分分析

主成分分析目标

- 线性降维方法：通过线性变换，用一组正交向量来表示原特征，新的特征向量是原特征向量的线性组合
- 记原特征向量 $x_1, \dots, x_p$，$\xi_i = \sum_{j=1}^{p} a_{ij} x_j$表示原始特征向量的线性组合，矩阵形式$\xi = A^T x$, 关键是寻找特征变换矩阵$A$来产生新的正交向量$\xi_i$（线性正交变换）



主成分分析

模型求解

- 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

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

$$\min -\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right)$$ → 基于最大重构性

等价

$$\max_{\mathbf{W}} \ \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right)$$
$$\text{s.t.} \ \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I},$$ → 基于最大可分性

算法流程

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 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

主成分分析

- PCA舍弃信息的效果
  - 降维：达到使样本的采样密度增大的目的
  - 去噪：当数据受噪声影响时，最小的特征值所对应的特征向量往往与噪声有关
- SVD与PCA
  - 对样本矩阵X作矩阵的SVD分解：

$$X = U\Sigma V^T$$

$$X^T X = (U\Sigma V^T)^T (U\Sigma V^T) = (V\Sigma U^T)(U\Sigma V^T)$$

$$U^T U = I_m$$

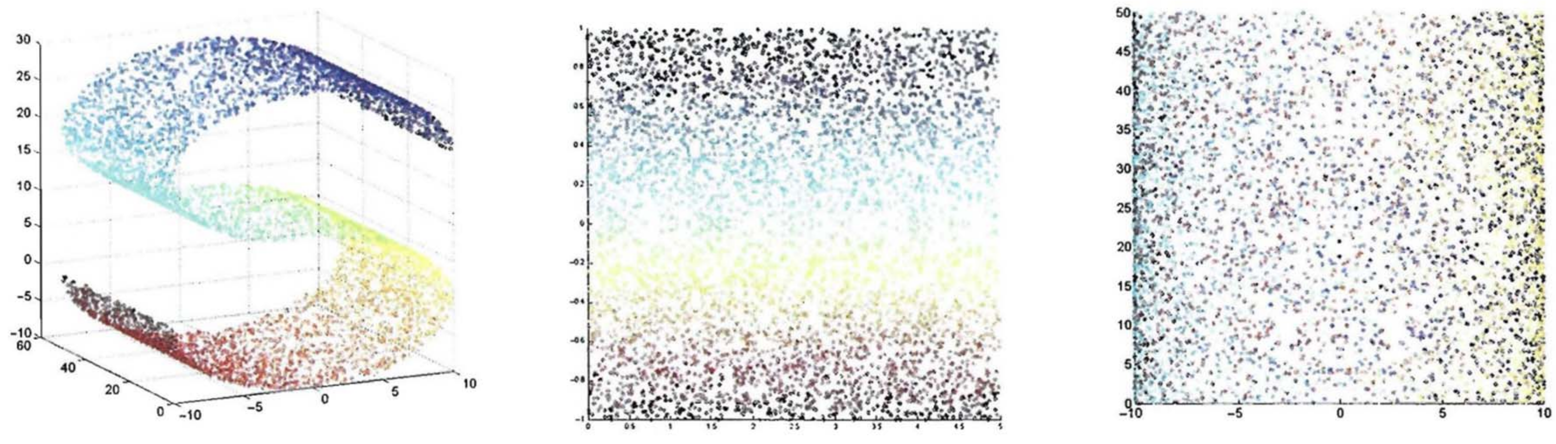
$$X^T X = V\Sigma^T \Sigma V^T$$

特征值分解！

  - 取V^T中对应于k个最大特征值的k行，作为投影向量

核化线性降维

- 思考：样本点从三维空间中的矩形区采样后以S形曲面嵌入到三维空间的情况下（本身就不是线性分布），如何降维？
  - 若直接使用线性降维方法对三维空间观察到的样本点进行降维，则将丢失
  - 本真低维空间：拥有“原本的低维结构”的低维空间



(a) 三维空间中的观察　(b) 本真二维结构　(c) PCA 降维结果

核化线性降维

- 核函数已知
  - 例如：Sigmoid、高斯核等
  - 将样本映射到高维空间，再在高维空间中使用线性降维的方法（如PCA）
- 核函数未知：核主成分分析（Kernelized PCA，KPCA）
  - 只知道如何计算高维空间中的样本内积（大多数情况下）
  - KPCA提出：空间中的任一向量，都可以由该空间中的所有样本线性表示。证明如下：其中z_i为样本点在高维特征空间的坐标，W为高维特征空间新基

$$\left(\sum_{i=1}^{m} z_i z_i^T\right) W = \lambda W$$

$$W = \frac{1}{\lambda}\left(\sum_{i=1}^{m} z_i z_i^T\right) W$$

$$= \sum_{i=1}^{m} z_i \boxed{\frac{z_i^T W}{\lambda}} = \sum_{i=1}^{m} z_i \alpha_i$$

核化线性降维

- 核函数未知：核主成分分析（Kernelized PCA，KPCA）
  - 根据“空间中的任一向量，都可以由该空间中的所有样本线性表示”，将高维特征空间中的投影向量w_i使用所有高维样本点线性表出后，PCA求解

$$\left(\sum_{i=1}^{m}\emptyset(x_i)\emptyset(x_i)^T\right)\ W=\lambda W$$

$$W=\sum_{i=1}^{m}\emptyset(x_i)\ \alpha_i$$

$$有\kappa(x_i,x_j)=\emptyset(x_i)^T\emptyset(x_j)$$

$$化简得，\mathbf{KA}=\lambda A$$

  - 显然，化简结果是特征值分解问题，取最大的d'个特征值对应的特征向量
  - 只需要对和矩阵K进行特征分解，便可得到w_i的系数α
  - 但是：为获得投影后坐标，KPCA需对所有样本求和，因此它的计算开销较大

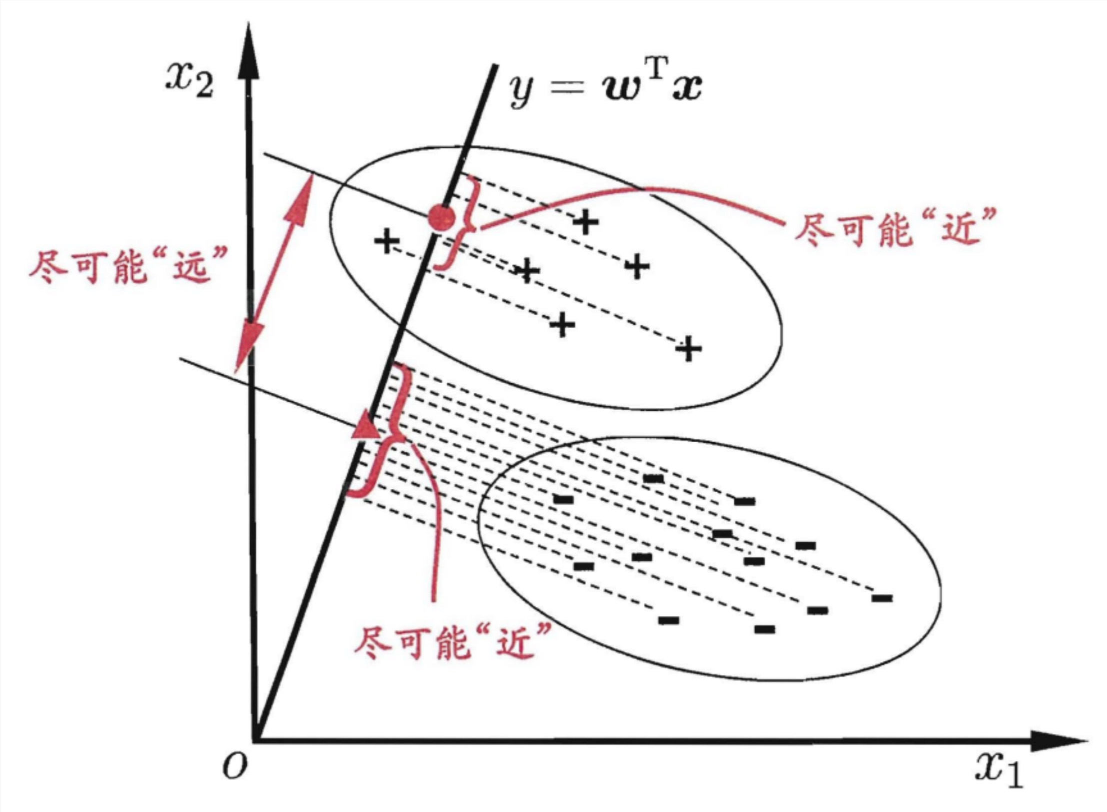
LDA回顾

线性判别分析(Linear Discriminant Analysis，LDA)

- 给定训练样本集，将样本投影到一条使得同类样本的投影点尽可能接近、异类样本投影点尽可能远离。考虑最大化下面的目标：

$$J = \frac{||\boldsymbol{w}^{\mathrm{T}}\mu_0 - \boldsymbol{w}^{\mathrm{T}}\mu_1||_2^2}{\boldsymbol{w}^{\mathrm{T}}\Sigma_0 w + \boldsymbol{w}^{\mathrm{T}}\Sigma_1 w}$$

$$= \frac{\boldsymbol{w}^{\mathrm{T}}(\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}} w}{\boldsymbol{w}^{\mathrm{T}}(\Sigma_0 + \Sigma_1) w}$$



LDA回顾

- 线性判别分析(Linear Discriminant Analysis，LDA)
  - 定义类间散度矩阵和类内散度矩阵

$$\mathbf{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}$$

$$\mathbf{S}_w = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 = \sum_{x \in X_0} (\boldsymbol{x} - \boldsymbol{\mu}_0)(\boldsymbol{x} - \boldsymbol{\mu}_0)^{\mathrm{T}} + \sum_{x \in X_1} (\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^{\mathrm{T}}$$

  - 优化目标重写为：最大化$\boldsymbol{S}_b$和$\boldsymbol{S}_w$的广义瑞利商

$$J = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{w}}$$

分子分母都是关于w的二次项，因此解与w的长度无关，只与其方向有关

LDA回顾

- **线性判别分析**(Linear Discriminant Analysis，LDA)

$$\min_{w} \; -w^T S_b w \qquad \text{s.t. } w^T S_w w = 1$$

  - 拉格朗日乘子法求解：

$$S_b w = \lambda S_w w$$

其中λ是拉格朗日乘子，又有：

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}}$$

不妨令：

$$S_b w = \lambda(\mu_0 - \mu_1)$$

解得：

$$w = S_w^{-1}(\mu_0 - \mu_1)$$

Fisher判别分析

- 核Fisher判别分析(Kernel Fisher Discriminant Analysis, KFDA)，是LDA对应的核方法，将LDA拓展到非线性问题中
- 核心思想为，将样本通过核函数Φ映射到新的特征空间中，再对映射后的样本特征做Fisher判别分析。目标函数变为

$$J=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{\Phi}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\boldsymbol{w}}^{\boldsymbol{\Phi}}\boldsymbol{w}}$$

$$\boldsymbol{S}_{\boldsymbol{b}}^{\boldsymbol{\Phi}}=\left(\mu_0^{\Phi}-\mu_1^{\Phi}\right)\left(\mu_0^{\Phi}-\mu_1^{\Phi}\right)^T \qquad \mu_i^{\Phi}=\frac{1}{n_i}\sum_{j=1}^{n_i}\Phi\left(x_j^i\right)$$

$$\boldsymbol{S}_{\boldsymbol{w}}^{\boldsymbol{\Phi}}=\sum_{x\in X_0}\left(\Phi(x)-\mu_0^{\Phi}\right)\left(\Phi(x)-\mu_0^{\Phi}\right)^{\mathrm{T}}+\sum_{x\in X_1}\left(\Phi(x)-\mu_1^{\Phi}\right)\left(\Phi(x)-\mu_1^{\Phi}\right)^T$$

Fisher判别分析

- 将数据x_i映射到$\Phi(x_i)$，再应用LDA计算的方式计算量巨大，并且在特征空间为无限维时不可解
- 因此利用核函数方法重构特征空间的点乘运算

$$k(x,y)=\phi(x)\cdot\phi(y)$$

- 则有

$$\boldsymbol{w}=\sum_{i=1}^{n}\boldsymbol{\alpha_i\phi(x_i)}$$

$$\boldsymbol{w^T\mu_i^{\phi}}=\frac{1}{n_i}\sum_{j=1}^{n}\sum_{k=1}^{n_i}\boldsymbol{\alpha_j}k\left(x_j,x_k^i\right)=\boldsymbol{\alpha^T M_i}$$

Fisher判别分析

- 令

$$(M_i)_j = \frac{1}{n_i}\sum_{k=1}^{n_i} k\left(x_j, x_k^i\right)$$

- 则优化目标函数可以重写为

$$J = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}$$

其中

$$M = (M_1 - M_0)(M_1 - M_0)^T$$
$$N = K_0\left(I - 1_{n_0}\right)K_0^T + K_1\left(I - 1_{n_1}\right)K_1^T$$

K为核函数$k(x_n, y_m)$组成的矩阵，1_{n_j}的所有元素为$1/n_j$

Fisher判别分析

- 对优化目标函数求导取零解得

$$\boldsymbol{\alpha} = \boldsymbol{N}^{-1}(\boldsymbol{M}_1 - \boldsymbol{M}_0)$$

- 值得注意的是，在实际应用中，$\boldsymbol{N}$往往是奇异的，因此需要加上一个极小单位矩阵

$$\boldsymbol{N}_\epsilon = \boldsymbol{N} + \epsilon \boldsymbol{I}$$

- 在解得$\boldsymbol{\alpha}$后，对一个新给的数据点，其Fisher投影为

$$y(x) = \big(w \cdot \phi(x)\big) = \sum_{i}^{n} \alpha_i k(x_i, x)$$

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 与Isomap和LLE同为**流形学习**方法
    - 高维空间的低维流形在局部仍保持欧氏空间的性质
  - 不同之处
    - Isomap利用测地线距离，尝试保持**任意两样本点**之间的**距离关系**
    - LLE尝试保持**邻域内样本**之间的**线性关系**
    - LE尝试保持**邻域内样本**之间的**距离关系**
- 局部保持投影（Locality Preserving Projections，LPP）
  - LE是**非线性映射**，且无法应用于未见数据
  - LPP以LE为基础，限制LE为**线性映射**，可通过线性映射处理未见数据

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 问题表示：给定m个样本$x_1, x_2, \dots, x_m \in \mathbb{R}^n$，找到其低维近似$y_1, y_2, \dots, y_m \in \mathbb{R}^l\ (l \ll n)$，并保持原始高维数据在局部的亲疏关系。即如果$x_i$和$x_j$相距很近，则它们的低维近似$y_i$和$y_j$也足够近
  - 目标函数：

$$\min_{y} \ \frac{1}{2}\sum_{i,j}\left\|y_i - y_j\right\|^2 w_{ij}$$

    - w_{ij}是样本点邻接矩阵W的元素，反映了高维空间中x_i和x_j的亲疏关系
    - $\left\|y_i - y_j\right\|^2$表示低维空间中样本点之间的欧氏距离（流形的局部）
    - w_{ij}越大，表示x_i和x_j关系越密切，此时要求$\left\|y_i - y_j\right\|^2$越小

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 邻接矩阵W反映了原始的高维数据在其邻域内的亲疏关系
    - 邻域 $Neighbor(x_i)$的确定方式：
      - ϵ–近邻：$x_j \in Neighbor(x_i)$，如果$\|x_i - x_j\|^2 < \epsilon$
      - k近邻：$x_j \in Neighbor(x_i)$，如果x_j是x_i的k近邻，或者x_i是x_j的k近邻
    - 亲疏关系的量化：
      - 热核法（Heat Kernel）

$$w_{ij} = exp\left(-\frac{\|x_i - x_j\|^2}{t}\right) \ \ if \ \ x_j \in Neighbor(x_i) \ \ else \ \ 0$$

      - 简单法（Simple Minded）

$$w_{ij} = 1 \ \ if \ \ x_j \in Neighbor(x_i) \ \ else \ \ 0$$

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 一些符号的定义
    - 前面的构造过程保证W为对称阵
    - D为对角阵，L为对称阵

$$W=\begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mm} \end{bmatrix}, \qquad D=\begin{bmatrix} \sum_{j=1}^{m} w_{1j} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{j=1}^{m} w_{mj} \end{bmatrix}, \qquad L=D-W$$

    - $Y=[y_1,y_2,\dots,y_m]$

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 目标函数化简

$$
\begin{aligned}
\frac{1}{2}\sum_{i,j}\left\|y_i - y_j\right\|^2 w_{ij} &= \frac{1}{2}\sum_{i,j}\left(y_i^T y_i + y_j^T y_j - 2y_i^T y_j\right)w_{ij} \\
&= \frac{1}{2}\left(\sum_{i,j} y_i^T y_i w_{ij} + \sum_{i,j} y_j^T y_j w_{ij} - 2\sum_{i,j} y_i^T y_j w_{ij}\right) \\
&= \sum_i \left(\sum_j w_{ij}\right) y_i^T y_i - \sum_{i,j} y_i^T y_j w_{ij} \\
&= \sum_i d_{ii}\, y_i^T y_i - \sum_{i,j} y_i^T y_j w_{ij} \\
&= \mathrm{tr}\left(YDY^T\right) - \mathrm{tr}\left(YWY^T\right) \\
&= \mathrm{tr}\left(Y(D-W)Y^T\right) = \mathrm{tr}\left(YLY^T\right)
\end{aligned}
$$

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 目标函数等价于

$$\min_{y} \frac{1}{2}\sum_{i,j}\left\|y_i - y_j\right\|^2 w_{ij} \leftrightarrow \min_{y} \operatorname{tr}\left(YLY^T\right)$$

  - 问题：可以无限地减小$Y = [y_1, y_2, \dots, y_m]$的模长，取得平凡解0
    - 与支持向量机的决策超平面$y = w^T x + b$的尺度问题类似
    - 解决方案：为$Y = [y_1, y_2, \dots, y_m]$加入尺度约束，即$\operatorname{tr}\left(YDY^T\right) = 1$
  - 最终的优化问题表述为：

$$\min_{y} \operatorname{tr}\left(YLY^T\right)$$
$$\text{s.t. } \operatorname{tr}\left(YDY^T\right) = 1$$

LE和LPP

- 拉普拉斯特征映射（Laplacian Eigenmaps，LE）
  - 优化问题求解：拉格朗日乘子法

$$\mathcal{L}(Y,\lambda)=\mathrm{tr}\left(YLY^T\right)-\lambda\left(\mathrm{tr}\left(YDY^T\right)-1\right)$$

令 $\frac{\partial\mathcal{L}}{\partial Y}=2(YL-\lambda YD)=0$ 得

$$YLD^{-1}=\lambda Y$$

$$D^{-1}LY^T=\lambda Y^T$$

故 Y^T 的列向量为 $D^{-1}L$ 的特征向量，λ 为对应的特征值

  - 由 $YLY^T=\lambda YDY^T$ 得 $\mathrm{tr}\left(YLY^T\right)=\lambda\,\mathrm{tr}\left(YDY^T\right)=\lambda$
    - 故应该取 $D^{-1}L$ 最小的若干个特征值对应的特征向量作为 Y^T

LE和LPP

从LE到局部保持投影（LPP）

- LE的问题：对于新样本数据的加入，该模型并不能给出其子空间的表示形式
- LPP对LE的改进：将原本的隐式非线性映射变为显式线性映射

$$Y = A^T X, \qquad A \in \mathbb{R}^{n \times l}$$

从而可以应用于新的未见样本

局部保持投影（LPP）推导

- 优化问题：

$$\min_A \operatorname{tr}\left(A^T X L X^T A\right)$$
$$\text{s.t. } \operatorname{tr}\left(A^T X D X^T A\right) = 1$$

- 下面仍使用拉格朗日乘子法求解该优化问题

LE和LPP

局部保持投影（LPP）推导

- 拉格朗日乘子

$$\mathcal{L}(A,\lambda) = \mathrm{tr}\left(A^T XLX^T A\right) - \lambda\left(\mathrm{tr}\left(A^T XDX^T A\right) - 1\right)$$

令 $\frac{\partial \mathcal{L}}{\partial A} = 2\left(XLX^T A - \lambda XDX^T A\right) = 0$ 得

$$XLX^T A = \lambda XDX^T A$$

当 XDX^T 可逆时

$$(XDX^T)^{-1} XLX^T A = \lambda A$$

故 A 的列向量为 $(XDX^T)^{-1} XLX^T$ 的特征向量，λ 为对应的特征值

- 由 $A^T XLX^T A = \lambda A^T XDX^T A$ 得 $\mathrm{tr}\left(A^T XLX^T A\right) = \lambda\, \mathrm{tr}\left(A^T XDX^T A\right) = \lambda$
- 故应该取 $(XDX^T)^{-1} XLX^T$ 最小的若干个特征值对应的特征向量作为 A

§11.3 流形学习

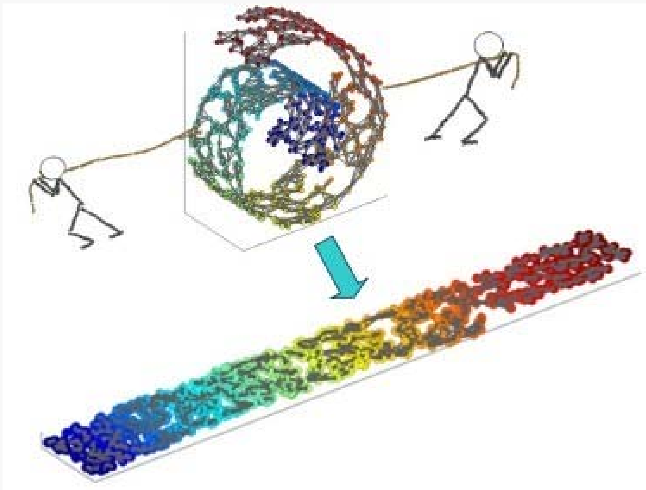
一、等距特征映射（IsoMap）

二、局部线性嵌入（LLE）

三、多维尺度变换（MDS）

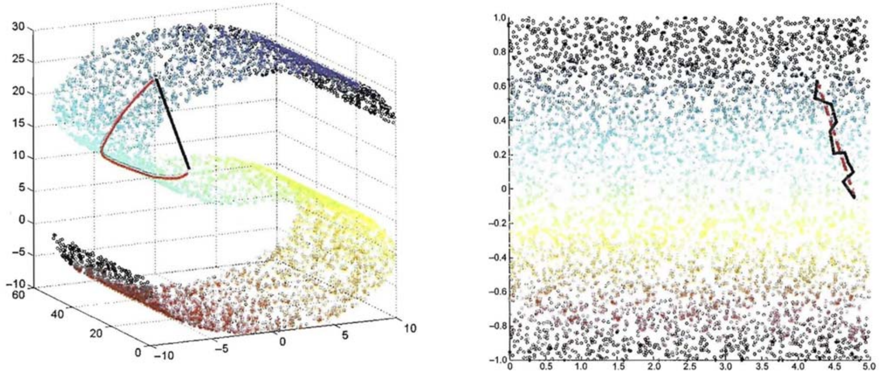
问题引入

- 流形学习
  - 流形：在局部与欧式空间同胚的空间，即在局部与欧式空间具有相同的性质，能用欧氏距离计算样本之间的距离
  - 借助拓扑流形概念的降维方法
  - 直观上：一个流形可以看作 一个d维的空间在一个D维的空间中（D>d）被扭曲之后的结果



等度量映射

- 等度量映射(Isometric feature mapping, Isomap)
  - 出发点：当数据在高维空间中存在某种复杂的结构分布时，直接用欧式距离有时候不能反映数据间的相互关系。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
  - 距离度量：利用流形在局部上与欧式空间同胚的性质，可以使用近邻距离来逼近测地线距离 → 最短路径问题(Dijkstra, Floyd算法等)



(a) 测地线距离与高维直线距离　　(b) 测地线距离与近邻距离

等度量映射

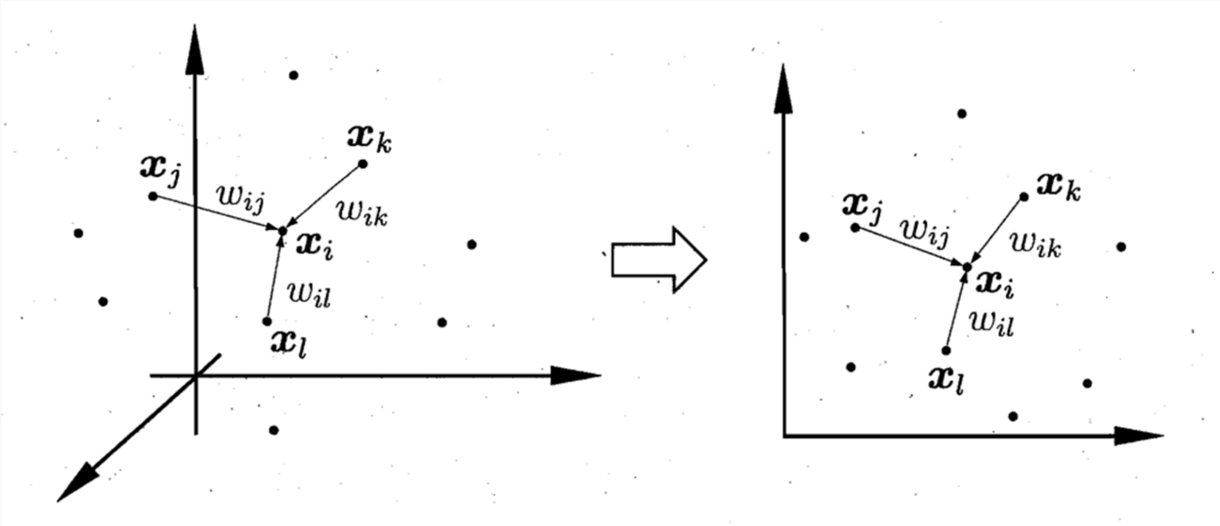
- 等度量映射(Isometric feature mapping, Isomap)
  - 算法思想：
    - 当样本分布较密集的区域，假定样本空间结构可在该局部用欧式距离度量
    - 对于两个相距较远的点，寻找一系列两两相邻的样本点构成连接二者的路径，用最短路径上局部距离之和度量二者距离
    - 得到重新定义样本“距离矩阵”后，用MDS映射到低维空间
  - 近邻图构建常见问题：
    - 邻域范围指定过大：“短路问题”，本身距离很远却成了近邻
    - 邻域范围指定过小：“断路问题”，有些样本点无法可达

局部线性嵌入

- 局部线性嵌入(Locally Linear Embedding, LLE)
  - 不同于Isomap，LLE算法关注邻域内的线性关系，假定样本xi的坐标可以通过它的邻域样本线性表出：

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

  - 线性关系保持不变，即邻域重构系数不变



局部线性嵌入

- 局部线性嵌入(Locally Linear Embedding, LLE)
  - 算法思想
    - 根据近邻关系计算出所有样本的邻域重构系数w，目标函数为重建损失

$$min_{w_1,w_2,\ldots,w_m} \sum_{i=1}^{m} \left\| x_i - \sum_{j\in Q_i} w_{ij} x_j \right\|_2^2$$

$$\text{s.t.} \ \sum_{i=1}^{m} w_{ij} = 1$$

    - 根据邻域重构系数不变，去求解低维坐标

令 $Z = (z_1, z_2, \ldots, z_m) \in R^{d' \times m}, (W)_{ij} = w_{ij}$， 则低维空间的目标函数可写为：

$$\sum_Z tr\left(ZMZ^T\right), \text{s.t.} \ ZZ^T = I$$

    - M特征值分解后最小的d’个特征值对应的特征向量组成Z

多维尺度变换

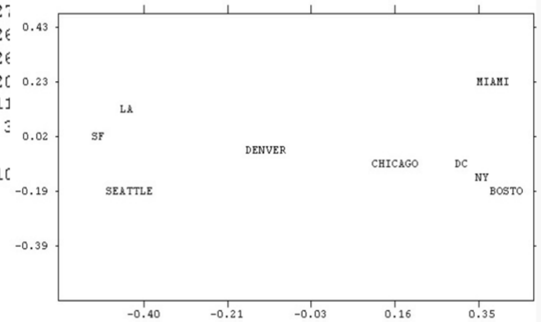
MDS基本出发点

- 根据**样本之间的距离关系**或**不相似度关系**在低维空间里生成对样本的表示
- 把样本之间的距离关系或不相似关系在二维或三维空间里表示出来

方法

- 给定距离，求（相对）坐标
- 例：地图坐标（右图）

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	27	
3	DC	429	233	0	1075	671	2684	2799	26	
4	MIAMI	1504	1308	1075	0	1329	3273	3053	26	
5	CHICAGO	963	802	671	1329	0	2013	2142	20	
6	SEATTLE	2976	2815	2684	3273	2013	0	808	11	
7	SF	3095	2934	2799	3053	2142	808	0	3	
8	LA	2979	2786	2631	2687	2054	1131	379		
9	DENVER	1949	1771	1616	2037	996	1307	1235	10	



多维尺度变换

MDS算法分析

- 定义：距离矩阵矩阵$D \in \mathbb{R}^{m\times m}$；低维空间中的表示$Z \in R^{d'\times n}$，低维空间维数$d' < d$；任意两个样本在低维空间中的欧式距离等于原始空间中的距离，即

$$\left\|z_i - z_j\right\| = Dist(ij)$$

- 目标：根据高维任意两点间距离$dist_{ij}^2$计算低维空间距离矩阵B

令$B = Z^T Z \in R^{m\times m}$，其中B为降维后样本的内积矩阵，$b_{ij} = {z_i}^T z_j$

$$\begin{aligned} dist_{ij}^2 &= \| z_i \|^2 + \| z_j \|^2 - 2{z_i}^T z_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

多维尺度变换

MDS算法分析

- 结合“降维后的样本坐标矩阵Z中心化”特性，推导矩阵B

$$B=\begin{bmatrix} z_1 \\ \dots \\ z_m \end{bmatrix} * \begin{bmatrix} z_1 & \dots & z_m \end{bmatrix} = \begin{bmatrix} z_1z_1 & z_1z_2 & \dots & z_1z_m \\ z_2z_1 & z_2z_2 & \dots & z_2z_m \\ \dots & \dots & \dots & \dots \\ z_mz_1 & z_mz_2 & \dots & z_mz_m \end{bmatrix}$$

和为零向量

和为零向量

$$\sum_{i=1}^{m} dist_{ij}^2 = \mathrm{tr}(\mathbf{B}) + m b_{jj} \,,$$

$$\sum_{j=1}^{m} dist_{ij}^2 = \mathrm{tr}(\mathbf{B}) + m b_{ii} \,,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} dist_{ij}^2 = 2m \, \mathrm{tr}(\mathbf{B}) \,,$$

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

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

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

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

多维尺度变换

MDS算法过程描述

- S1：根据公式，求解$dist_{i\cdot}^2$ $dist_{\cdot j}^2$和$dist_{\cdot\cdot}^2$
- S2：计算矩阵 B
- S3：对矩阵B做特征值分解
  - 对矩阵B做特征值分解(eigenvalue decomposition)，$B = V\Lambda V^T$，其中$\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$为特征值构成的对角矩阵， $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_d$)，V为特征向量矩阵.假定其中有$d^*$个非零特征值，它们构成对角矩阵$\Lambda_* = diag(\lambda_1, \lambda_2, \ldots, \lambda_{d^*})$，令$V_*$表示相应的特征向量矩阵,则Z可表达为

$$Z = \Lambda_*^{1/2} V_*^T \in R^{d^* \times m}$$

- S4：取$\tilde{\Lambda}$为d'个最大特征值所构成的对角矩阵，$\tilde{V}$为相应的特征向量矩阵
- 输出：矩阵Z的每行是一个样本的低维坐标

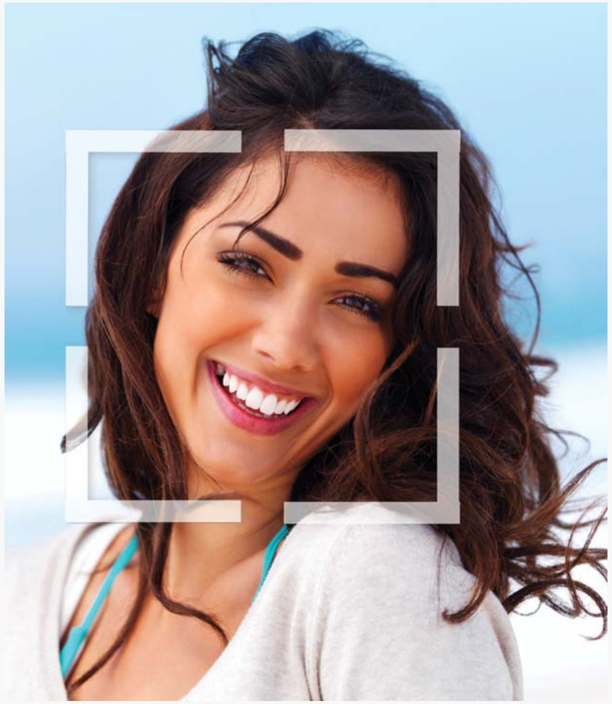
§11.4 应用举例

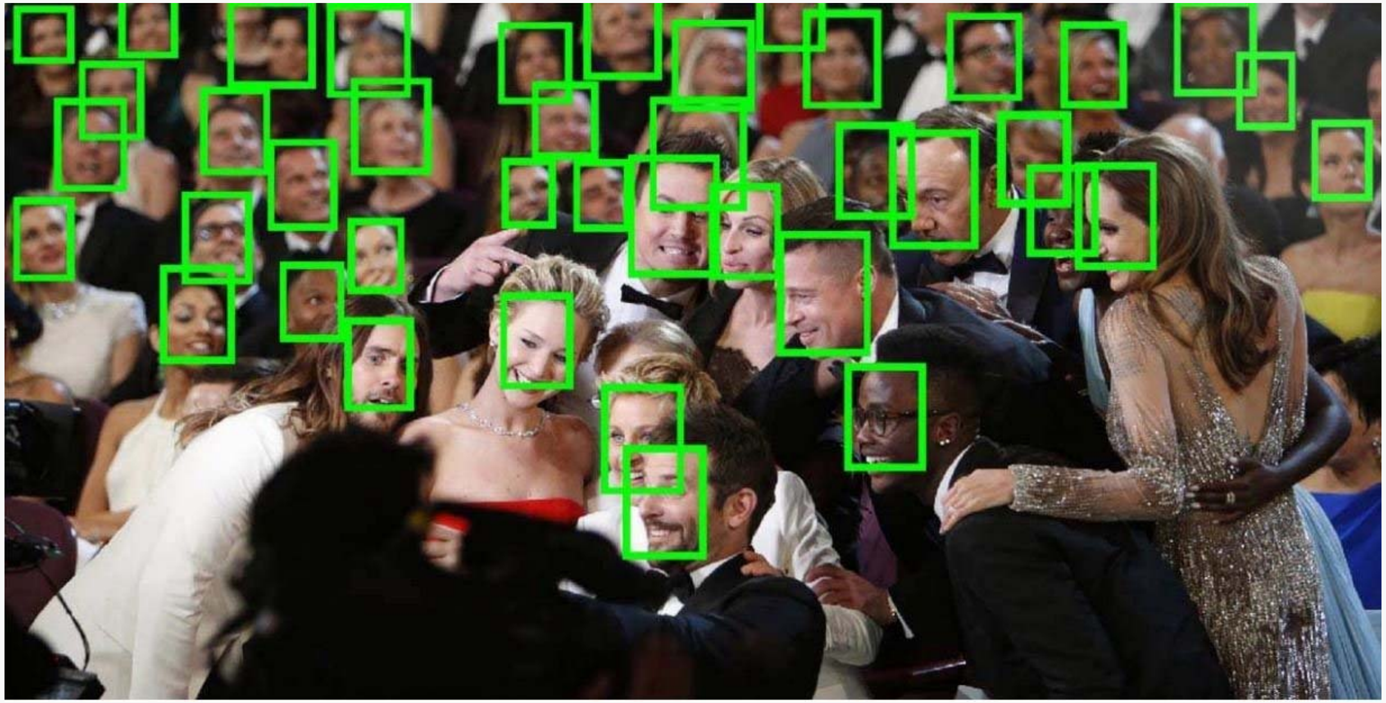
一、人脸特征表示

人脸检测

人脸检测

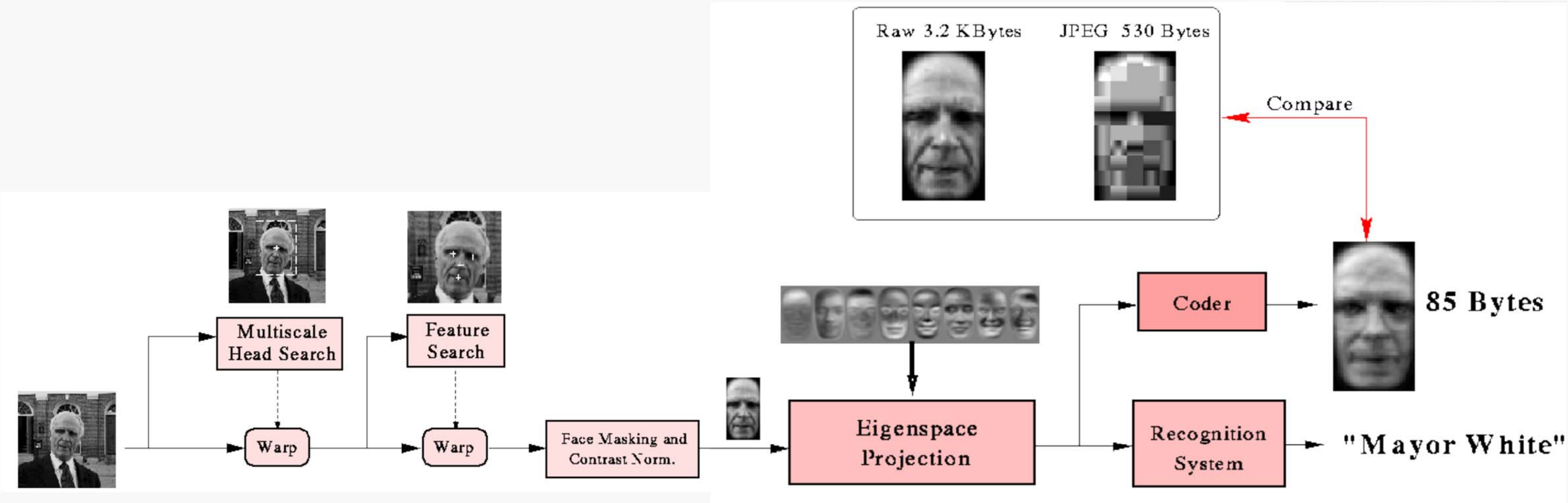
- **任务定义：** 有无人脸，具体位置
- **困难：** 光照、人脸朝向、阴影、尺寸等





人脸特征表示：EigenFaces

- PCA在人脸识别中的应用举例
  - S1：对人脸数据集预处理（图像归一化和裁剪）
  - S2：本征脸提取、表示和基于本征脸的分类

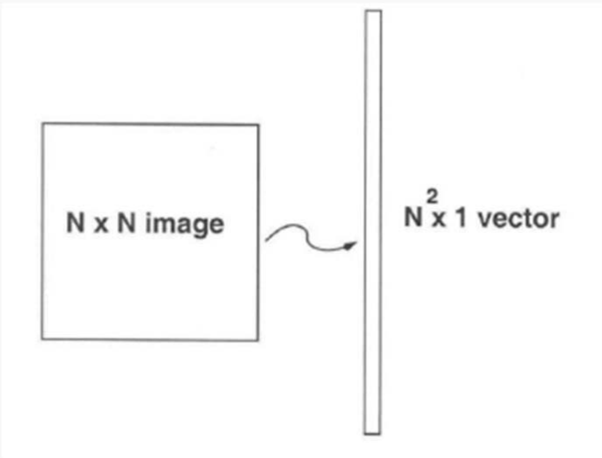


M. Turk & A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience, vol.3, no.1, pp.71-86, 1991.

人脸特征表示：EigenFaces

实现算法

- 样本集$X_i \in R^{N^2}$, i=1, ⋯, M，用PCA进行降维
- 总体散布矩阵$\sum = \frac{1}{M}\sum_{i=0}^{M-1}(X_i-\mu)(X_i-\mu)^T = \frac{1}{M}XX^T$
- 存在问题：$N^2 \times N^2$维矩阵，求其正交归一的本征向量，但计算困难



人脸特征表示：EigenFaces

实现算法

- 解决方法：

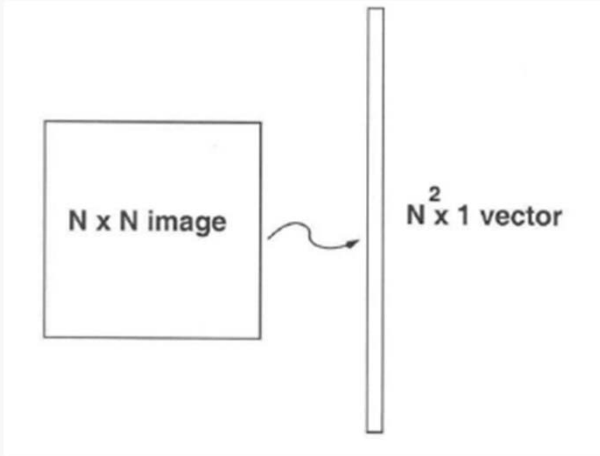
考察M×M（M为样本数，M<<$N^2 \times N^2$)维矩阵

$R = X^T X$

- 其特征方程是$X^T X v_i = \lambda_i v_i$
- 两边同乘以X： $X X^T X v_i = \lambda_i X v_i$
- $\sum X v_i = \lambda_i X v_i$
- 记$u_i = X v_i$，有$\sum u_i = \lambda_i u_i$

所以，矩阵$X^T X$和$X X^T$具有相同的特征值，而特征相关具有关系

$$u_i = X \mathrm{v_i}$$



人脸特征表示：EigenFaces

实现算法

易求得，$\sum$的归一化的本征向量是

$$u_i = \frac{1}{\sqrt{\lambda_i}} X v_i \text{ ，} \quad i=1,2,\cdots M$$

- 注意，因为矩阵$\sum$的秩最多为M，所以最多只有M个本征值和本征向量
- 每一个本征向量仍然是一个N^2维向量，即N×N维图像，仍然具有类似人脸的样子，因此被称作“本征脸”（eigenfaces）
- 按照本征值从大到小排列

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$$

- 并从前向后取相应的本征脸，即构成对原图像的最佳的降维表示

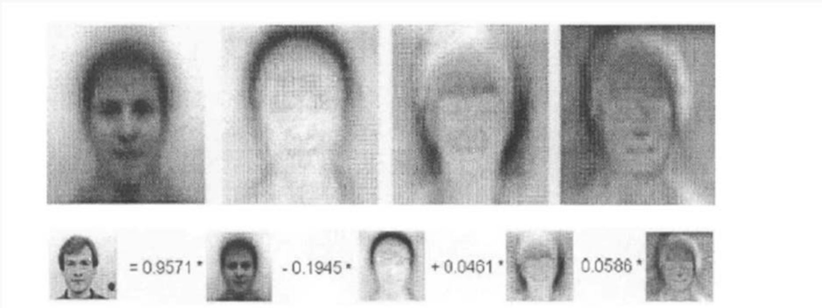
人脸特征表示：EigenFaces

实现算法

- 原图像可以表示成特征脸的线性组合（在特征脸空间中的点）

$$y_i = U^T x_i,\quad \text{i=1,}\cdots\text{M}$$

$$\hat{x}_i = \hat{U}\hat{y}_i^{\,T}$$ 其中，$\hat{y}_i$为d维（d<p），$\hat{U}$为p×d维



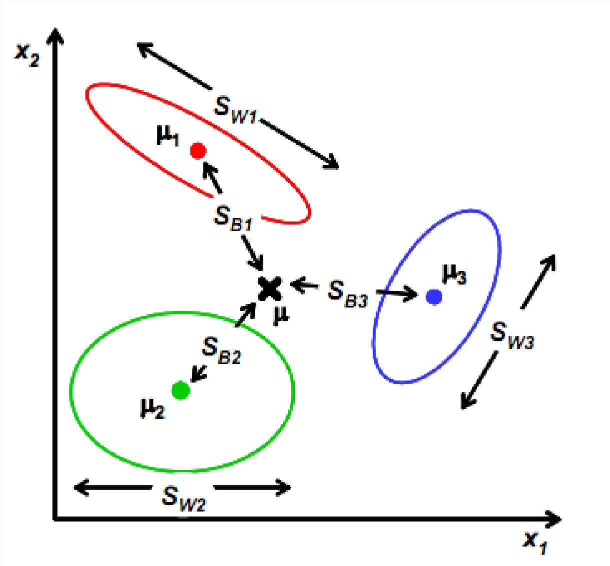
- 比如选取前k个特征向量，时

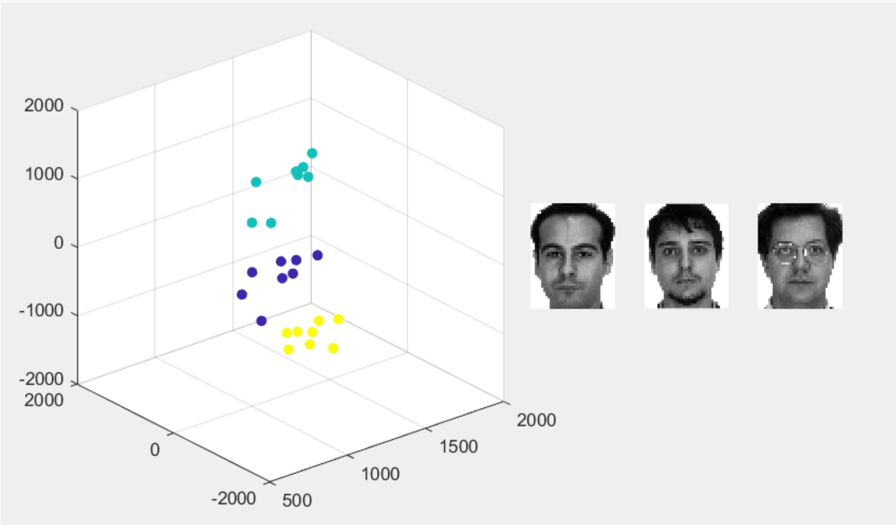
$$\sum_{i=0}^{k-1} \lambda_i \,/\, \sum_{i=0}^{M-1} \lambda_i \geq \alpha$$

- 比如α=99%即可以保持原样本99%的信息
- 对原图像的表示$\hat{x}_i - \mu = \sum_{j=1}^{k} y_{ij} u_j$

人脸特征表示：FisherFaces

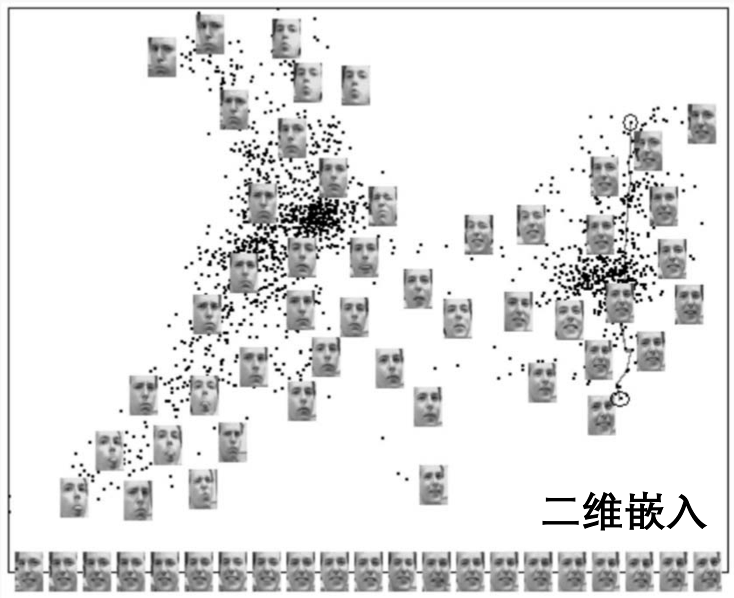
- 对比EigenFace
  - EigenFace基于PCA降维，FisherFace基于LDA降维
  - PCA是无监督方法，目标是最小化类内散度、最大化类间散度
  - LDA是有监督方法，目标是更好的分类
  - 在小数据上面两种方法相差不大，在大数据上LDA有着更明显的优势

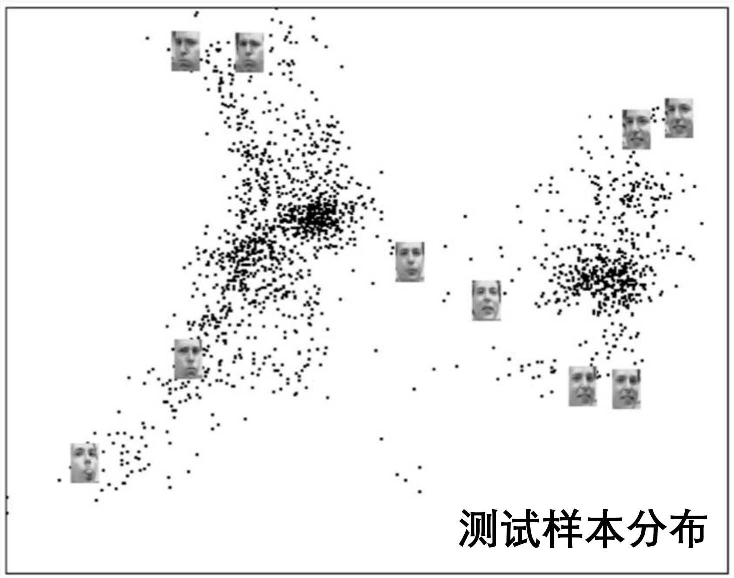




人脸特征表示：LaplacianFaces

- 对比EigenFaces与FisherFaces
  - EigenFace和FisherFace只关注人脸空间的欧几里得结构
  - LPP则利用LPP降维，获取能保留人脸关键局部信息的子空间
  - 可以看到，测试样本都能找到表示其内在属性（如位姿、表情）的最优坐标





思考题

- 降维中涉及的投影矩阵通常要求是正交的，试述正交、非正交投影矩阵用于降维的优缺点
- 答：正交有两个好处：
  - 1. 降维和重构变换计算方便。比如，在PCA中，设新坐标系中基矢为$\{w_1, w_2, \ldots, w_{d'}\}$，样本$x$在新坐标系中的坐标为$z$，通过$z$重构的样本坐标为$\hat{x} = Wz$。现在假设$\{w_i\}$是一组线性无关的基矢，但是未必正交归一，那么为了满足“最近重构”，需要$min_z|\hat{x} - x| = min_z|Wz - x|$，该问题有解析解：$z^* = (W^TW)W^Tx$。如果$\{w_i\}$彼此正交归一，便有$W^TW = I$，于是$z^* = W^Tx$。
  - 2. 变换后的Z的不同坐标之间是“去相关”的。我们已经知道，在PCA中，变换后，在新的特征空间中，不同特征之间是“不相关”的，也就是协方差矩阵ZZ^T是对角化的，非对角元素为零。

思考题

- 现在，假设有一组基矢$\{w_i'\}$不是正交化的，设为W'，它可以由正交化的W线性表出：$W' = WA$，从“最近重构”的角度，两者的重构效果应该等同：$W'z' = Wz$，于是$z' = A^{-1}$，$Z'Z^T = A^{-1}ZZ^TA^{-1^T}$，此时协方差的非对角元就未必为零了。
- 其实至于不同特征之间“去相关”有什么好处，现在没有相关的实践应用，不好体会，留待以后有所体会的时候再回来补充吧。
- 关于正交的缺点方面，大概也就是“去相关”后的缺点吧，某些情况下也许特征之间完全去相关未必是好事。同样需要慢慢实践、体会。现在想到的例子，比如：一个人的身高和体重是正相关的，通过PCA方法大概可以得到“身体年龄”和“肥瘦度”两个相互独立的特征，但是，或许在一个特定任务中，直接用身高和体重两个特征更容易一些。