模式识别

主讲：崔林艳&邹征夏

单位：宇航学院

开课专业：飞行器控制与信息工程

第五章 特征选择与特征提取

CONTENTS PAGE

第五章 特征选择与特征提取

CONTENTS PAGE

5.1 基本概念

5.1.1 特征概念回顾

5.1.2 识别系统回顾

5.1.3 研究特征的原因

5.1 基本概念

5.1.1 特征概念回顾

- 特征和特征值：特征是可以用来体现类别之间相互区别的某个或某些数学测度，也称作属性，测度的值称为特征值。
- 特征向量和特征空间：由被识别的对象（样本）确定一组基本特征，组成该样本的特征向量。样本的特征构成了样本的特征空间，空间的维数就是特征的个数，而每一个样本就是特征空间中的一个点。

5.1 基本概念

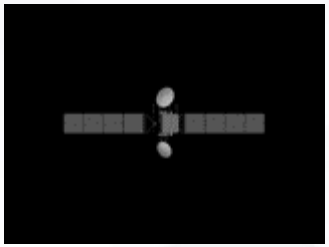
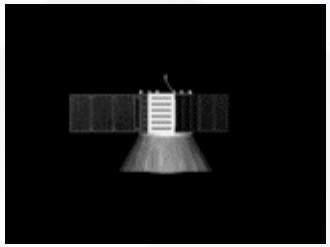
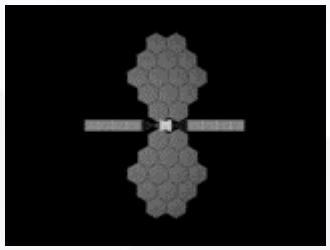
5.1.1 特征概念回顾

- 特征形成与计算：由被识别的对象（即样本）确定一组基本特征，组成该样本的特征向量。
  - 当识别对象是波形或图像时，这些特征可以是计算得到的。
  - 当识别对象是实物或某种过程时，这些特征可以是由仪表或传感器测量出来的。

5.1 基本概念

5.1.1 特征概念回顾

- 原始特征：根据应用领域相关知识决定采用的特征。
- 例如：实验用的卫星图像大小160 x 120，如果采用全部灰度值的话，原始特征即为19200维。
- 如果改为计算卫星的面积、周长、形状、纹理等有效特征，原始特征可能不超过100维。

5.1 基本概念

5.1.1 特征概念回顾

- 特征选择与提取：
  - 由目视识别经验和训练样本确定一组能够体现类别间相互区别的测度（即原始特征）。
  - 从原始特征中尽量挑选出一些最有效的特征，以达到降低特征空间维数的目的。
  - 直接搜索挑选出最优组合——特征选择
  - 数学变换优化成另一组特征——特征提取

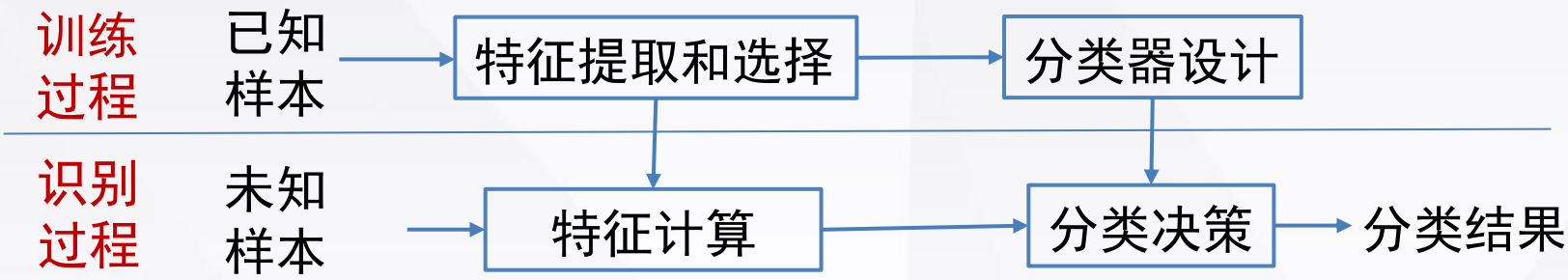
5.1 基本概念

5.1.1 特征概念回顾

- 特征选择与提取：
  - 特征选择（feature selection）是指从原始特征中挑选出一组最有代表性、分类性能好的特征。
  - 特征提取（feature extraction）则是通过映射（或变换）的方法将高维特征映射到低维空间，在低维空间表示样本。
  - 相似点：达到的效果一样，即试图去减少特征数据集中的属性(或者称为特征)的数目
  - 方式方法不同：特征提取主要是通过属性间的关系，如组合不同的属性得到新的属性，改变了原来的特征空间；而特征选择的方法是从原始特征数据集中选择出子集，是一种包含的关系，没有更改原始的特征空间。

5.1 基本概念

5.1.2 识别系统回顾



特征提取、特征选择、特征计算是模式识别系统的关键过程。

5.1.3 研究特征的原因

- 模式识别系统的成败，首先取决于所采用的特征集能否较好地反映有待于研究的分类问题。
- 如何设计和选用特征是设计模式识别系统最关键的步骤之一。
- 特征选择和特征提取是试图降低特征空间的维数、减少系统设计和使用难度的关键环节。

第五章 特征选择与特征提取

CONTENTS PAGE

5.2 图像特征

5.2.1 颜色特征

5.2.2 形状特征

5.2.3 纹理特征

5.2.4 空间关系特征

5.2 图像特征

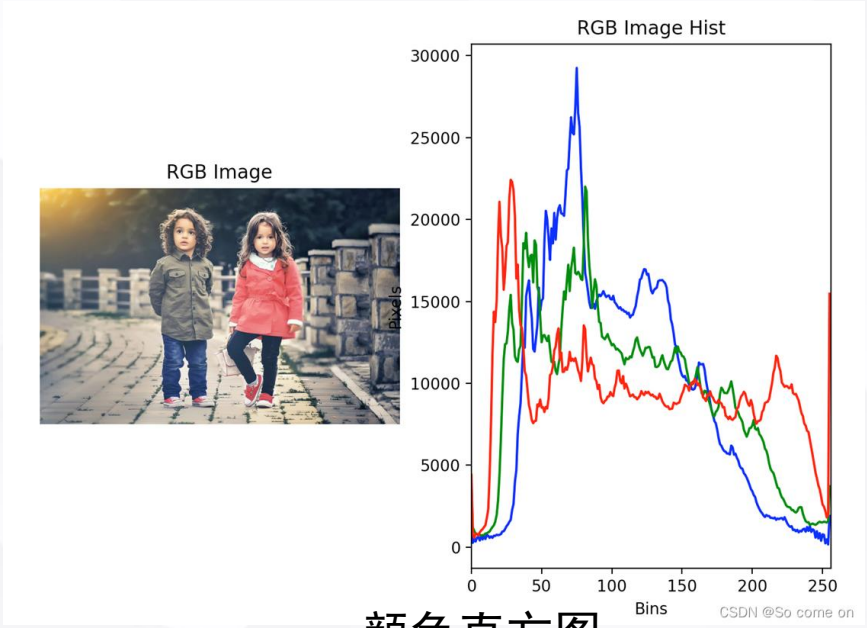
5.2.1 颜色特征

- 颜色特征描述图像或图像区域对应景物的表面性质。
- 颜色特征是基于像素值的特征。
- 颜色特征是一种全局特征，通常对图像或图像区域的方向、大小等变化不敏感。颜色特征通常不能很好的描述图像中对象的局部特征。

5.2 图像特征

5.2.1 颜色特征

- 颜色直方图：描述一幅图像中颜色的全局分布，无法描述图像中颜色的局部分布及每种色彩所处的空间位置。
- 是颜色全局分布的统计结果。
  - 灰度直方图
  - 颜色直方图



颜色直方图

5.2 图像特征

5.2.1 颜色特征

- 颜色矩：是一种简单有效的颜色特征表示方法。
- 矩来自于力学中的概念，用于表征物质的空间分布。
- 对于图像分析与识别，图像的灰度对应于力学中物体的质量。
  - 一阶矩（均值）
  - 二阶矩（方差）
  - 三阶矩（偏度）
  - 四阶矩（峰度）

5.2 图像特征

5.2.1 颜色特征

P_{ij}表示彩色图像第j个像素的第i个颜色分量
N表示图像中的像素个数

a）一阶颜色矩（均值）

$$E_i = \sum_{N}^{j=1} \frac{1}{N} P_{ij}$$

反映图像的整体明暗程度。值越大，图像越亮。

b）二阶颜色矩（方差）

$$\sigma_i = \sqrt{(\frac{1}{N} \sum_{N}^{j=1} (P_{ij} - E_i)^2)}$$

反映图像的颜色分布范围，值越大，颜色分布范围越广。

c）三阶颜色矩（偏度）

$$S_i = \sqrt[3]{(\frac{1}{N} \sum_{N}^{j=1} (P_{ij} - E_i)^3)}$$

反映图像颜色分布的对称性。

当S_i=0时，图像的颜色分布是对称的；
当S_i<0时，颜色分布左偏或负偏；
当S_i>0时，颜色分布右偏或正偏；

5.2 图像特征

5.2.2 形状特征

- 形状特征用来表示图像中感兴趣的区域
- 形状特征的主要形式：
  - 区域特征：图像形状的整个区域
  - 轮廓特征：图像形状的外边界

5.2 图像特征

5.2.2 形状特征

① 区域特征：

- 几何参数特征
  - 面积
  - 周长
  - 半径（直径）
  - 长度
  - 宽度
  - 高度
- 不变矩特征
  - 在图像平移、缩放、旋转时均保持不变
  - 具有全局特性

Hu不变矩

$$
\begin{aligned}
M1 &= y_{20} + y_{02} \\
M2 &= (y_{20} - y_{02})^2 + 4y_{11}^2 \\
M3 &= (y_{30} - 3y_{12})^2 + (3y_{21} - y_{03})^2 \\
M4 &= (y_{30} + y_{12})^2 + (y_{21} + y_{03})^2 \\
M5 &= (y_{30} - 3y_{12})(y_{30} + y_{12})((y_{30} + y_{12})^2 - 3(y_{21} + y_{03})^2) \\
&\quad + (3y_{21} - y_{03})(y_{21} + y_{03})(3(y_{30} + y_{12})^2 - (y_{21} + y_{03})^2) \\
M6 &= (y_{20} - y_{02})((y_{30} + y_{12})^2 - (y_{21} + y_{03})^2) \\
&\quad + 4y_{11}(y_{30} + y_{12})(y_{21} + y_{03}) \\
M7 &= (3y_{21} - y_{03})(y_{30} + y_{12})((y_{30} + y_{12})^2 - 3(y_{21} + y_{03})^2) \\
&\quad - (y_{30} - 3y_{12})(y_{21} + y_{03})(3(y_{30} + y_{12})^2 - (y_{21} + y_{03})^2)
\end{aligned}
$$

5.2.2 形状特征

② 轮廓特征：

- Hough变换——直线、圆
- HOG描述子（ Histogramof Oriented Gradients）
- 傅立叶描述子Fourier Descriptor

5.2.3 纹理特征

- 纹理特征也是一种全局特征，用来描述图像中对应物体的表面性质。
- 纹理特征不能完全反映出物体的本质属性，仅仅利用纹理特征无法获得高层次图像内容。
- 纹理特征的特点
  - 对包含多个像素点的区域进行统计计算
  - 应具有旋转不变性
  - 容易受分辨率变化影响

5.2.3 纹理特征

- 灰度共生矩阵
- 自相关函数（图像的能量谱函数）
  - 直接计算图像的自相关函数（能量谱函数），可以得到图像纹理的粗细度、方向性等参数。
- LBP局部二值模式（Local Binary Pattern）
  - 利用相邻像素间的关系，将0-255灰度级的像素值映射到重新定义的灰度级，然后通过统计直方图来反映图像的全局纹理信息。
  - 具有光照不变性。
  - 统计与结构相结合的纹理分析方法。

5.2.4 空间关系特征

- 所谓空间关系，是指图像中分割出来的多个目标之间的相互空间位置，或者相对方向关系。
- 空间位置信息
  - 相对空间位置信息：目标之间的相对情况，如上下左右关系等
  - 绝对空间位置信息：目标之间的距离大小以及方位
- 常见的空间关系
  - 连接/邻接关系
  - 交叠/重叠关系
  - 包含/包容关系

第五章 特征选择与特征提取

CONTENTS PAGE

5.3 特征选择

5.3.1 问题与思路

5.3.2 特征评价准则

5.3.3 最优搜索

5.3.4 次优搜索

5.3.5 启发式搜索

5.3 特征选择

5.3.1 问题与思路

- 原始特征的问题：
  - 有很多特征可能与要解决的分类问题关系不大，但却在后续分类器设计中影响分类器性能。
  - 即使很多特征与分类问题关系密切，但特征过多导致计算量大、推广能力差。当样本数有限时容易出现病态矩阵等问题。

5.3 特征选择

5.3.1 问题与思路

- 特征选择的原因：
  - 模式识别系统的成败，首先取决于所采用的特征是否较好的反映模式的特性以及模式的分类问题。
  - 希望在保证分类效果前提下，采用尽可能少的特征完成分类。

5.3 特征选择

5.3.1 问题与思路

- 特征选择问题：
  - 已知给定的M个原始特征
  - 从中优选出m个特征（m ＜ M）
  - 即：从给定的M个原始特征中搜索到最优的特征组合
- 两方面解决：
  - 特征选择的准则：定义类别可分离性准则J_{ij}，用来衡量在一组特征下第i类和第j类之间的可分程度。
  - 特征搜索算法：在允许的时间内找出最优的那组特征。

5.3 特征选择

5.3.2 特征评价准则

➢ 评价准则的基本要求

① 类别可分性判据J与错误率（或错误率的上界）有单调关系，J最大时，错误率最小，这样才能较好地反映分类目标。

② 当特征独立时，判据J对特征应该具有可加性，即

$$J_{ij}(x_1, x_2, \cdots, x_d) = \sum_{k=1}^{d} J_{ij}(x_k)$$

J_{ij}表示使用括号中的特征组合时的第i类与第j类可分性判别函数，该值越大，两类的分离程度就越大，$x_1, x_2, \cdots, x_d$是一系列特征变量。

③ 判据应该具有以下度量特性：

$$J_{ij} \geq 0, i \neq j; J_{ij} = 0, i = j; J_{ij} = J_{ji}$$

④ 理想的判据应该对特征具有单调性，即加入新的特征不会使判据减小，即 $J_{ij}(x_1, x_2, \cdots, x_d) \leqslant J_{ij}(x_1, x_2, \cdots, x_d, x_{d+1})$

5.3.2 特征评价准则

① 基于类内类间距离的可分性判据

- 基本思想：各类样本可以分开是因为它们位于特征空间中的不同区域，显然这些**区域之间距离越大，类别可分性就越大**。
- **距离度量有多种**，如欧氏距离、各种平均平方距离：

$$J_1 = tr\left(\boldsymbol{S}_w + \boldsymbol{S}_b\right)$$

$$J_2 = tr\left(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\right)$$

$$J_3 = \ln\frac{\left|\boldsymbol{S}_b\right|}{\left|\boldsymbol{S}_w\right|}$$

$$J_4 = \frac{tr\left(\boldsymbol{S}_b\right)}{tr\left(\boldsymbol{S}_w\right)}$$

$$J_5 = \frac{\left|\boldsymbol{S}_b - \boldsymbol{S}_w\right|}{\left|\boldsymbol{S}_w\right|}$$

- ✓ 总类内离散度：$\mathbf{S}_w$
- ✓ 总类间离散度：$\mathbf{S}_b$

5.3.2 特征评价准则

② 基于信息熵的可分性判据

- 把类别$\omega_i, i=1,\cdots,c$看作一系列随机事件，它的发生依赖于随机向量x，给定x后的后验概率是$P(\omega_i \mid \boldsymbol{x})$。
- 如果根据x（这里代指特征）能完全确定ω，则ω就没有不确定性，对ω本身的观察就不会再提供信息量，此时熵为0，特征最有利于分类；
- 如果x（这里代指特征）完全不能确定ω，则ω不确定性最大，对ω本身的观察所提供信息量最大，此时熵为最大，特征最不利于分类。

（1）Shannon熵：
$$H=-\sum_{i=1}^{c}P(\omega_i|x)\log_2 P(\omega_i|x)$$

（2）平方熵：
$$H=2\left[1-\sum_{i=1}^{c}P^2(\omega_i \mid \boldsymbol{x})\right]$$

5.3.2 特征评价准则

② 基于信息熵的可分性判据

（1）Shannon熵：$$H=-\sum_{i=1}^{c}P(\omega_i|x)\log_2 P(\omega_i|x)$$

（2）平方熵：$$H=2\left[1-\sum_{i=1}^{c}P^2(\omega_i|x)\right]$$

在熵的基础上，对特征x的所有取值积分，就得到基于信息熵的可分性判据：

$$J_E=\int H(x)p(x)\mathrm{d}x$$

J_E 越小，可分性越好。

5.3 特征选择

5.3.1 问题与思路

5.3.2 特征评价准则

5.3.3 最优搜索

5.3.4 次优搜索

5.3.5 启发式搜索

5.3 特征选择

5.3.3 最优搜索

- 特征选择问题：
  - 已知给定的M个原始特征
  - 从中优选出m个特征（m < M）
  - 即：从给定的M个原始特征中搜索到最优的特征组合
- 全局搜索次数：

$$C_M^m = \frac{M!}{(M-m)!m!}$$ （穷举法）

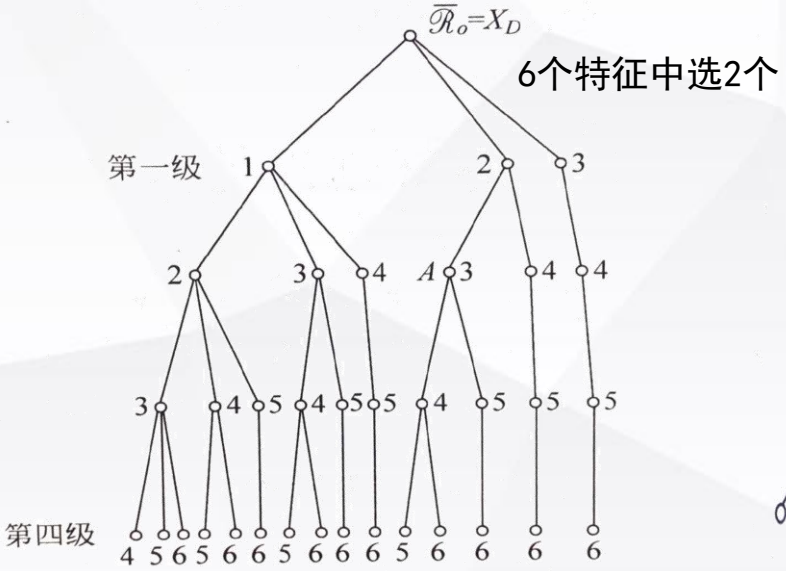
  - 运算量巨大！

5.3 特征选择

5.3.3 最优搜索

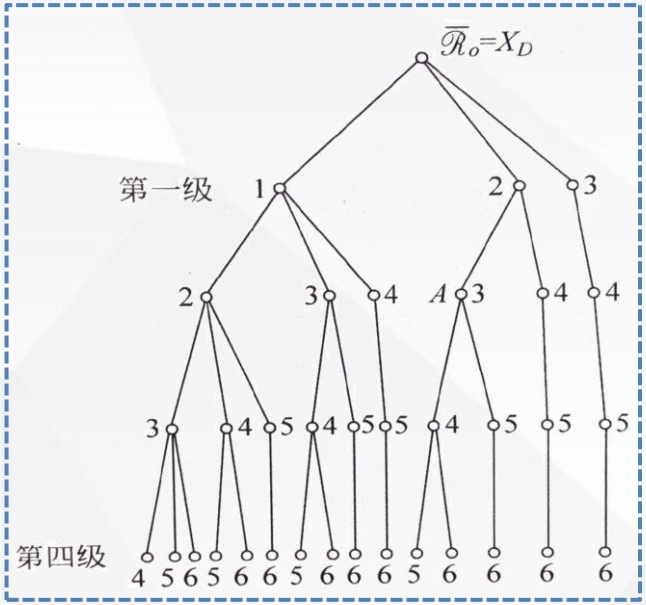
- 分支定界算法（Branch and Bound，简称B&B算法）

  - 是特征选择中的一种全局最优搜索算法；
  - 基本思想：将所有可能特征组合构建成一个树状结构，按照特定的规则对树进行搜索和回溯，尽早达到最优解，而不必遍历树的所有分支。



5.3 特征选择

5.3.3 最优搜索



- 利用了可分性判据中的单调性，即：

$$J_{ij}(x_1,x_2,\cdots,x_d)\leqslant J_{ij}(x_1,x_2,\cdots,x_d,x_{d+1})$$

- B&B算法步骤：

① 从包含所有特征（根节点）开始；

② 每一级去掉一个特征，逐步去掉不被选中的特征（可用特征的序号标记一个节点，同一层中按对准则函数的影响大小，从左到右排列）；

③ 底层的每个叶节点代表特征选择的一个组合；

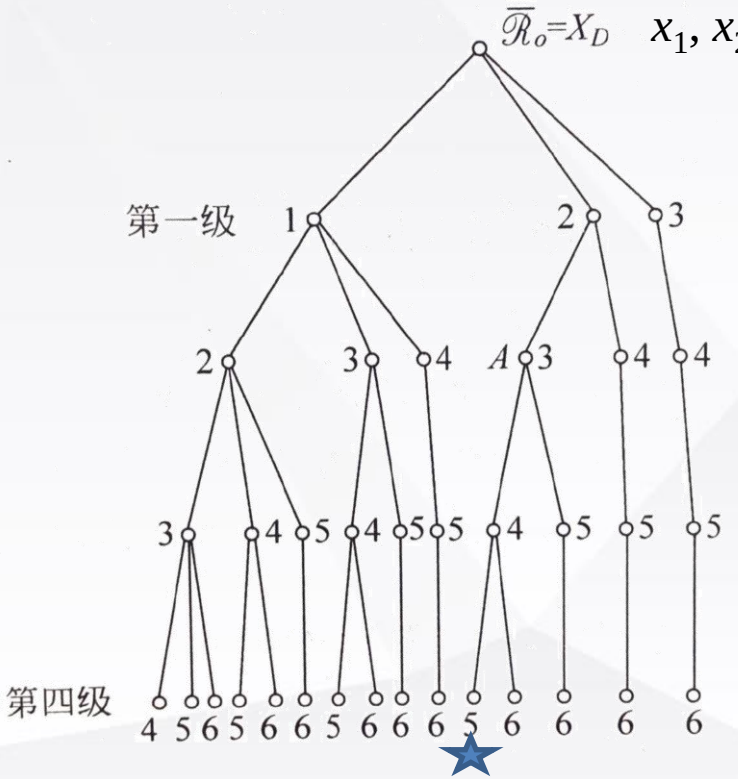
④ 到达叶节点后算法向上回溯，每回溯一级，应将相应节点上去掉的特征再收回来；

⑤ 定义准则函数阈值，用来结束搜索过程。

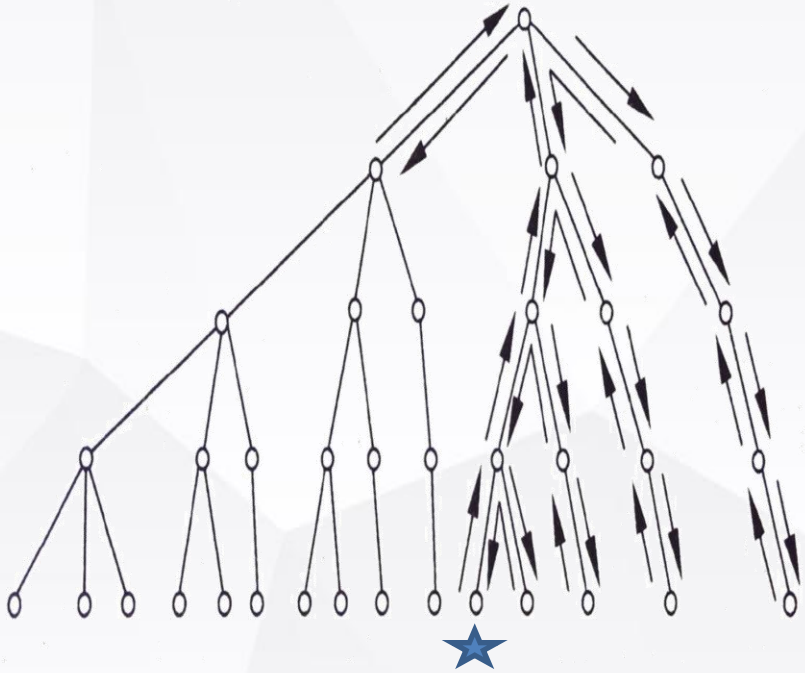
5.3 特征选择

5.3.3 最优搜索

- B&B算法举例：6个特征中选2个



完整的树



实际搜索过程

x_1, x_6

5.3 特征选择

5.3.3 最优搜索

- B&B算法特点：
  - 要求准则函数（即可分性判据）对特征具有单调性
  - 树的构建生长过程就是特征选择的过程
  - 回溯过程保证算法可以遍历所有可能组合
  - 算法有可能提前达到准则函数阈值，因此运算量有可能有效减少。

5.3.4 次优搜索

- 目标：相对最优搜索，继续减小计算量
- 常用次优搜索策略
  - ① 单独最优特征的组合：
  - ② SFS（Sequential forward selection，顺序前向搜索）
  - ③ SBS（Sequential backward selection，顺序后向搜索）
  - ④ L-R方法

5.3 特征选择

5.3.4 次优搜索

① 单独最优特征的组合

- 对每个特征单独计算类别可分性判据
- 对单个特征的判据值进行排序
- 选择前m个单独最优特征进行组合

- 为什么是次优？
  - 单独最优特征的组合不一定是最优的

5.3.4 次优搜索

② SFS（Sequential Forward Selection，顺序前向搜索）

- 考虑了一定的特征间组合的因素，属于**局部最优搜索**；
- 第一个特征选单独最优的特征；
- 第二个特征从其余所有特征中选择与第一个特征组合在一起后准则最优的特征；
- 后面每一个特征都选择与已有入选的特征组合起来最优的特征，以此类推，直到m个特征。

5.3.4 次优搜索

③ SBS（Sequential Backward Selection，顺序后向搜索）

- 属于局部最优搜索；
- 从所有特征开始逐一剔除不被选中的特征；
- 每次剔除的特征都是使得剩余的特征的准则函数值最优的特征；
- 直到m个特征为止。

5.3 特征选择

5.3.4 次优搜索

④ L-R方法

- 结合SFS和SBS，局部最优搜索
- 引入回溯——增L减R（L>R）：每次首先按照SFS方式逐步增选L个特征，然后再按照SBS方式逐步剔除R个特征，以此类推，直到选择到所需要数目的特征;
- 引入回溯——减R增L（L<R）：每次首先按照SBS逐步剔除R个特征，然后再从已经被剔除的特征中按照SFS方式逐步选择L个特征，直到剩余的特征数目达到所需数目；
- 特点：运算量大；回溯使得已选的特征可以剔除，剔除的特征可以再选。

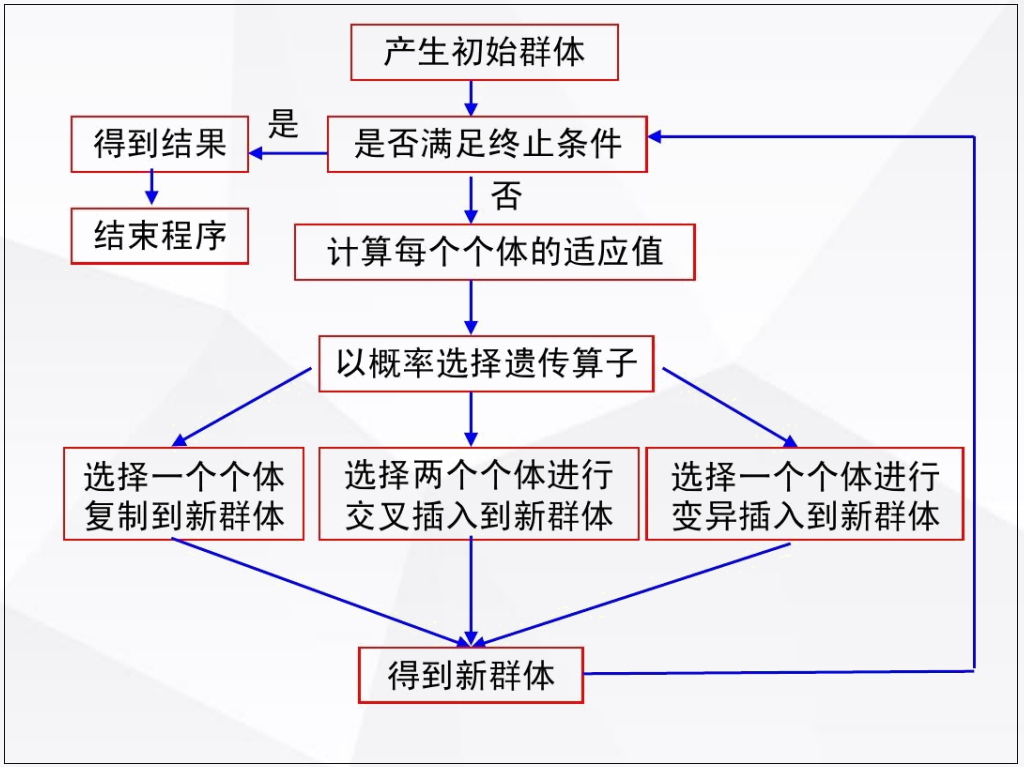
5.3.5 启发式搜索

- 特征维数较大时，最优搜索和次优搜索这类直接搜索方法计算量仍然太大，需要更好的搜索方法。
- 启发式搜索：利用问题拥有的启发信息来引导搜索，达到减少搜索范围、降低问题复杂度的目的。
  - 遗传算法
  - 模拟退火算法

5.3 特征选择

5.3.5 启发式搜索

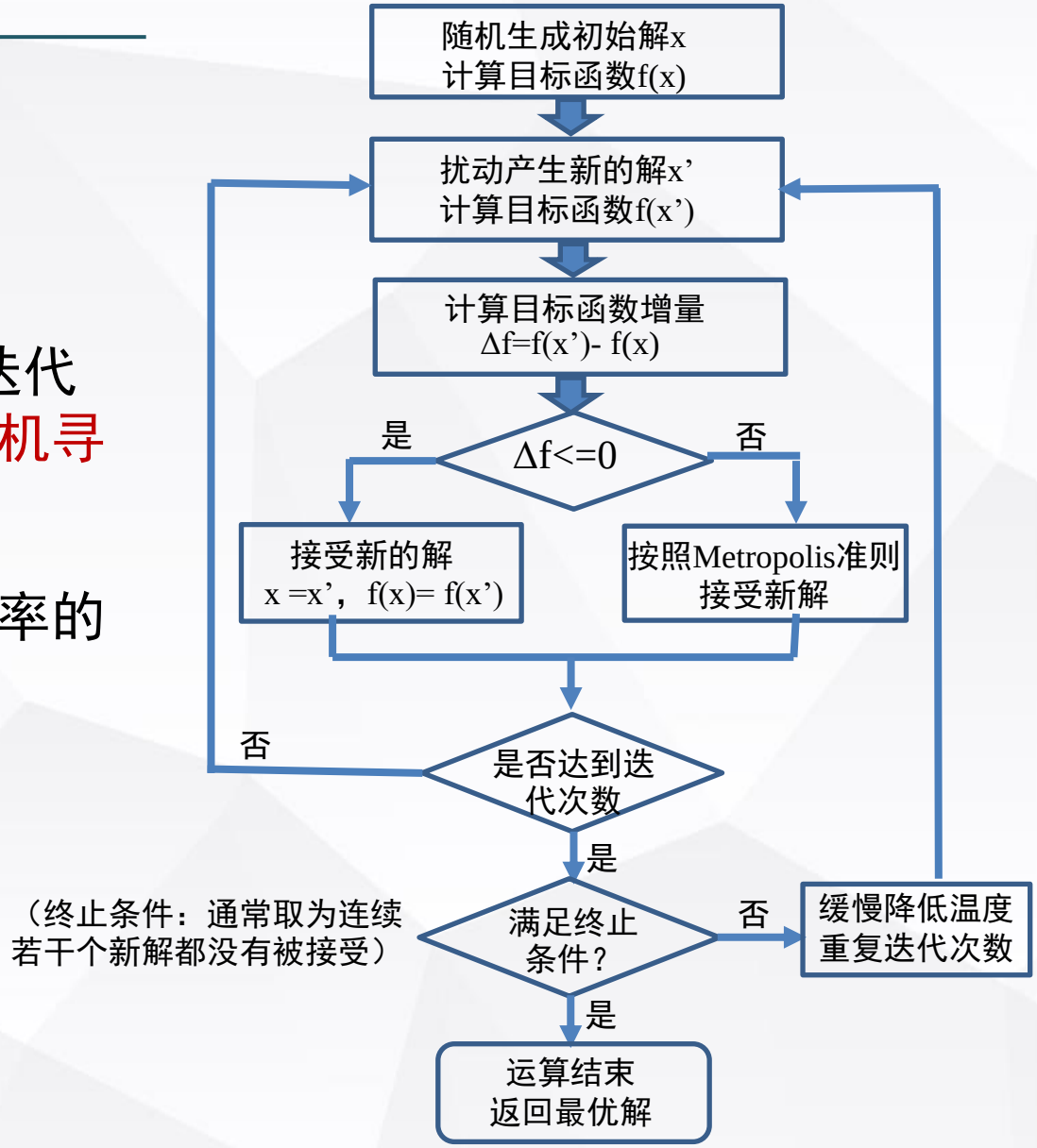
- 遗传算法（随机搜索算法）
  - 一种通过模拟生物选择和进化过程的搜索寻优方法
  - 不能保证全局最优



5.3 特征选择

5.3.5 启发式搜索

- 模拟退火算法
  - 基于Monte-Carlo迭代求解策略的一种随机寻优算法
  - 理论上算法具有概率的全局优化性能



第五章 特征选择与特征提取

CONTENTS PAGE

5.4 特征提取

5.4.1 问题与思路

5.4.2 基于类别可分性判据的特征提取方法

5.4.3 LDA

5.4.4 PCA

5.4.5 K-L变换

5.4.6 流形学习

5.4 特征提取

5.4.1 问题与思路

- 重温原始特征的问题：
  - 有很多特征可能与要解决的分类问题关系不大，但却在后续分类器设计中影响分类器性能。
  - 即使很多特征与分类问题关系密切，但特征过多导致计算量大、推广能力差。当样本数有限时容易出现病态矩阵等问题。

5.4 特征提取

5.4.1 问题与思路

- 特征提取的原因：（与特征选择一样）
  - 模式识别系统的成败，首先取决于所采用的特征是否较好的反映模式的特性以及模式的分类问题。
  - 原始特征依赖于具体应用问题和相关专业知识
  - 希望在保证分类效果前提下，采用尽可能少的特征完成分类。

5.4 特征提取

5.4.1 问题与思路

- 特征提取问题：
  - 已知给定的D个原始特征
  - 经过数学变换得到d个特征（d< D）
- 特征提取目的：
  ① 降低特征空间维数，使后续分类器设计在计算上更容易实现；
  ② 消除特征之间可能存在的相关性，减少特征中与分类无关的信息，使新的特征更有利于分类。

5.4 特征提取

5.4.1 问题与思路

- 数学变换形式可以多样：

① 最常采用的特征变换是线性变换（本节重点介绍），若$\boldsymbol{x} \in R^D$是D维原始特征，变换后的d维新特征$\boldsymbol{z} \in R^d$：

$$\boldsymbol{z} = \boldsymbol{W}^T \boldsymbol{x}$$

其中，$\boldsymbol{W}$是$D \times d$矩阵，称为变换矩阵。

② 有些情况下也可以采用非线性变换$\boldsymbol{z} = \boldsymbol{W}(\boldsymbol{x})$，此时$\boldsymbol{W}(.)$为非线性变换。

- 特征提取就是根据训练样本求适当的$\boldsymbol{W}$，使得某种特征变换的准则最优。

5.4.2 基于类别可分性判据的特征提取方法

- 如果采用类别可分性判据作为衡量新特征的准则，则特征提取问题就是求最优的W^*，使得：

$$z = W^T x$$

$$W^* = \arg\max_{\{W\}} J\left(W^T x\right)$$

① 如果采用基于类内类间距离的可分性判据J_1~J_5，经过W的特征变换之后，类内离散度矩阵和类间离散度矩阵分别变为：

a) 特征变换后的类内离散度矩阵：

$$W^T S_w W$$

b) 特征变换后的类间离散度矩阵：

$$W^T S_b W$$

5.4.2 基于类别可分性判据的特征提取方法

② 此时，特征提取的问题就是求W^*，使得下列准则最优

$$J_1(W)=tr\left[W^T\left(S_w+S_b\right)W\right]$$

$$J_2(W)=tr\left[\left(W^TS_wW\right)^{-1}\left(W^TS_bW\right)\right]$$

$$J_3(W)=\ln\frac{\left|W^TS_bW\right|}{\left|W^TS_wW\right|}$$

$$J_4(W)=\frac{tr\left(W^TS_bW\right)}{tr\left(W^TS_wW\right)}$$

$$J_5(W)=\frac{\left|W^T\left(S_w+S_b\right)W\right|}{\left|W^TS_wW\right|}$$

准则形式不同，但得到的最优变换矩阵W^*是相同的！

5.4.2 基于类别可分性判据的特征提取方法

② 此时，特征提取的问题就是求W^*，使得下列准则最优

$$J_1(W)=tr\left[W^T\left(S_w+S_b\right)W\right]$$

$$J_2(W)=tr\left[\left(W^TS_wW\right)^{-1}\left(W^TS_bW\right)\right]$$

$$J_3(W)=\ln\frac{\left|W^TS_bW\right|}{\left|W^TS_wW\right|}$$

$$J_4(W)=\frac{tr\left(W^TS_bW\right)}{tr\left(W^TS_wW\right)}$$

$$J_5(W)=\frac{\left|W^T\left(S_w+S_b\right)W\right|}{\left|W^TS_wW\right|}$$

- 设矩阵$S_w^{-1}S_b$的特征值为$\lambda_1,\lambda_2,\ldots,\lambda_D$，按照大小顺序排列为：

$$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_D$$

- 选择前d个特征值对应的特征向量作为W，即：

$$W=\left[u_1,u_2,\ldots,u_d\right]$$

- 以上W就是在这些准则下的最优变换阵。

5.4.2 基于类别可分性判据的特征提取方法

➢ 以J_1准则为例给出推导过程：

$$J_1(\boldsymbol{W}) = tr\left[\boldsymbol{W}^T\left(\boldsymbol{S}_w + \boldsymbol{S}_b\right)\boldsymbol{W}\right] \quad \text{（公式1）}$$

■ 要最大化J_1，无论选取什么样的$\boldsymbol{W}$，只要把它再乘以一个系数，则准则函数值总会再变大，但是变换的方向并没有改变，这就是尺度问题。为了解决以上问题，引入一个约束条件：

$$tr\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right) = c \quad \text{（公式2）}$$

■ 设c=1，此时优化问题变为：

$$\begin{aligned} &max J_1(\boldsymbol{W}) \\ &\text{s.t.} \quad tr\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right) = 1 \end{aligned} \quad \text{（公式3）}$$

■ 利用拉格朗日方法将以上有约束的优化问题转化为无约束问题，拉格朗日函数是：

5.4.2 基于类别可分性判据的特征提取方法

➢ 以J_1准则为例给出推导过程：

（公式4） $g(\boldsymbol{W}) = J_1(\boldsymbol{W}) - tr\left[\boldsymbol{\Lambda}\left(\boldsymbol{W}^T \boldsymbol{S}_W \boldsymbol{W} - \boldsymbol{I}\right)\right]$

$\boldsymbol{I}$是单位矩阵，$\boldsymbol{\Lambda}$是对角阵（对角线元素是拉格朗日乘子）。

■ 在拉格朗日函数极值点上，应该满足$\frac{\partial g(\boldsymbol{W})}{\partial \boldsymbol{W}} = 0$，由此得到：

$$\boldsymbol{S}_W^{-1}\left(\boldsymbol{S}_W + \boldsymbol{S}_b\right)\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda} \quad \text{（公式5）}$$

■ 整理，得到： $\boldsymbol{S}_W^{-1}\boldsymbol{S}_b\boldsymbol{W} = \boldsymbol{W}\left(\boldsymbol{\Lambda} - \boldsymbol{I}\right)$ （公式6）

可见，$\boldsymbol{W}$由$\boldsymbol{S}_W^{-1}\boldsymbol{S}_b$的特征向量组成，$\boldsymbol{\Lambda} - \boldsymbol{I}$等于$\boldsymbol{S}_W^{-1}\boldsymbol{S}_b$对应的特征值$\lambda_i$组成的对角阵，即：

$$\boldsymbol{\Lambda} = I + \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix}$$

5.4.2 基于类别可分性判据的特征提取方法

- 以J_1准则为例给出推导过程：
  - 此时J_1准则可以表示为：

$$J_1(\boldsymbol{W}) = tr\left[\boldsymbol{W}^T\left(\boldsymbol{S}_w + \boldsymbol{S}_b\right)\boldsymbol{W}\right]$$

$$= tr\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\boldsymbol{\Lambda}\right)$$

$$= tr\left(\boldsymbol{\Lambda}\right)$$

$\boldsymbol{S}_W^{-1}\left(\boldsymbol{S}_W + \boldsymbol{S}_b\right)\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda}$ （公式5）

$tr\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right) = 1$ （公式2）

  - 那么，对于$D \times d$变换矩阵

（公式7） $J_1(\boldsymbol{W}) = \sum_{i=1}^{d}\left(1+\lambda_i\right)$

$$\boldsymbol{\Lambda} = I + \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix}$$

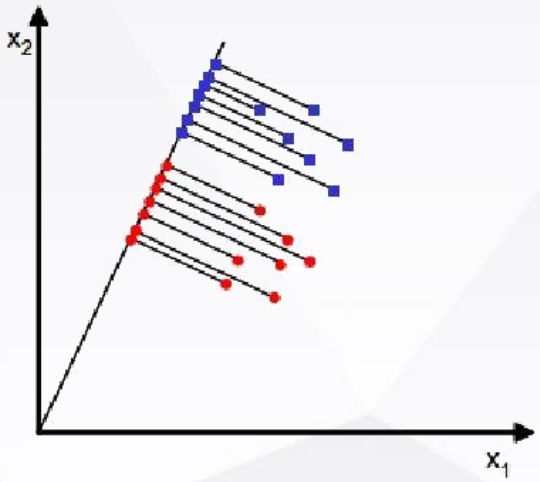
- 推导小结：最优的变换阵$\boldsymbol{W}$就是由$\boldsymbol{S}_W^{-1}\boldsymbol{S}_b$的前$d$个特征值所对应的特征向量组成的，而$J_1$准则值由公式(7)定义，其中$\lambda_i, i = 1,\ldots,d$为$\boldsymbol{S}_W^{-1}\boldsymbol{S}_b$的从大到小排列的前$d$个特征值。

5.4 特征提取

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

- Fisher投影准则回顾（二分类）：
  - 已知给定的D个原始特征
  - 经过投影得到1个特征
  - 求解广义Rayleigh商

$$J_F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$



  - 得到最佳投影向量

$$\mathbf{w}^* = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

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

- 多类FLDA：
  - 已知给定的D个原始特征
  - 投影到d维空间（d<D）,对应的基向量组成的矩阵为$\boldsymbol{W}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots\boldsymbol{u}_d]$，$\boldsymbol{W}\in R^{D\times d}$
  - 优化目标变为最大化如下公式：

$$\mathbf{J}(\boldsymbol{W})=\frac{\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}}$$

替换优化目标为

$\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}$ 和 $\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}$均为矩阵，不再是标量，无法作为一个标量函数来优化

5.4.2节

$$J_1(\boldsymbol{W})=tr\left[\boldsymbol{W}^T\left(\boldsymbol{S}_b+\boldsymbol{S}_w\right)\boldsymbol{W}\right]$$

$$J_2(\boldsymbol{W})=\frac{tr\left(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right)}{tr\left(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right)}$$

$$J_3(\boldsymbol{W})=\ln\frac{\left|\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right|}$$

$$J_4(\boldsymbol{W})=tr\left(\frac{\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}}\right)$$

$$J_5(\boldsymbol{W})=\frac{\left|\boldsymbol{W}^T\left(\boldsymbol{S}_b+\boldsymbol{S}_w\right)\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right|}$$

5.4 特征提取

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

➢ 多类FLDA ：

$$J_1(\boldsymbol{W}) = tr\left[\boldsymbol{W}^T\left(\boldsymbol{S}_b + \boldsymbol{S}_w\right)\boldsymbol{W}\right]$$

$$J_2(\boldsymbol{W}) = \frac{tr\left(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right)}{tr\left(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right)}$$

$$J_3(\boldsymbol{W}) = \ln\frac{\left|\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right|}$$

$$J_4(\boldsymbol{W}) = tr\left(\frac{\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}}\right)$$

$$J_5(\boldsymbol{W}) = \frac{\left|\boldsymbol{W}^T\left(\boldsymbol{S}_b + \boldsymbol{S}_w\right)\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right|}$$

（参见5.4.2节推导过程）

- 设矩阵$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的特征值为$\lambda_1,\lambda_2,\ldots,\lambda_D$，按照大小顺序排列为：

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$$

- 选择前d个特征值对应的特征向量作为$\boldsymbol{W}$，即：

$$\boldsymbol{W} = \left[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d\right]$$

- 以上$\boldsymbol{W}$就是在这些准则下的最优变换阵。

5.4 特征提取

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

- 多类FLDA算法流程：

输入：原始样本集$A=\{(\boldsymbol{x}_i, y_i)\}$，样本$i=1,\ldots,n$，$\boldsymbol{x}_i \in R^D$（$D$维度向量），$y_i$代表样本类别，包含$C$类。需要将特征降维到$d$维；

输出：降维后的样本集A'

步骤1：计算原始样本集类内离散度矩阵$\boldsymbol{S}_w$；

步骤2：计算原始样本集类间离散度矩阵$\boldsymbol{S}_b$；

步骤3：计算矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$；

步骤4：计算 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 的最大的d个特征值及对应的特征向量$\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots\boldsymbol{u}_d$，得到投影矩阵$\boldsymbol{W}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots\boldsymbol{u}_d]$；

步骤5：对原始样本集中的每个样本特征$\boldsymbol{x}_i$，转换为新的样本特征$\boldsymbol{z}_i = \boldsymbol{W}^T\boldsymbol{x}_i$，得到新的样本集$A'=\{(\boldsymbol{z}_i, y_i)\}$，样本$i=1,\ldots,n$

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

- FLDA问题讨论：

该特征提取方法是有监督，还是无监督方法？

答案：有监督，假设特征向量的类别标签已知，在降维过程中可以使用类别的先验知识（计算类内、类间离散度矩阵）；

5.4.4 PCA（Principal Components Analysis ）

- 一种典型的线性降维方法；

$$z = W^T x$$

- 通过正交变换将原始数据（D维）投影到新的维度空间（d维），用较少的维度包含了原始数据中的绝大部分信息。即，PCA提取出空间原始数据中的主要特征，减少数据冗余。

- 是一种无监督特征提取方法（特征向量类别未知）。

5.4.4 PCA（Principal Components Analysis ）

- PCA出发点：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，应具有如下性质：
  - 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PCA的两种等价推导。

5.4 特征提取

5.4.4 PCA（Principal Components Analysis ）

① 最近重构性：

- **对样本进行中心化，**有$\sum_i x_i = 0$
- 假定投影变换后得到的新坐标系为$\{\boldsymbol{u}_1, \boldsymbol{u}_2, ...\boldsymbol{u}_D\}$
- $\boldsymbol{u}_i$ 是标准正交基向量，$\boldsymbol{u}_i^T \boldsymbol{u}_j = \begin{cases} 1, & i = j \\ 0 & i \neq j \end{cases}$
- 降维到d维，则样本点$\boldsymbol{x}_i$在低维坐标系中的投影是

$$\boldsymbol{z}_i = \left(z_{i1}; z_{i2}; ...; z_{id}\right)$$

- $z_{ij} = \boldsymbol{u}_j^T \boldsymbol{x}_i$ 是 $\boldsymbol{x}_i$在低维坐标下第j维的坐标
- $\boldsymbol{x}_i$ 的重构 $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d} z_{ij} \boldsymbol{u}_j$

5.4.4 PCA（Principal Components Analysis ）

① 最近重构性：

- 在含n个样本点的训练集上重构误差

$$\boldsymbol{W}=\left(\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_d\right)$$

$$\sum_{i=1}^{n}\left\|\sum_{j=1}^{d}z_{ij}\boldsymbol{u}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{n}\boldsymbol{z}_i^T\boldsymbol{z}_i-2\sum_{i=1}^{n}\boldsymbol{z}_i^T\boldsymbol{W}^T\boldsymbol{x}_i+const$$

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

- 优化目标：

$$\min_{\boldsymbol{W}} \quad -tr\left(\boldsymbol{W}^T\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{W}\right) \qquad s.t. \qquad \boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{I}$$

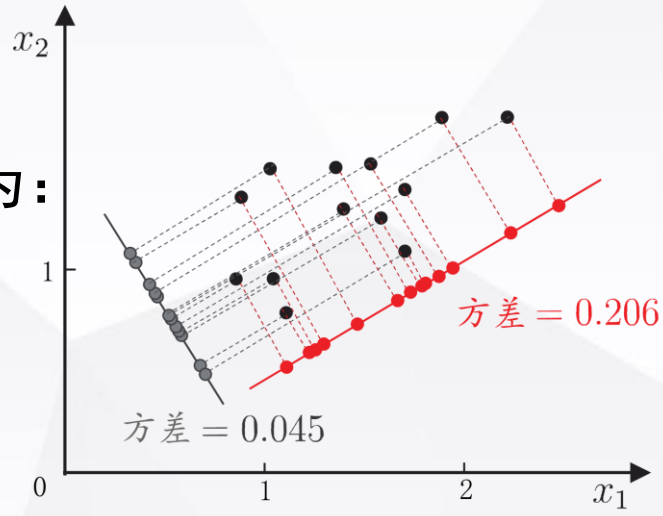
协方差矩阵，可表示为$\boldsymbol{X}\boldsymbol{X}^T$

5.4.4 PCA（Principal Components Analysis ）

② 最大可分性：

- 所有样本点的投影能尽可能分开，则投影后样本点方差最大
- 样本点 $\boldsymbol{x}_i$ 的投影为 $\boldsymbol{W}^T\boldsymbol{x}_i$，投影后样本点的协方差矩阵为：

$$\sum_i \boldsymbol{W}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{W}$$



- 优化目标（使训练集上方差最大）为：

$$\max_{\boldsymbol{W}} \quad tr\left(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W}\right)$$

$$s.t. \quad \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$$

$\boldsymbol{X}\boldsymbol{X}^T$ ：样本协方差矩阵

5.4.4 PCA（Principal Components Analysis ）

最近重构性

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

等价

最大可分性

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

③ PCA求解：

- 拉格朗日乘子法，并令关于W的偏导数为零，得到：

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

- 对样本协方差矩阵 XX^{T} 进行特征值分解
- 取前d个特征值对应的特征向量构成PCA的解

$$W=(u_1,u_2,...,u_d)$$

5.4 特征提取

$$XX^TW = \lambda W$$

- 对样本协方差矩阵 XX^T进行特征值分解
- 取前d个特征值对应的特征向量构成PCA的解 $W = (u_1, u_2, ..., u_d)$

5.4.4 PCA

④ PCA算法流程：

输入：D维的样本集A，要降维到的维数d

输出：降维后的样本集A’。

步骤1：对原样本集A中所有的样本进行中心化；

步骤2：计算中心化后样本的协方差矩阵；

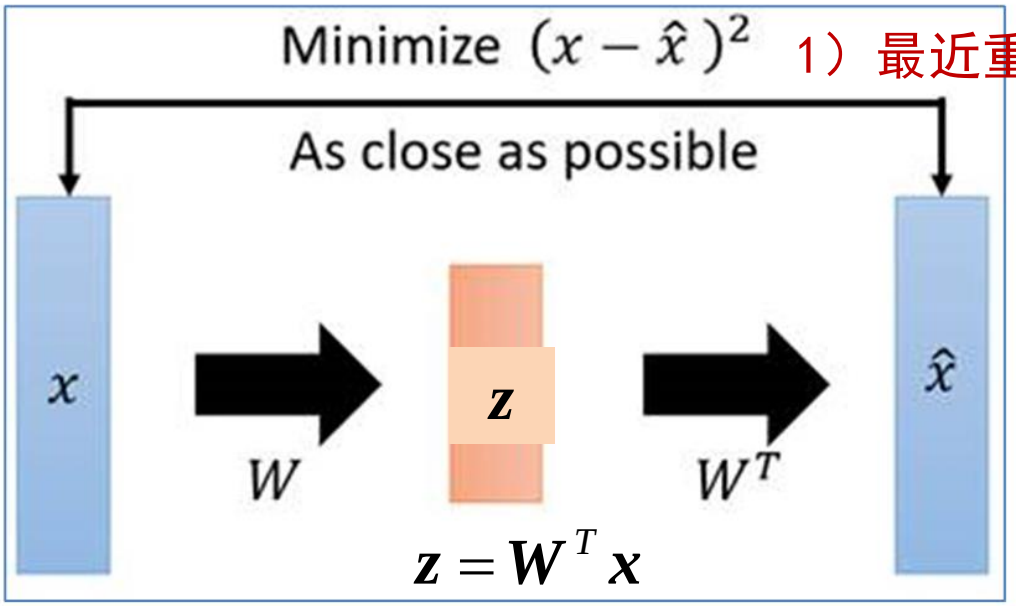
步骤3：对以上协方差矩阵进行特征值分解；

步骤4：取出最大的d个特征值对应的特征向量，并将所有特征向量标准化后，组成特征向量矩阵；

步骤5：利用特征向量矩阵，将原样本集A中的每个样本转化为新的样本，输出新的样本集A’。

5.4.4 PCA

- 用示意图去理解PCA



1）最近重构性

$$\hat{x} = \left(W^T\right)^T z = Wz$$

2）最大可分性：投影后方差最大 $\sum_i z_i z_i^T = \sum_i W^T x_i x_i^T W$

优化目标函数：$\max_W \quad tr\left(W^T X X^T W\right)$

5.4.4 PCA

⑤ PCA特点：

- 求解协方差矩阵的特征值对应的特征向量
- 一个矩阵有多个特征值，一个特征值对应一个主成分
- 特征值最大（方差最大）的主成分包含原始数据的信息最多，从而可对新特征排列——“重要性”
- 各个主成分之间线性无关，即新特征之间不相关——正交变换
- 不考虑样本的类别——非监督方法

5.4 特征提取

5.4.4 PCA

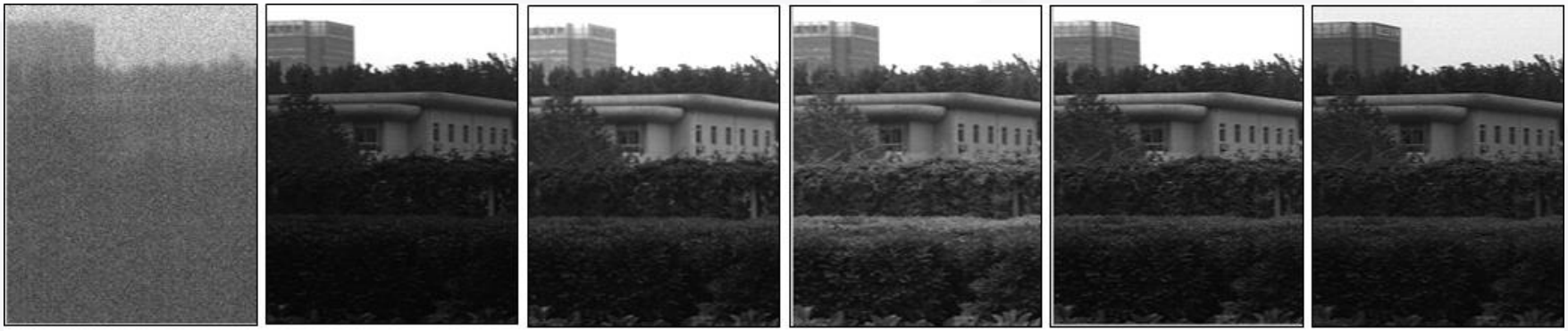
⑥ PCA应用举例

10个人的人脸图像数据集

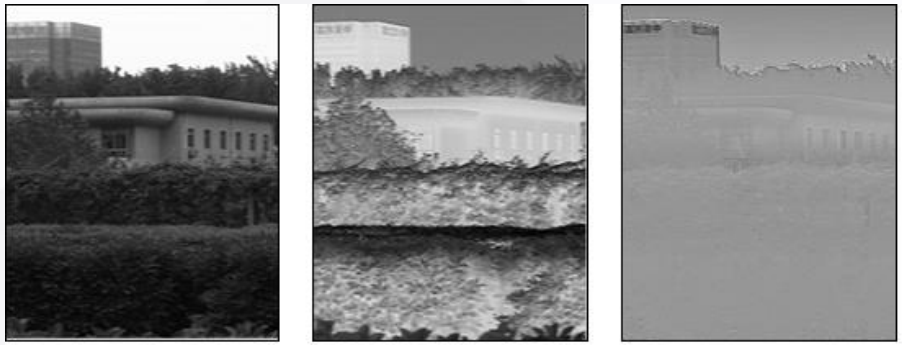
PCA变换后得到的特征脸

5.4.4 PCA

⑥ PCA应用举例



520波段高光谱图像中的6个波段



对520个数据进行PCA变换后，第1、2、10个主成分

5.4 特征提取

5.4.5 K-L变换

- K-L变换（Karhunen-Loeve）是由H.Karhunen和M.Leoeve等人最早提出来的，用于处理随机过程中的连续信号的去相关问题。
- 1933年，霍特林（Hotelling）提出了一种离散信号的去相关线性变换，称为霍特林变换，它实际是K-L级数展开的离散等效方法。因此，霍特林变换也可以称为离散K-L变换。
- 习惯上，不论对连续或离散信号，这种去相关变换，统称为Karhuner-Loeve(K-L)变换或Hotelling变换。
- K-L变换是常用的特征提取方法。

5.4.5 K-L变换

K-L变换原理

- 模式识别中的一个样本可以看作是随机向量的一次实现。
- 对一个D维随机向量$\boldsymbol{x}$，可以用确定的完备归一化正交向量系$\{\boldsymbol{u}_i\}_{i=1}^{\infty}$来展开

$$\boldsymbol{x}=\sum_{i=1}^{\infty}a_i\boldsymbol{u}_i, a_i\in\boldsymbol{R} \quad \text{公式（1）}$$

  - ✓ $\boldsymbol{x}$表示成$\boldsymbol{u}_i$的线性组合，其中：$\boldsymbol{u}_i^T\boldsymbol{u}_j=\begin{cases}1, i=j\\0, i\neq j\end{cases}$
  - ✓ a_i是线性组合的系数，将以上公式两边同时左乘$\boldsymbol{u}_i^T$，得到$a_i=\boldsymbol{u}_i^T\boldsymbol{x}$。

- 如果用有限的d项($d<D$)来逼近$\boldsymbol{x}$，即 $\hat{\boldsymbol{x}}=\sum_{i=1}^{d}a_i\boldsymbol{u}_i$

$$x=\sum_{i=1}^{\infty}a_i u_i, a_i \in R$$

$$\hat{x}=\sum_{i=1}^{d}a_i u_i$$

5.4.5 K-L变换

K-L变换原理

$$a_i = u_i^T x$$

- 则与原向量x的均方误差为：

$$\xi = E\left[(x-\hat{x})^T(x-\hat{x})\right] = E\left[\left(\sum_{i=d+1}^{\infty}a_i u_i\right)^T\left(\sum_{i=d+1}^{\infty}a_i u_i\right)\right]$$

$$= E\left(\sum_{i=d+1}^{\infty}u_i^T x x^T u_i\right) = \sum_{i=d+1}^{\infty}u_i^T E\left(xx^T\right)u_i = \sum_{i=d+1}^{\infty}u_i^T R u_i$$

公式（2）

其中，$R = E\left(xx^T\right)$ 为样本自相关矩阵。不同的正交向量系对应不同的均方误差，其选择应该使得均方误差最小。

- 即求解如下优化问题

$$\min \xi = \sum_{i=d+1}^{\infty}u_i^T R u_i \qquad \text{s.t } u_i^T u_i - 1 = 0, \forall i$$

公式（3）

5.4.5 K-L变换

$$\min \xi = \sum_{i=d+1}^{\infty} \boldsymbol{u}_i^T \boldsymbol{R} \boldsymbol{u}_i \qquad \text{s.t } \boldsymbol{u}_i^T \boldsymbol{u}_i - 1 = 0, \forall i \qquad \text{公式（3）}$$

K-L变换原理

- 利用拉格朗日乘子法求解

$$J(\boldsymbol{u}_i) = \sum_{i=d+1}^{\infty} \boldsymbol{u}_i^T \boldsymbol{R} \boldsymbol{u}_i - \sum_{i=d+1}^{\infty} \lambda_i \left(\boldsymbol{u}_i^T \boldsymbol{u}_i - 1 \right) \qquad \text{公式（4）}$$

对$\boldsymbol{u}_i$求导，并令导数为零，有

$$\left(\boldsymbol{R} - \lambda_i \boldsymbol{I} \right) \boldsymbol{u}_i = 0, i = d+1, \ldots, \infty$$

即， $\boldsymbol{u}_i$是矩阵$\boldsymbol{R}$的特征向量且满足

$$\boldsymbol{R} \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \qquad \lambda_i \text{是矩阵} \boldsymbol{R} \text{的特征值} \qquad \text{公式（5）}$$

以上推导表明：当利用自相关矩阵R的特征值对应的特征向量展开$\boldsymbol{x}$时，截断误差最小。

5.4.5 K-L变换

$$\min \xi = \sum_{i=d+1}^{\infty} \boldsymbol{u}_i^T \boldsymbol{R} \boldsymbol{u}_i \qquad \text{s.t } \boldsymbol{u}_i^T \boldsymbol{u}_i - 1 = 0, \forall i \quad \text{公式（3）}$$

K-L变换原理

- 此时，选择前d项估计$\boldsymbol{x}$时，结合公式(2)、(3)和(5)，引起的均方误差为：

$$\xi = \sum_{d+1}^{\infty} \boldsymbol{u}_i^T \boldsymbol{R} \boldsymbol{u}_i = \sum_{d+1}^{\infty} \boldsymbol{u}_i^T \lambda_i \boldsymbol{u}_i = \sum_{d+1}^{\infty} \lambda_i \quad \text{公式（6）}$$

- ✓ λ_i决定了截断的均方误差，λ_i值越小，则均方误差越小。
- ✓ 当用$\boldsymbol{x}$的K-L展开式中的前d项估计$\boldsymbol{x}$时，展开式中的$\boldsymbol{u}_i$应该是自相关矩阵$\boldsymbol{R}$的前d个较大特征值所对应的特征向量。
- ✓ $\boldsymbol{u}_i$，i=1, 2, …,d 组成了新的特征空间，样本$\boldsymbol{x}$在这个新空间中的展开系数$a_i = \boldsymbol{u}_i^T \boldsymbol{x}$, i=1, 2, …,d 组成了样本新的特征向量。
- ✓ 以上特征提取方法称为K-L变换，矩阵$\boldsymbol{R}$被称为K-L变换的产生矩阵。

5.4.5 K-L变换

- 产生矩阵$\boldsymbol{R}$可以有多种形式，如

自相关矩阵：$\boldsymbol{R} = E\left(\boldsymbol{x}\boldsymbol{x}^T\right) \approx \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$

无类别标签样本集

协方差矩阵：$\boldsymbol{\Sigma} = E\left[\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\right]$ ⟹ 此时K-L变换与PCA等价

总类内离散度矩阵：$S_w = \sum_{i=1}^{c} P_i \boldsymbol{\Sigma}_i$

$\boldsymbol{\Sigma}_i = E\left[\left(\boldsymbol{x}_i-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}_i\right)^T\right]$

有类别标签样本集

5.4 特征提取

作业

从原理上分析PCA和K-L变换的异同

5.4 特征提取

5.4.1 问题与思路

5.4.2 基于类别可分性判据的特征提取方法

5.4.3 LDA

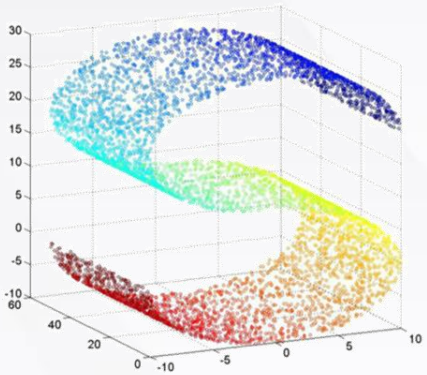
5.4.4 PCA

5.4.5 K-L变换

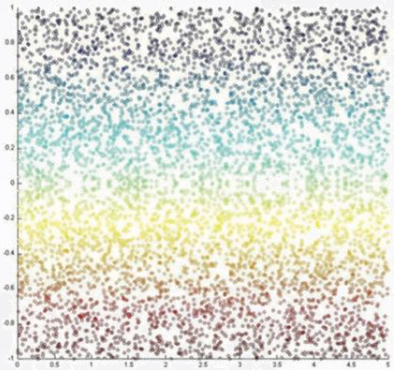
5.4.6 流形学习

5.4.6 流形学习

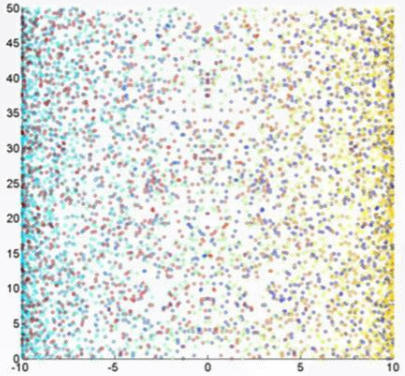
- 线性降维的问题： $z = W^T x$



样本三维分布



理想二维分布

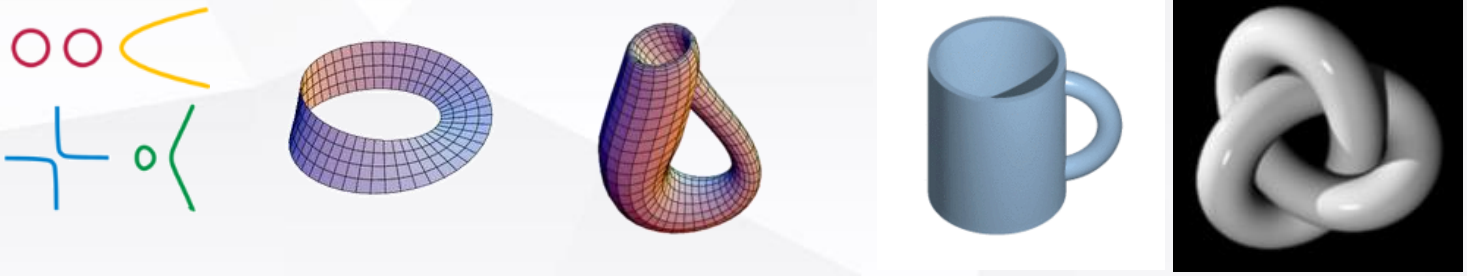


PCA结果

- 需要研究非线性降维方法

5.4.6 流形学习

- 流形的数学概念
  - “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。
  - 流形是欧几里得空间中的曲线、曲面等概念的推广。



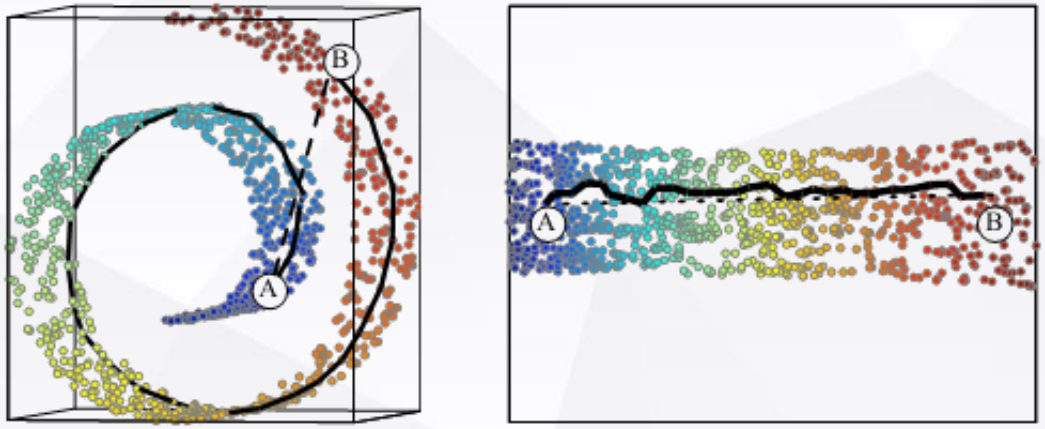
5.4.6 流形学习

- 流形学习（ manifold learning ）是一类借鉴了拓扑流形概念的降维方法。
- 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，能对高维数据进行可视化展示，因此流形学习也可被用于可视化。

5.4.6 流形学习

- 流形学习
  - 寻找低维嵌入结构
  - 发现映射关系

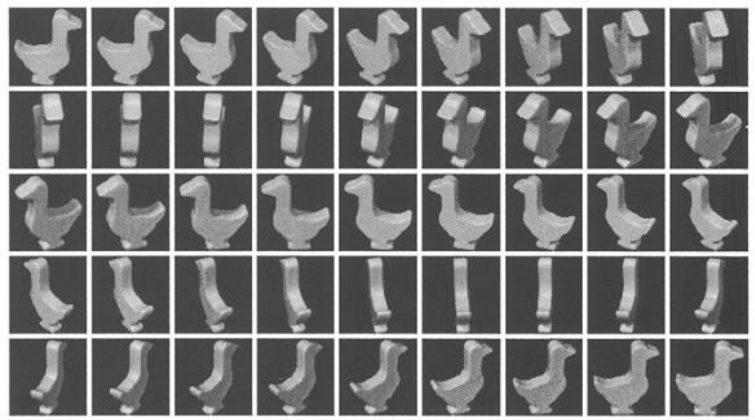
设$\mathrm{Y} \subset \mathbb{R}^d$是一个低维流形，$f: Y \to \mathbb{R}^D$是一个光滑嵌入，其中 $D > d$，Y中的数据集为$\{y_i\}$，经过f映射为观察空间的数据$\{x_i = f(y_i)\}$，流形学习就是给定$\{x_i\}$的条件下重构f和$\{y_i\}$



5.4 特征提取

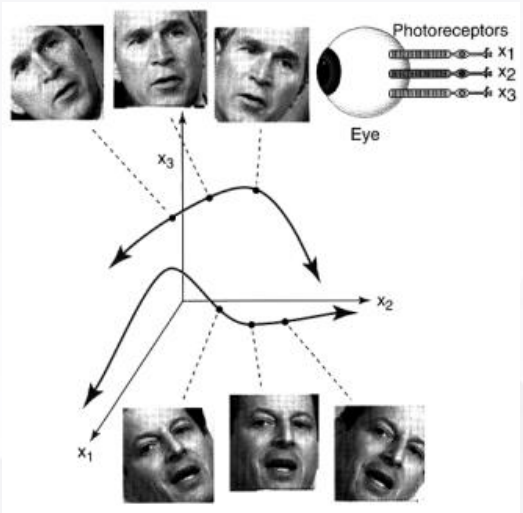
5.4.6 流形学习

以360度围绕同一个目标旋转成像

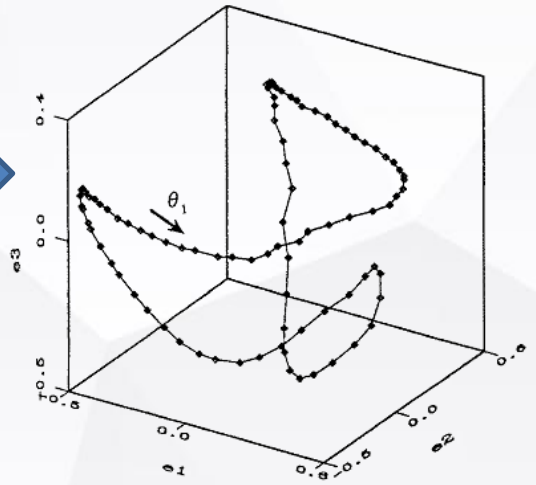


2000年在Science杂志上发表的研究成果指出：

人的视觉感知以流形方式存在



旋转图像序列的低维流形结构

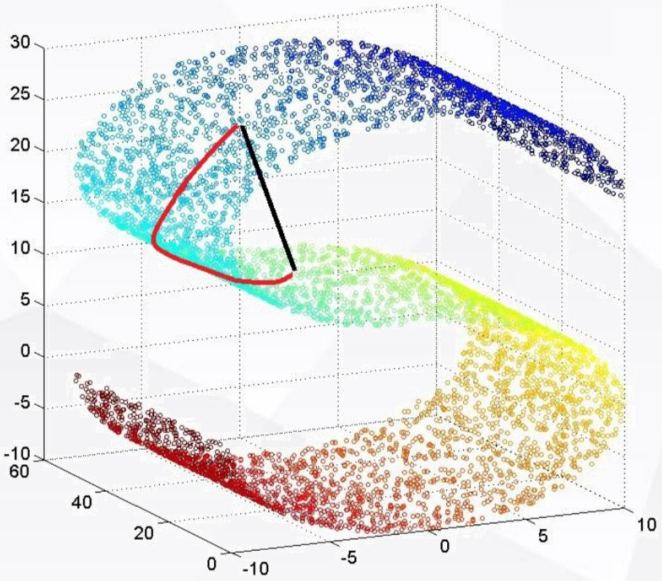


图像序列按成像顺序分布在一条三维空间中的闭合曲线（反映了成像时的旋转角度变化）

5.4.6 流形学习

① 等度量映射（Isometric Mapping，Isomap）方法

- 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不可达）。而低维嵌入流形上两点间的本真距离是**“测地线”(geodesic)距离。**

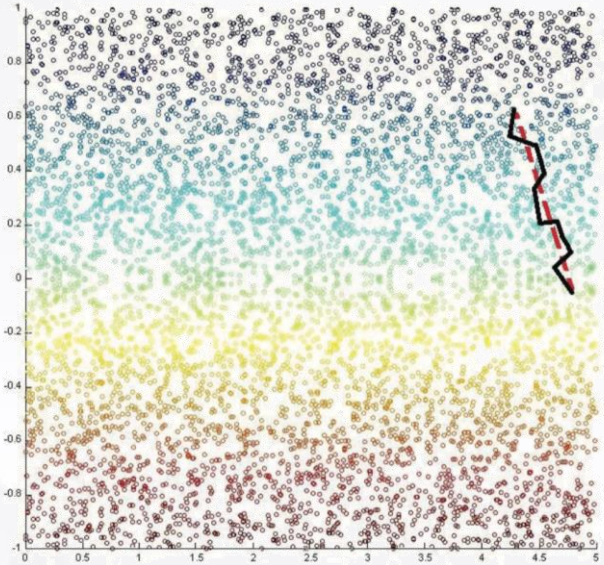


红色曲线是距离最短的路径，即S曲面上的测地线，测地线距离是两点之间的本真距离。

5.4.6 流形学习

① 等度量映射（Isometric Mapping，Isomap）方法

- **测地线距离的计算：**
  - 利用流形在局部上与欧氏空间同胚的性质，对每个点基于欧氏距离找出其近邻点；
  - 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；
  - 此时，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。



- 最短路径的计算可通过Dijkstra算法或Floyd算法实现

5.4.6 流形学习

① 等度量映射（Isometric Mapping，Isomap）方法

输入： 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
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: $\boldsymbol{x}_i$ 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: 调用最短路径算法计算任意两样本点之间的距离 $\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$;
6: 将 $\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 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 $Z=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m\}$.

5.4.6 流形学习

➢ 补充：MDS法（multi-dimensional scaling，多维尺度法）

- 是一种很经典的数据映射方法，将定义在多维空间中的样本间关系按比例缩放到二维或三维空间中展示出来。
- 该方法出发点：不是直接将样本从一个空间映射到另外一个空间，而是根据样本之间的距离关系或不相似度关系在低维空间（二维或三维空间）里生成对样本的一种表示，实现原特征空间到低维表示空间的变换。

5.4.6 流形学习

➢ 补充：MDS法（multi-dimensional scaling，多维尺度法）

■ 分为度量型（metric）和非度量型(non-metric)两种类型

a）度量型MDS把样本间的距离或不相似度看作一种定量的度量，希望在低维空间里的表示能够尽可能保持这种度量关系。

b）非度量型MDS把样本间的距离或不相似度关系仅仅看作是一种定性的关系，在低维空间里的表示只需要保持这种关系的顺序。

5.4.6 流形学习

➢ 补充：MDS法（multi-dimensional scaling，多维尺度法）

■ 度量型MDS。

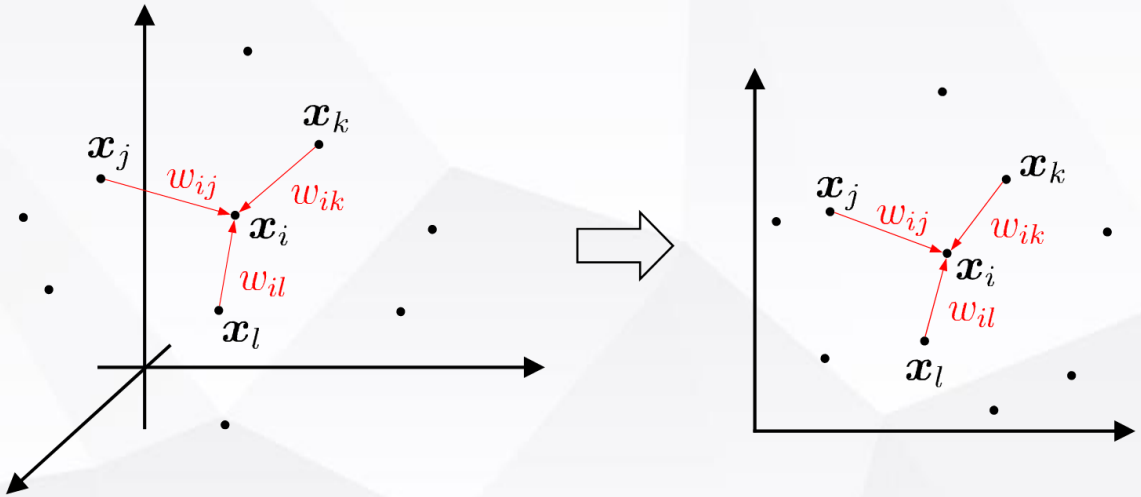
✓ 已知一组样本两两之间的不相似度量δ_{ij}（可以是某种距离度量），要用某个低维空间中一组点来表示这组样本，这组点在低维空间中两两之间的距离是d_{ij}，希望所得到的低维空间表示能使d_{ij}尽可能地代表δ_{ij}。

✓ 可以采用δ_{ij}（称作给定距离）的平方与d_{ij}（称作表示距离）的平方之间的平均误差$\sum_{i,j}(\delta_{ij}^2 - d_{ij}^2)$作为目标函数。

5.4.6 流形学习

② 局部线性嵌入（Locally Linear Embedding，LLE）方法

- 基本思想：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空间中继续保持。

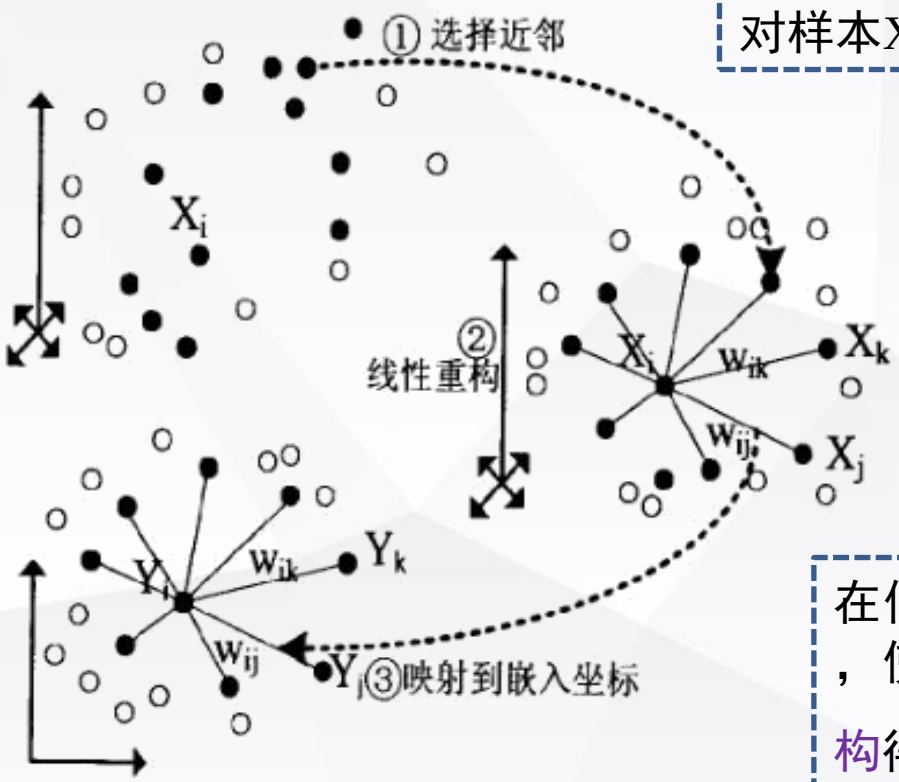


$$\boldsymbol{x}_i = w_{ij}\boldsymbol{x}_j + w_{ik}\boldsymbol{x}_k + w_{il}\boldsymbol{x}_l$$

5.4.6 流形学习

② 局部线性嵌入（Locally Linear Embedding，LLE）

➢ 算法基本步骤示意图



对样本X_i选择一组邻域样本

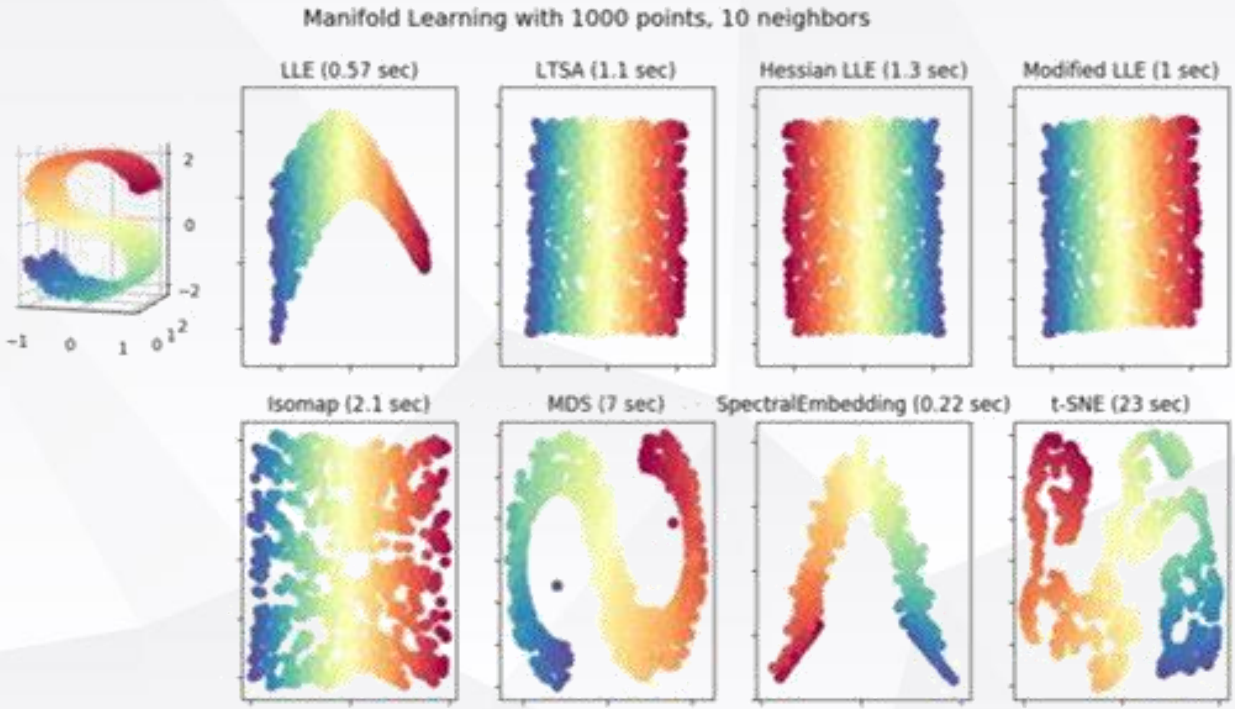
用这组邻域样本的线性加权组合重构X_i，得到一组使得如下重构误差最小的权值w_{ij}

$$\left| X_i - \sum_j w_{ij} X_j \right|$$

在低维空间里求向量Y_i及其邻域的映射，使得对所有样本用同样的权值进行重构得到的误差$\left| Y_i - \sum_j w_{ij} Y_j \right|$最小

5.4.6 流形学习

- 不同流形学习方法的效果差异



5.4.6 流形学习

- 不同流形学习方法的效果差异