—武大本科生课程

第11讲　特征提取和选择

(Lecture 11 Feature Extraction and Selection)

武汉大学计算机学院机器学习课程组

2023.04

第9章　特征提取和选择

内容目录 (以下红色字体为本讲3学时讲授内容)

9.1 基本概念

在模式识别领域，特征提取与选择尤为重要。特征提取的基本思想是将处于高维空间中的原始样本特征描述映射为低维特征特征描述。不同的模式识别应用，需要采用不同的特征提取与选择方法。

对于实际的模式识别问题，以人脸识别为例，一开始的原始特征可能很多，如ORL (https://cam-orl.co.uk/facedatabase.html)人脸数据库中，每幅图像的分辨率为112×92，如果将每个像素作为1维特征，则高达10304维(Figure 1)。若把所有的原始特征都作为分类特征送到分类器，不仅使得分类器复杂，分类判别计算量大，而且分类错误概率也不一定小；原始特征的特征空间有很大的冗余，完全可以用很小的空间相当好地近似表示图像，这一点与压缩的思想类似。因此有必要减少特征数目，以获取“少而精”的分类特征，即获取特征数目少且能使分类错误概率小的特征向量。

Figure 1 ORL face images

特征提取与具体问题有很大关系，目前还没有完备的理论能给出对任何问题都有效的特征提取方法。由于在许多实际问题中，那些最重要的特征往往不易找到，因此研究样本的特征模式、提取有效的特征成为构建模式识别系统的核心任务之一。一些特征及经典特征提取方法如下：

- 结构特征，如指纹的结构特征(斗型纹whorl，箕型纹loop)；
- 统计特征，如矩[3]，灰度共生矩阵(Co-occurrence Ma tr ix)就是矩的典型运用[4]等；
- 用LDA(Fisher Linear Discriminant Analysis)方法作特征压缩；
- 用PCA(Principal Component Analysis)方法作特征压缩；
- 用SVD(Singular Value Decomposition，奇异值分解)方法作特征压缩；
- 用流形学习方法作特征提取，如等度量映射(Isometric Mapping, Isomap)方法、局部线性嵌入方法(Locally Linear Embedding, LLE)等；
- 用傅立叶变换FT (频谱分析) 或小波变换WT (时频域分析) 的系数作为信号或图像的特征；
- ……

本章主要介绍PCA方法及灰度共生矩阵。

1. 对特征的要求

作为识别分类用的特征应具备以下几个条件：

(1)具有很大的识别信息量。即所提供的特征应具有很好的可分性，使分类器容易判别。

(2)具有可靠性。对那些模棱两可，似是而非不易判别的特征应该去掉。

(3)具有尽可能强的独立性。重复的、相关性强的特征只选一个，因为强的相关性并没有增加更多的分类信息，不能要。

(4)数量尽可能少，同时损失的信息尽量小。

模式识别中减少特征数目的方法有两种：一种是**特征提取**，另一种是**特征选择**。

- **原始特征**：**通过直接测量得到的特征称为原始特征**。比如人体的各种生理指标（描述其健康状况）；数字图像中的各像素点的亮度值（描述图像内容），都是原始特征。

- **特征提取**：**通过映射 (变换) 的方法把高维的特征向量转为低维的特征向量。**

$$A: X \rightarrow Y$$

- **特征选择**：**从原始特征中挑选出一些最有代表性、分类性能好的特征，以达到降低特征空间维数的目的。**

$$D\text{个特征} \rightarrow d\text{个特征}(d<D)$$

就是说，特征选择就是从已有的D个原始特征中挑选出d个特征组成一个特征子集，同时将D-d个对类别可分离性无贡献的或贡献不大的特征简单地忽略掉。

2. 特征的特点

模式识别的主要功能在于利用计算机实现人的类识别能力，它是一个与领域专门知识有关的问题。

研究领域不同，选择的特征也不同，但不论采用什么样的特征，都应该满足如下条件：

(1) 特征可获取

模式识别系统的主要处理设备是计算机，因此作为观察对象的数字化表达，观察对象应该是可以通过数据采集设备输入到计算机的。目前，市场上有各种传感设备和数字化设备，如采集图像信息的图像卡和采集语音信息的声卡等。作为特征，既可以是数字化表达的结果，也可以是在数字化表达基础上形成的参数性质的值，如图像分割后的子目标特征表达等。

(2) 类内稳定

选择的特征对同一类应具有稳定性。 由于模式类是由具有相似特性的若干个模式构成的，因此它们同属一类模式，其首要前提是特性相似，反映在取值上，就应该有较好的稳定性。

(3) 类间差异

选择的特征对不同的类应该有差异。若不同类的模式的特征值差异很小，则说明所选择的特征对于不同的类没有什么差异，作为分类的依据时，容易使不同的类产生混淆，使误识率增大。 一般来讲，特征的类间差异应该大于类内差异。

3. 特征的类别

特征是用于描述模式性质的一种量，从形式上看识别对象的特征可以分为三类：

(1) 物理特征

物理特征是比较直接、人们容易感知的特征，一般在设计模式识别系统时容易被选用。如为了描述指定班级中的某名学生，可以用以下物理特征：性别、 身高、 胖瘦、肤色等外在特征。物理特征虽然容易感知，却未必能非常有效地表征分类对象。

(2) 结构特征

结构特征的表达能力一般要高于物理特征，如汉字识别的成功实现离不开结构特征的选择。结构特征的表达是先将观察对象分割成若干个基本构成要素，再确定基本要素间的相互连接关系。

通过要素和相互连接关系表达对象，可以较好地表达复杂的图像信息，在实际中已经有较多的成功应用， 如指纹的识别就是基于结构信息完成的。结构信息对对象的尺寸往往不太敏感，如汉字识别时，识别系统对汉字大小不敏感，只对笔划结构信息敏感。

结构特征比物理特征要抽象一些，但仍属比较容易感知的特征，如人的指纹特征、 人脸的五官结构信息等，是目前认定人的身份的重要参数。

(3) 数学(数字)特征

一般来说，数学特征是为了表征观察对象而设立的特征，如给每名学生设立一个学号，作为标志每名学生的特征。由于学号是人为设定的，可保证唯一性，但这种特征是抽象的，不容易被人感知。而数学特征易于用机器定量描述与判别，如统计平均值、相关系数、协方差矩阵的特征值与特征向量、距离等都是数学特征。

4. 特征的形成

在设计一个具体的模式识别系统时，往往是先接触一些训练样本，由领域专家和系统工程师联合研究模式类所包含的特征信息，并给出相应的表述方法。这一阶段的主要目标是获取尽可能多的表述特征。在这些特征中，有些可能满足类内稳定、类间离散的要求，有的则可能不满足，不能作为分类的依据。根据样例分析得到一组表述观察对象的特征值，而不论特征是否实用，称这一步为特征形成，得到的特征称为原始特征。

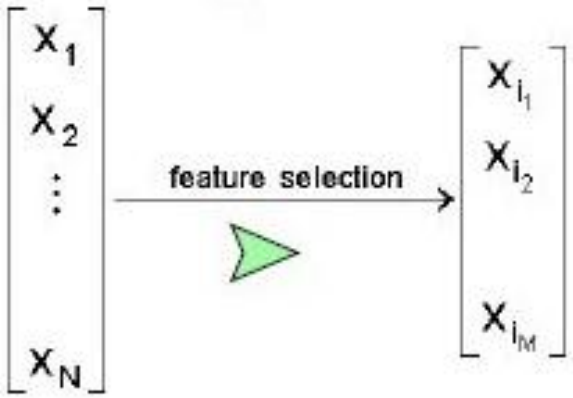
在这些原始特征中，有的特征对分类有效，有的则不起什么作用。 在得到一组原始特征后，若不加筛选， 全部用于分类函数确定，则有可能存在无效特征，这既增加了分类决策的复杂度，又不能明显改善分类器的性能。 为此，需要对原始特征集进行处理，去除对分类作用不大的特征，从而可以在保证性能的前提下，通过降低特征空间的维数来减少分类方法的复杂度。

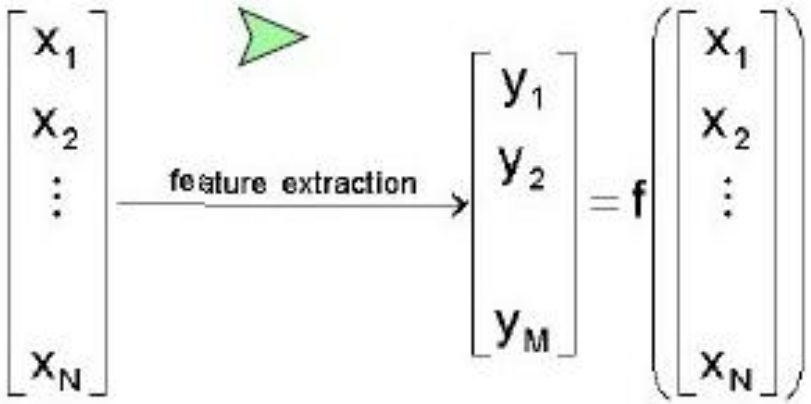
实现上述目的的方法有两种： 特征提取和特征选择。特征提取和特征选择都不考虑针对具体应用需求的原始特征形成过程，而是假设原始特征形成工作已经完成。 然而在实际工作中，原始特征的获得并不容易，因为人具有非常直观的识别能力，有时很难明确描述用于分类的特性依据。 如人脸的判定，人识别脸部特征非常容易，若用计算机来识别人脸，则需要得到多达上千个特征，难度很大。可以说，特征形成是模式识别过程中的重点和难点之一。

5. 特征提取和选择的作用

特征选择是指从一组特征中挑选出对分类最有利的特征，达到降低特征空间维数的目的。

特征提取是指通过映射 (或变换) 的方法获取最有效的特征，实现特征空间的维数从高维到低维的变换。 经过映射后的特征称为二次特征，它们是原始特征的某种组合，最常用的是线性组合。





从定义可以知，实现特征选择的前提是：确定特征是否有效的标准，在该标准下，寻找最有效的特征子集。用于特征选择的特征既可以是原始特征，也可以是经数学变换后得到的二次特征。需要注意，特征提取一定要进行数学变换，但数学变换未必就是特征提取。

特征提取和特征选择的主要目的都是在不降低或很少降低分类结果性能的情况下，降低特征空间的维数，其主要作用在于：

(1) 简化计算。 特征空间的维数越高，需占用的计算机资源越多，设计和计算也就越复杂。

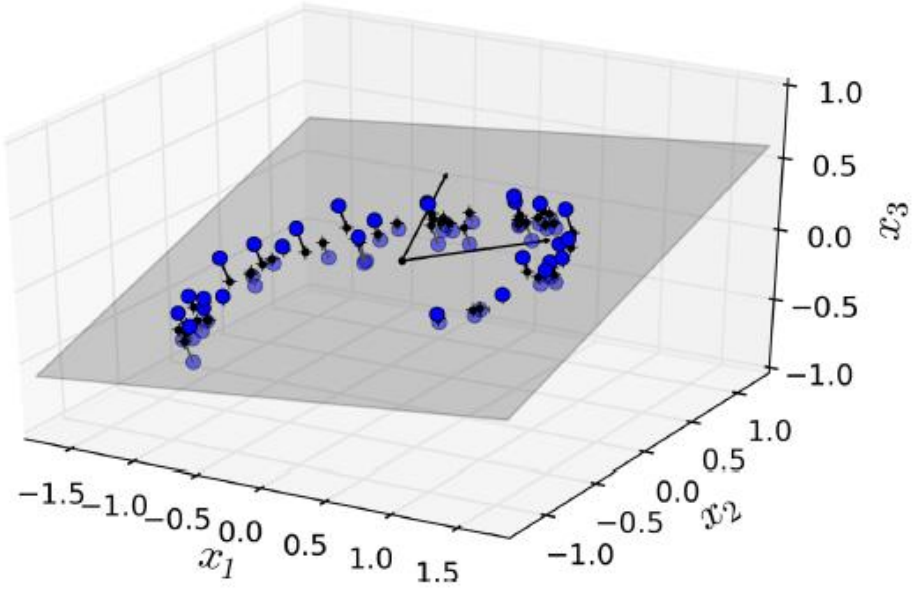
(2) 简化特征空间结构。 由于特征提取和选择是去除类间差别小的特征，保留类间差别大的特征，因此，在特征空间中，每类所占据的子空间结构可分离性更强， 从而也简化了类间分界面形状的复杂度。

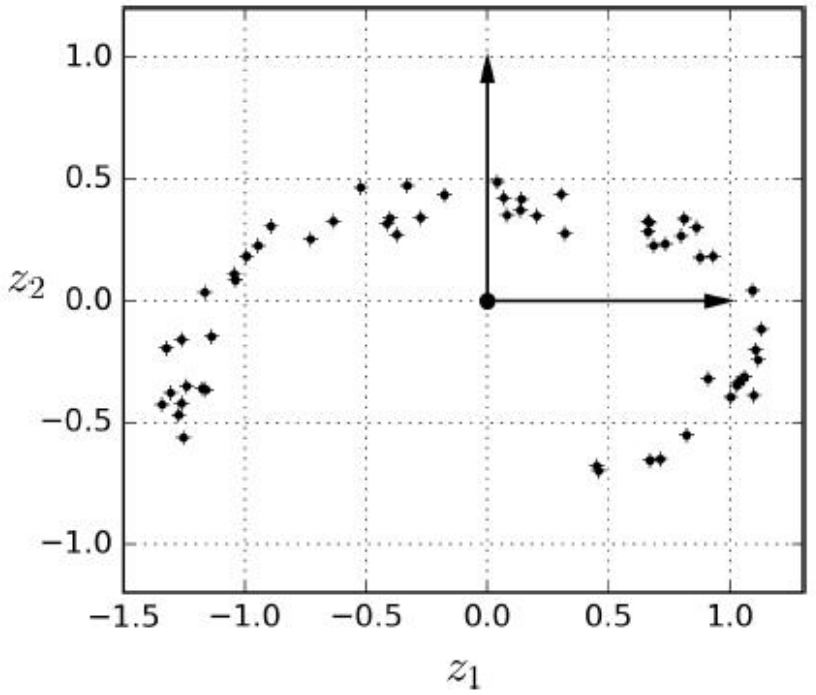
6. 降维的两大类方法

降维的第1大类方法：

投影 (特征提取，坐标变换，线性降维)

大多数实际问题中，样本在各个维度上的分布情况并不一致：在某个或某些维度，样本在一定范围内变动较大 (大方差)，但是在其他更多的维度，样本变动很小 (小方差)。这时，用线性或非线性变换，可以得到样本数据的低维表达。

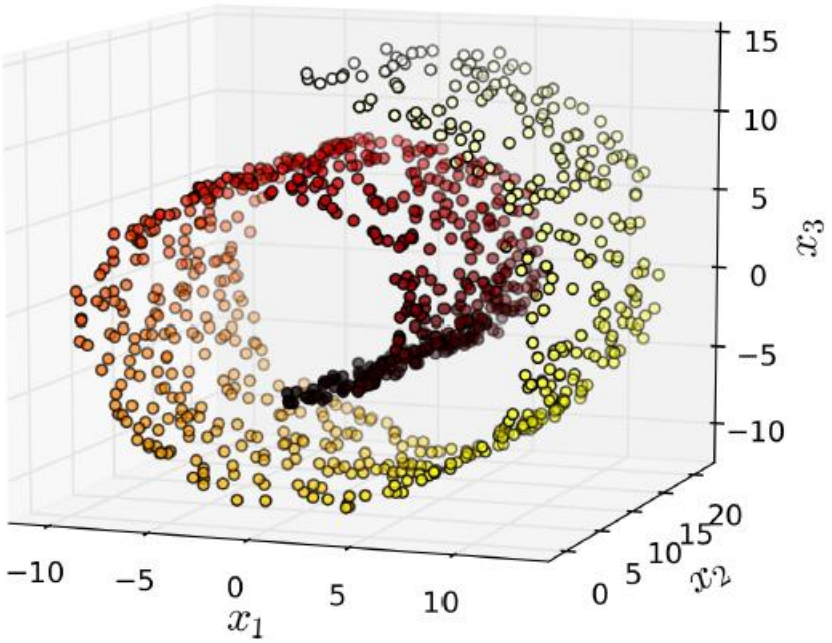


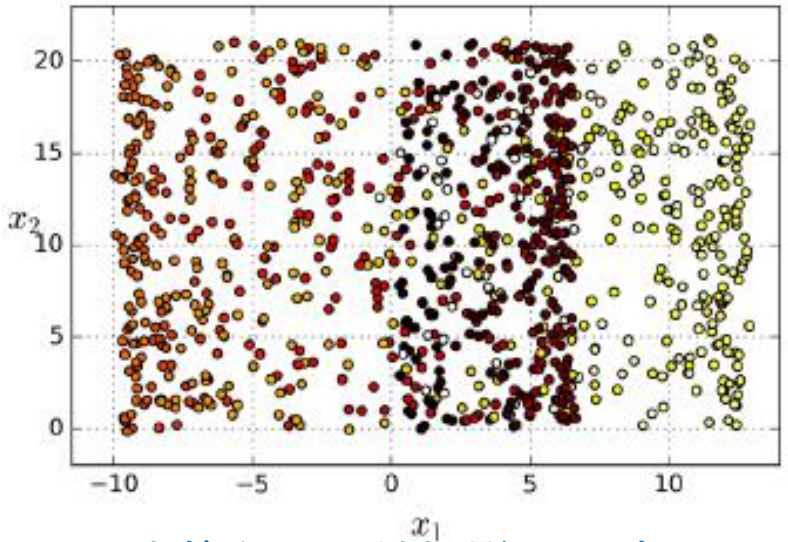


降维的第2大类方法：

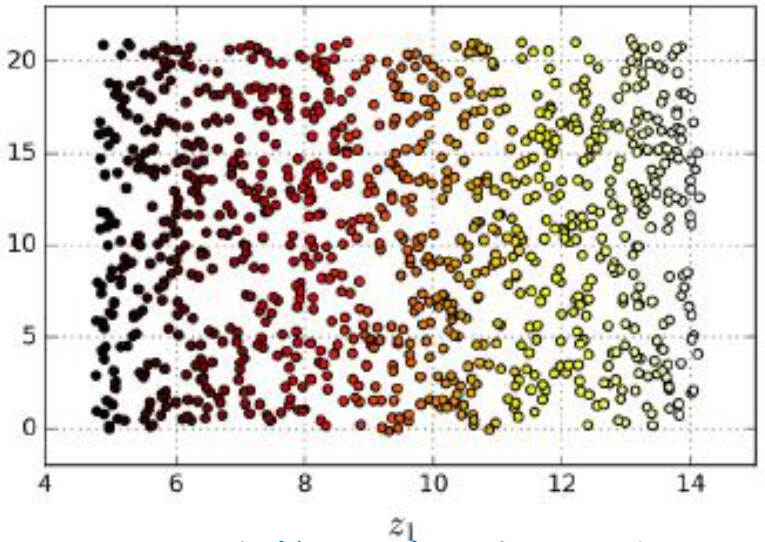
流形学习 (非线性降维)

投影 (线性变换) 并不总能有效：由于物理或几何性质，某些样本集虽然表现出低维性，但是其分布是扭曲或旋绕形态。





直接向x_1x_2轴投影，导致瑞士卷不同部分的重叠



用手“拉开”瑞士卷，再向x_1x_2轴投影，瑞士卷完美展开

9.2 类别可分性判据

判断特征有效性的一个自然想法是将分类器的错误概率作为度量的标准，即可用分类器错误概率(误差)最小的那组特征。理论上是正确的，但在实际应用中却存在很大的困难。

判断特征有效性的另一个想法是考虑依据某种类别可分性判据来评价。

类别可分性判据 (准则)： 刻画特征对分类的贡献。

设J_{ij}表示ω_i、ω_j两类间的可分性判据 (准则)，希望所构造的可分性判据J_{ij}满足以下要求：

1. 与分类器错误概率P(e) (或是错误概率的上、下界)有单调关系，J_{ij}最大值时，P(e)最小。

2. 非负性，即：

$$\begin{cases} J_{ij} > 0 & i \neq j \\ J_{ij} = 0 & i = j \end{cases}$$

3. 对称性, 即： $J_{ij} = J_{ji}$

该特性表明有效性判据对类别号没有方向性，而只强调对区分两类的贡献。

4. 特征独立时，判据满足可加性，即：

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

5. 单调性，当加入新特征时，判据不减少。

$$J_{ij}(x_1, x_2, \mathrm{L}\ , x_d) \leq J_{ij}(x_1, x_2, \mathrm{L}\ , x_d, x_{d+1})$$

所构造的可分性判据并不一定要求同时具有上述性质。

下面介绍常见的几种类别可分性判据。

1. 基于距离的可分性判据

基于距离的可分性判据直接依靠样本计算，直观简洁，物理概念清晰，因此目前应用较为广泛。基于距离的可分性判据的出发点是：各类样本之间的距离越大、类内离散度越小，则类别的可分性越好。

(1) 两类之间的距离

设两类为ω_i、ω_j， 分别有N_i、N_j个样本，即：

$$\omega_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \mathrm{L}\ , \mathbf{x}_{N_i}^i\}$$

$$\omega_j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \mathrm{L}\ , \mathbf{x}_{N_j}^j\}$$

两类之间的平均距离 $D_{\omega_i\omega_j}$ 可由下式定义：

$$D_{\omega_i\omega_j} = \frac{1}{N_i N_j} \sum_{r=1}^{N_i} \sum_{s=1}^{N_j} D(\mathbf{x}_r^i, \mathbf{x}_s^j)$$

其中，$D(x^i_r, x^j_s)$为某种定义下，x^i_r、x^j_s两个模式样本(向量)间的距离。 由点间距离的对称性可知，类间距离也具有对称性。

常用的点间距离有：欧氏距离、马氏距离、绝对距离(城市距离/曼哈顿距离、Hamming距离)、Minkowsky距离、切比雪夫距离等。

Hamming距离定义：

设$\mathbf{x}=(x_1,x_2,\ldots,x_d)^T$, $\mathbf{y}=(y_1,y_2,\ldots,y_d)^T$，有$D(\mathbf{x},\mathbf{y})=\Sigma|xi-yi|$;

在一个码组集合中，任意两个码字之间对应位上码元取值不同的位的数目定义为这两个码字之间的Hamming距离。即$D(\mathbf{x},\mathbf{y})=\sum x[i]\oplus y[i]$，这里$i=1,\ldots,d$，$\mathbf{x}$, $\mathbf{y}$都是d位的编码，$\oplus$ 表示异或。

例如，(00)与(01)的Hamming距离是1，(110)和(101)的Hamming距离是2。

在一个码组集合中，任意两个编码之间Hamming距离的最小值称为这个码组的最小Hamming距离。最小Hamming距离越大，码组抗干扰能力越强。

当取欧氏距离时，两类均方距离为：

$$D^2_{\omega_i\omega_j}=\frac{1}{N_iN_j}\sum_{r=1}^{N_i}\sum_{s=1}^{N_j}(\mathbf{x}_r^i-\mathbf{x}_s^j)^T(\mathbf{x}_r^i-\mathbf{x}_s^j)$$

(2) 各类模式之间的总的平均距离

设N个模式分别属于m类，$\omega_i=\{x^i_k, k=1, 2, \ldots, N_i\}$，$i=1, 2, \ldots, m$，各类模式之间的总的样本平均距离定义为：

$$J(\mathbf{x})=\frac{1}{2}\sum_{i=1}^{m}\hat{P}(\omega_i)\sum_{j=1}^{m}\hat{P}(\omega_j)\frac{1}{N_iN_j}\sum_{r=1}^{N_i}\sum_{s=1}^{N_j}D(\mathbf{x}_r^i,\mathbf{x}_s^j) \qquad (9.2.1)$$

其中，$\hat{P}(\omega_i)$是先验概率$P(\omega_i)$的估计，即：

$$\hat{P}(\omega_i)=N_i/N,\quad i=1,2,\ldots,m$$

$D(\mathbf{x}_r^i,\mathbf{x}_s^j)$是$\mathbf{x}_r^i$和$\mathbf{x}_s^j$间的距离.

当取欧氏距离时，总的均方距离为

$$J(\mathbf{x})=D_t^2(\mathbf{x})=\frac{1}{2}\sum_{i=1}^{m}\hat{P}(\omega_i)\sum_{j=1}^{m}\hat{P}(\omega_j)\frac{1}{N_iN_j}\sum_{r=1}^{N_i}\sum_{s=1}^{N_j}(\mathbf{x}_r^i-\mathbf{x}_s^j)^T(\mathbf{x}_r^i-\mathbf{x}_s^j)$$

(9.2.2)

(3) 类内离散度矩阵

类内离散度矩阵(散布矩阵， scatter matrix)表示各模式样本在本类的样本均值向量周围散布的情况。设 $\omega_i=\{x^i_k, k=1, 2, \ldots, N_i\}$， μ_i为ω_i类的样本均值向量，则ω_i类类内离散度矩阵/散布矩阵/协方差矩阵为：

$$S_{wi}=\frac{1}{N_i}\sum_{k=1}^{N_i}(\mathbf{x}^i_k-\boldsymbol{\mu}_i)(\mathbf{x}^i_k-\boldsymbol{\mu}_i)^T \qquad (9.2.3)$$

显然有：

$$tr(S_{wi})=\frac{1}{N_i}\sum_{k=1}^{N_i}(\mathbf{x}^i_k-\boldsymbol{\mu}_i)^T(\mathbf{x}^i_k-\boldsymbol{\mu}_i)$$

上式表明，类内离散度矩阵S_{wi}的迹等于该类类内均方欧氏距离.

说明：在“矩阵论(Matrix Theory)”中，N×N阶矩阵A的迹记为tr(A)，它定义为A的主对角线的元素之和。

(4) 多类情况下的多类类内、类间及总体离散度矩阵

设N个样本分别属于m类，$\omega_i=\{x^i_k, k=1, 2, \ldots, N_i\}$, i=1,2,…,m; S_{wi} 为ω_i类类内离散度矩阵。

多类类内离散度矩阵定义为:

$$S_w=\sum_{i=1}^{m}P(\omega_i)S_{wi}=\sum_{i=1}^{m}P(\omega_i)\frac{1}{N_i}\sum_{k=1}^{N_i}(\mathbf{x}_k^i-\boldsymbol{\mu}_i)(\mathbf{x}_k^i-\boldsymbol{\mu}_i)^T$$

类间离散度矩阵定义为:

$$S_b=\sum_{i=1}^{m}P(\omega_i)(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T=\frac{1}{N}\sum_{i=1}^{m}N_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T$$

总体离散度矩阵定义为:

$$S_t=\frac{1}{N}\sum_{l=1}^{N}(\mathbf{x}_l-\boldsymbol{\mu})(\mathbf{x}_l-\boldsymbol{\mu})^T$$

上面三式中,$P(\omega_i)$为ω_i类的概率, $\boldsymbol{\mu}_i$为ω_i类的样本均值向量, $\boldsymbol{\mu}$为总的样本均值向量, 它们分别是如下统计量:

$$P(\omega_i) = \frac{N_i}{N}$$

$$\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{k=1}^{N_i}\mathbf{x}_k^i$$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \sum_{i=1}^{m}P(\omega_i)\boldsymbol{\mu}_i = \frac{1}{N}\sum_{i=1}^{m}\sum_{k=1}^{N_i}\mathbf{x}_k^i$$

可证明:

$$S_t = S_w + S_b$$

$$J(\mathbf{x}) = tr(S_t) = tr(S_w + S_b) \qquad (9.2.4)$$

S_w, S_b和S_t为对称矩阵,而任意对称阵可经正交变换对角化,且对角线上元素为特征值.

由离散度矩阵的定义可知，此时对角线上的元素具有方差、均方距离等含义，且各分量不相关。正交变换为相似变换，变换后矩阵迹不变、行列式值也不变。因此，可以在原特征空间中用S_w、 S_b、 S_t的迹或行列式构造出许多可分性判据。

在概率统计中，$\boldsymbol{\mu}_i$、$\boldsymbol{\mu}$和S_w、 S_b 、S_t的定义或公式总结如下(*：了解)：

$$\boldsymbol{\mu}_i = \int \mathbf{x} p(\mathbf{x} \mid \omega_i) d\mathbf{x}$$

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\mathbf{S}_b = \sum_{i=1}^{m} P(\omega_i)(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

$$\mathbf{S}_w = \sum_{i=1}^{m} P(\omega_i) E_i\left[(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T\right] = \sum_{i=1}^{m} P(\omega_i) \int (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T p(\mathbf{x} \mid \omega_i) d\mathbf{x}$$

$$\mathbf{S_t = S_w + S_b}$$

为了使所使用的特征能够有效地进行分类，我们希望类间离散度S_b尽量大，同时类内离散度S_w尽量小，从直观上看可以构造下面各种判据：

$$J_1 = \frac{|\mathbf{S}_b|}{|\mathbf{S}_w|}$$

$$J_2 = tr(\mathbf{S}_w^{-1}\mathbf{S}_b)$$

$$J_3 = \ln\left[\frac{|\mathbf{S}_b|}{|\mathbf{S}_w|}\right]$$

$$J_4 = \frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}$$

$$J_5 = \frac{|\mathbf{S}_w + \mathbf{S}_b|}{|\mathbf{S}_w|}$$

为了有效地分类，它们的值越大越好。

基于距离的可分性判据虽然简单直观，但只是对于类间无重叠的情况效果较好，若类间存在重叠，则效果会受到影响。 基于概率的可分性判据能够较好地解决类间有重叠的问题。

2. 基于概率密度函数的可分性判据(*：选学)

基于概率密度函数的可分性判据主要考虑的是两类的概率分布情况。考虑图9.3所示两种极端情况，容易看出，图9.3(a)中两类是完全可分的，图9.3(b)中两类是完全不可分的， 两类概率密度函数的重叠程度反映了两类的可分性。因此，可以利用类条件概率密度函数构造可分性判据。

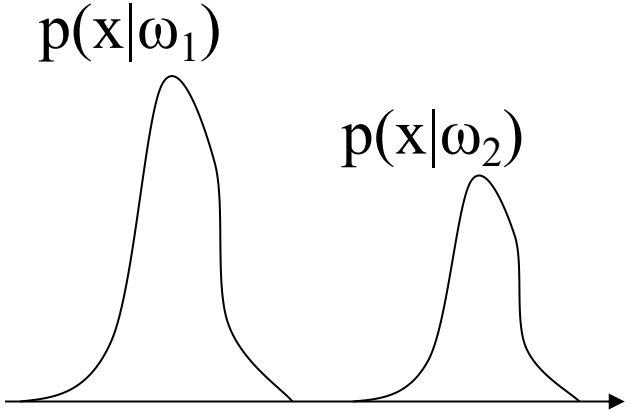


图9.3(a)

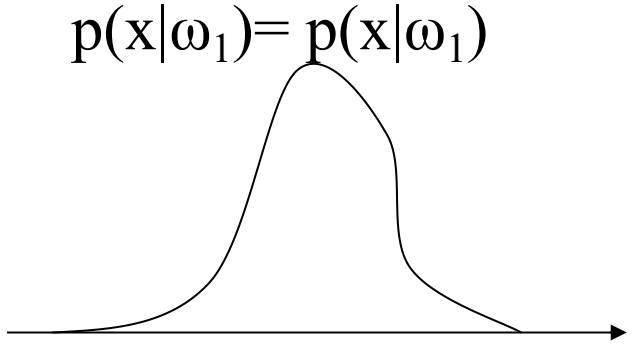


图9.3(b)

基于类条件概率密度函数$p(x|\omega_1)$、 $p(x|\omega_2)$的可分性判据J_p满足下面四个条件：

(1) 非负性

$$J_p \geq 0$$

(2) 对称性 ：相对于两个概密具有对称性。

$$J_p[p(\mathbf{x}|\omega_1), p(\mathbf{x}|\omega_2)] = J_p[p(\mathbf{x}|\omega_2), p(\mathbf{x}|\omega_1)]$$

(3) 最大值：当两类完全可分时，J_p具有最大值。

(4) 最小值：当两类完全不可分时，J_p具有最小值，即$J_p=0$。

设两类ω1和ω2的概率密度函数分别为$p(x|\omega 1)$、$p(x|\omega 2)$，$\mathbf{x}=(x_1,x_2,\ldots,x_n)^T$，下面构造三种典型的基于概率密度函数的可分性判据。

(1)巴氏(Bhattacharyya)判据J_B

Bhattacharyya判据计算式定义：

$$J_B = -\ln \int [p(\mathbf{x}\mid\omega_1)p(\mathbf{x}\mid\omega_2)]^{\frac{1}{2}} d\mathbf{x}$$

在最小错误概率判决准则下，最小错误概率Pe为：

$$P_e \le [P(\omega_1)P(\omega_2)]^{\frac{1}{2}} \exp(-J_B)$$

(证明此略)

(2) 切诺夫(Chernoff)判据J_C

Chernoff判据定义为：

$$J_c = -\ln \int p^s(\mathbf{x}\,|\,\omega_1)\, p^{1-s}(\mathbf{x}\,|\,\omega_2)\, d\mathbf{x} \qquad s \in [0,1]$$

由定义式可见，当s=1/2时，Chernoff界限距离就是巴氏 (Bhattacharyya) 距离。

一般情况下J_C的计算比较困难，当ω_1、 ω_2的类条件概率密度函数都是正态分布，即$p(\mathbf{x}|\omega 1)\sim N(\mu_i, \Sigma_i)$和$p(\mathbf{x}|\omega 2)\sim N(\mu_j, \Sigma_j)$时，可以推导出：

$$J_C = \frac{1}{2}s(1-s)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T[(1-s)\boldsymbol{\Sigma}_i + s\boldsymbol{\Sigma}_j]^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\frac{\left|(1-s)\boldsymbol{\Sigma}_i + s\boldsymbol{\Sigma}_j\right|}{\left|\boldsymbol{\Sigma}_i\right|^{1-s}\left|\boldsymbol{\Sigma}_j\right|^{s}} \tag{9.2.5}$$

当$\Sigma = \Sigma_i = \Sigma_j$时：

$$J_C = \frac{1}{2}s(1-s)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

思考题：

设有两类正态分布样本集，各类均值向量和方差分别为

$$\boldsymbol{\mu}_1=(0\ 0)^{\mathrm{T}},\Sigma_1=\begin{pmatrix}1 & 0\\ 0 & 0.25\end{pmatrix};\quad \boldsymbol{\mu}_2=(0\ 0)^{\mathrm{T}},\Sigma_2=\begin{pmatrix}0.25 & 0\\ 0 & 1\end{pmatrix}$$

试求两类的巴氏距离J_{B} 。

解：当J_C定义中的s=1/2时，J_C距离就是J_B距离。

$$s=1/2\text{时，}\ J_C=J_B=\frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\left[\frac{1}{2}\boldsymbol{\Sigma}_1+\frac{1}{2}\boldsymbol{\Sigma}_2\right]^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\frac{1}{2}\ln\frac{\left|\frac{1}{2}\boldsymbol{\Sigma}_1+\frac{1}{2}\boldsymbol{\Sigma}_2\right|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}$$

又由于$\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$，故$J_B=\frac{1}{2}\ln\frac{\left|\frac{1}{2}\boldsymbol{\Sigma}_1+\frac{1}{2}\boldsymbol{\Sigma}_2\right|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}=\frac{1}{2}\ln\frac{\frac{1.25\times1.25}{2}}{\sqrt{0.25}\times\sqrt{0.25}}$

$$=\frac{1}{2}\ln[1.25\times1.25\div0.25]=\frac{1}{2}\ln[(\frac{5}{4}\times\frac{5}{4})\div(\frac{1}{4})]=\ln\frac{5}{2}\approx0.92$$

(3) 散度J_D

在最小错误率Bayes决策中，对于两类的分类问题，最大后验概率判决准则可以通过似然比$p(\mathbf{x}|\omega 1)/p(\mathbf{x}|\omega 2)$和阈值$p(\omega 2)/p(\omega 1)$的比较实现，显然似然比对于分类来说是一个重要的度量。 对于给定的阈值$p(\omega 2)/p(\omega 1)$，$p(\omega 1|\mathbf{x})/p(\omega 2|\mathbf{x})$越大， 对类$\omega 1$来讲可分性越好，该比值反映了两类类条件概率密度函数的重叠程度。

为了保证概率密度函数完全重叠时判据为零，应对该比值取对数。 于是，可构造出D-判据J_D。

$\omega 1$类相对于$\omega 2$类的平均可分性信息定义为：

$$I_{12} = E\left[\ln \frac{p(\mathbf{x}\mid\omega_1)}{p(\mathbf{x}\mid\omega_2)}\right] = \int p(\mathbf{x}\mid\omega_1)\ln\frac{p(\mathbf{x}\mid\omega_1)}{p(\mathbf{x}\mid\omega_2)}\,d\mathbf{x}$$

ω_2类相对于ω_1类的平均可分性信息定义为：

$$I_{21} = E\left[\ln \frac{p(\mathbf{x}\mid\omega_2)}{p(\mathbf{x}\mid\omega_1)}\right] = \int p(\mathbf{x}\mid\omega_2)\ln\frac{p(\mathbf{x}\mid\omega_2)}{p(\mathbf{x}\mid\omega_1)}\,d\mathbf{x}$$

对于ω_1和ω_2两类总的平均可分性信息称为散度(Divergence), 其定义为：

$$J_D = I_{12} + I_{21}$$

$$= \int [p(\mathbf{x}\,|\,\omega_1) - p(\mathbf{x}\,|\,\omega_2)] \ln \frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} d\mathbf{x}$$

从数学构造上看，上式是合理的，式中被积函数两概密之差和两概密之比能反映两概密的重叠程度，同时被积函数中两因式总是同号，故其乘积非负。

3. 基于熵函数的可分性判据(*：选学)

由信息论知，对于一组概率分布而言，分布越均匀，平均信息量越大，分类的错误概率越大；分布越接近0-1分布，平均信息量越小，分类的错误概率越小，可分性越好。 因此，可以建立基于熵函数的可分性判据，其中熵函数表征平均信息量。

（具体内容，此略）

9.3 基于可分性判据的特征提取

设有n个原始特征构成的特征向量$\mathbf{x}=(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，特征提取就是对$\mathbf{x}$作线性变换，产生$d$维向量$\mathbf{y}=(y_1, y_2, \cdots, y_d)^{\mathrm{T}}$，$d \le n$，即：

$$\mathbf{y} = \boldsymbol{W}^{\mathrm{T}}\mathbf{x}$$

式中，$\boldsymbol{W}=\boldsymbol{W}_{n\times d}$称为特征提取矩阵或简称变换矩阵，$\mathbf{y}$称为二次特征。

基于可分性判据的特征提取就是在一定的准则函数(可分性判据)下，如何求最优的变换矩阵$\boldsymbol{W}^*$。

1. 基于距离可分性判据的特征提取方法

前面研究了基于距离的可分性判据，得到了相应判据，它们都反映了一个基本思想，即类内距离小和类间距离大的要求。 下面我们以*J2*准则 $J_2 = tr(\mathbf{S}_w^{-1}\mathbf{S}_b)$为例讨论特征提取的方法。

设S_w和S_b为原始特征空间的多类类内离散度矩阵和类间离散度矩阵，S_w^*和S_b^*为变换后特征空间的多类类内离散度矩阵和类间离散度矩阵，W为变换矩阵。 则有：

$$\mathbf{S}_w^* = \mathbf{W}^T\mathbf{S}_w\mathbf{W}$$

$$\mathbf{S}_b^* = \mathbf{W}^T\mathbf{S}_b\mathbf{W}$$

在变换域中，J_2为：

$$J_2(\mathbf{W}) = tr[(\mathbf{S}_w^*)^{-1}\mathbf{S}_b^*] = tr[(\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})]$$

若$\mathbf{W}$为非奇异矩阵, 可得$tr[(\mathbf{S}_w^*)^{-1}\mathbf{S}_b^*] = tr[\mathbf{S}_w^{-1}S_b]$, J_2是不变的。

对矩阵作相似变换特征值不变, 其行列式值不变, 其迹不变, 一个方阵的迹等它的所有特征值之和。

设$\mathbf{W}_e$标准正交阵, 用W_e对对称阵$\mathbf{S}_w^{-1}S_b$作相似变换使其成为对角阵:

$$\mathbf{W}_e^{-1}\mathbf{S}_w^{-1}S_b\mathbf{W}_e=\mathbf{W}_e^{T}\mathbf{S}_w^{-1}S_b\mathbf{W}_e=\begin{pmatrix}\lambda_1 & \mathrm{K} & 0\\ \mathrm{M} & \mathrm{O} & \mathrm{M}\\ 0 & \mathrm{K} & \lambda_n\end{pmatrix}=diag(\lambda_1,\ldots,\lambda_n)$$

其中, $\lambda_i(i=1,2,\ldots,\mathrm{n})$为$\mathbf{S}_w^{-1}S_b$的特征值, $\mathbf{W}_e$的列向量$\mathbf{w}_i$为矩阵$\mathbf{S}_w^{-1}S_b$相应于λ_i的特征向量。

设此处$\mathbf{W}_e$的列向量的排列已作适当调整, 使矩阵$\boldsymbol{S}^{-1}{}_w\boldsymbol{S}_b$的特征值$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$。

由此可得出，在d给定后，取前d个较大的特征值所对应的特征向量$\mathbf{w}_i(i=1,2,\ldots,d)$构造特征提取矩阵，即：

$$\mathbf{W}=(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_d)$$

对$\mathbf{x}$作变换$\mathbf{y}=\mathbf{W}^T\mathbf{x}$，这时对于给定的$d$所得到的$J_2^*$

$$J_2^*(\mathbf{W})=\sum_{i=1}^{d}\lambda_i$$ 达到最大值

此法对于J_4判据也适用。

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

$$\lambda_1 \geq \lambda_2 \geq \mathrm{L} \geq \lambda_n$$

相应的正交化、归一化的特征向量为：

$$\mathbf{w}_1, \mathbf{w}_2, \mathrm{L}\ , \mathbf{w}_n$$

选前d个特征向量作为变换矩阵：

$$\mathbf{W} = \left[\mathbf{w}_1, \mathbf{w}_2, \mathrm{L}\ , \mathbf{w}_d\right]_{n\times d}$$

2. 基于概率密度函数可分性判据的特征提取方法 (*: 选学)

基于概率密度函数的可分性判据的方法需要知道各类的概率密度函数的解析形式，难度较大，计算量也较大。一般地，只有当概率密度函数为某些特殊的函数形式时才便于使用，这里只研究多元正态分布的两类问题。

对于基于概率密度函数可分性判据的特征提取方法而言，通常选用的变换仍为线性变换，设n维原始特征向量$\mathbf{x}$经线性变换后的二次特征向量为$\mathbf{y}$，即：

$$\mathbf{y} = \mathbf{W}^T \mathbf{x}$$

在映射后的特征空间内建立某种准则函数，使得它为变换矩阵$\mathbf{W}$的函数：

$$J_C = J_C(\mathbf{W})$$

其中，J_C为基于概率密度函数的可分性判据，如前面介绍的Bhattacharyya距离和Chernoff 距离等可分性判据。通过求解判据的极值点即可得到使映射后的特征组可分性最好的变换矩阵。 在$J_C(\mathbf{W})$可微的情况下，就是求解偏微分方程：

$$\frac{\partial J_C(\mathbf{W})}{\partial \mathbf{W}} = \mathbf{0}$$

这里以Chernoff距离为例，分析特征提取方法。 当两类都是正态分布时，两类的分布函数分别为：

$$p(\mathbf{x}\mid\omega_1)=\frac{1}{(2\pi)^{n/2}\left|\sum_1\right|^{1/2}}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\sum_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)]$$

$$p(\mathbf{x}\mid\omega_2)=\frac{1}{(2\pi)^{n/2}\left|\sum_2\right|^{1/2}}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\sum_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)]$$

变换后的判据J_C是$\mathbf{W}$的函数，记为$J_C(\mathbf{W})$，根据前面的(9.2.5)式并利用矩阵乘积求迹的性质，有：

$$J_C(\mathbf{W})=\frac{1}{2}s(1-s)tr\{\mathbf{W}^T\mathbf{M}\mathbf{W}[(1-s)\mathbf{W}^T\boldsymbol{\Sigma}_1\mathbf{W}+s\mathbf{W}^T\boldsymbol{\Sigma}_2\mathbf{W}]^{-1}\}$$

$$+\frac{1}{2}\ln\left|(1-s)\mathbf{W}^T\boldsymbol{\Sigma}_1\mathbf{W}+s\mathbf{W}^T\boldsymbol{\Sigma}_2\mathbf{W}\right|$$

$$-\frac{1}{2}(1-s)\ln\left|\mathbf{W}^T\boldsymbol{\Sigma}_1\mathbf{W}\right|-\frac{1}{2}s\ln\left|\mathbf{W}^T\boldsymbol{\Sigma}_2\mathbf{W}\right|$$

式中，$M=(\mu_1-\mu_2)(\mu_1-\mu_2)^T$。

因为$J_c(\mathbf{W})$是标量，可以对$\mathbf{W}$的各个分量求偏导，并令其为零，经简化可得矩阵方程：

$$\mathbf{MW}-[(1-s)\boldsymbol{\Sigma}_1\mathbf{W}+s\boldsymbol{\Sigma}_2\mathbf{W}][(1-s)\mathbf{W}^T\boldsymbol{\Sigma}_1\mathbf{W}+s\mathbf{W}^T\boldsymbol{\Sigma}_2\mathbf{W}]^{-1}\mathbf{W}^T\mathbf{MW}$$
$$+\boldsymbol{\Sigma}_1\mathbf{W}[I-(\mathbf{W}^T\boldsymbol{\Sigma}_1\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Sigma}_2\mathbf{W}]+\boldsymbol{\Sigma}_2\mathbf{W}[\mathbf{I}-(\mathbf{W}^T\boldsymbol{\Sigma}_2\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Sigma}_1\mathbf{W}]=\mathbf{0}$$

上式是关于$\mathbf{W}$的非线性方程，只能采用数值优化的方法得到近似最优解。

在以下两种特殊情况下可以得到最优的解析解。

(1) $\Sigma_1=\Sigma_2=\Sigma, \mu_1\neq\mu_2$

在此种情况下，最优特征提取矩阵是由$\Sigma^{-1}M$矩阵的特征向量构成的。又因为矩阵M的秩为1，故$\Sigma^{-1}M$只有一个非零特征值，对应于特征值为零的那些特征向量对$J_c(\mathbf{W})$没有影响，因此可以舍去，所以最优变换$\mathbf{W}$是$\Sigma^{-1}M$的非零特征值对应的特征向量v，不难得到：

$$\mathbf{W}=v=\Sigma^{-1}(\mu_1-\mu_2)$$

上面结果与Fisher线性判别式的解相同。

(2) $\Sigma_1 \neq \Sigma_2, \mu_1 = \mu_2$

在此种情况下，最优特征矩阵W^*是由$\Sigma^{-1}{}_2\Sigma_1$满足下列关系的前d个特征值所对应的特征向量构成的，此时$J_c(\mathbf{W})$取最大值。

$$(1-s)\lambda_1^s + s\lambda_1^{s-1}$$

$$\geq (1-s)\lambda_2^s + s\lambda_2^{s-1}$$

$$\geq \ldots \geq (1-s)\lambda_n^s + s\lambda_n^{s-1}$$

9.4 主成分分析 (PCA)

主成分分析(PCA, Principal Components Analysis)是一种有效的、广泛使用的无监督线性降维方法，也称为*K-L*变换(Karhunen-Loeve Transform)/主分量变换/Hotelling变换。 *K-L*变换是一种基于目标统计特性的最佳正交变换，它的最佳性体现在变换后产生的新的分量正交或不相关。

K-L变换分连续和离散两种情况，这里只讨论离散K-L变换法。

设n维随机向量$\mathbf{x}=(x_1, x_2, \ldots, x_n)^{\mathrm{T}}$，$\mathbf{x}$经标准正交矩阵$A$正交变换后成为向量$\mathbf{y}=(y_1, y_2, \ldots, y_n)^{\mathrm{T}}$，即：

$$\mathbf{y} = \mathbf{A}^T\mathbf{x} \qquad (9.4.1)$$

$\mathbf{y}$的自相关矩阵为：

$$\mathbf{R}_{\mathrm{y}} = E(\mathbf{y}\mathbf{y}^T) = E[\mathbf{A}^T\mathbf{x}\mathbf{x}^T\mathbf{A}] = \mathbf{A}^T\mathbf{R}_{\mathrm{x}}\mathbf{A}$$

其中，R_{x}为$\mathbf{x}$的自相关矩阵(auto-correlation matrix)，

即$R_{\mathrm{x}}=E(\mathbf{x}\mathbf{x}^T) \approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$，是对称矩阵。

选择矩阵$A=(a_1, a_2, \ldots, a_n)$，且满足：

$$\mathbf{R}_x\mathbf{a}_i = \lambda_i\mathbf{a}_i$$

这里，λ_i为自相关矩阵R_x的特征值，并且$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$，$a_i$为$\lambda_i$的正交基向量(特征向量)，即$a^T_i a_j=1(i=j)$，$a^T_i a_j=0(i\neq j; i, j=1, 2, \ldots, n)$。 R_y是对角矩阵：

$$\mathbf{R}_y = \mathbf{A}^T\mathbf{R}_x\mathbf{A} = diag(\lambda_1, \lambda_2, \mathrm{L}\ \ , \lambda_n)$$

若R_x是正定的，则它的特征值是正的。 此时变换式$y=A^Tx$称为K-L变换。

由式(9.4.1)式可得：

$$\mathbf{x}=(\mathbf{A}^{T})^{-1}\mathbf{y}=\mathbf{A}\mathbf{y}=(\mathbf{a}_1,\mathbf{a}_2,\mathrm{L}\ \ ,\mathbf{a}_n)\begin{pmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_n \end{pmatrix}=\sum_{i=1}^{n} y_i\mathbf{a}_i$$

选择**x**关于$\boldsymbol{a_i}$的展开式的前***d***项在最小均方误差准则下估计 $\hat{\mathbf{x}}$，此时估计式表示为：

$$\hat{\mathbf{x}}=\sum_{i=1}^{d} y_i\mathbf{a}_i\ ,\quad (1\le d\le n)$$

估计的均方误差为：

$$\varepsilon^2(d) = E\left[(\mathbf{x}-\hat{\mathbf{x}})^T(\mathbf{x}-\hat{\mathbf{x}})\right]$$

$$= E\left[\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|^2\right] = E\left[\left\|\sum_{i=d+1}^{n}\mathbf{a}_i y_i\right\|^2\right]$$

$$= \sum_{i=d+1}^{n} E[y_i^2] = \sum_{i=d+1}^{n} E[y_i\, y_i^T] = \sum_{i=d+1}^{n} E[(\mathbf{a}_i^T\mathbf{x})(\mathbf{x}^T\mathbf{a}_i)]$$

$$= \sum_{i=d+1}^{n} \mathbf{a}_i^T E[\mathbf{x}\mathbf{x}^T]\mathbf{a}_i = \sum_{i=d+1}^{n} \lambda_i$$

希望选择使估计的均方误差最小的特征向量，因此要选择相关矩阵$\boldsymbol{R}_{\mathbf{x}}$的$d$个最大的特征值对应的特征向量构成变换矩阵$\mathbf{A}$，这样得到的均方误差将会最小，是$n$-$d$个极小特征值之和。 可以证明，与$d$维向量中的其他$\mathbf{x}$逼近值相比，这个结果是最小均方误差解。 这就是**K-L**变换也称为主分量分析(**PCA**)的原因。

基于K-L变换/主成分分析的特征提取算法描述：

设$\mathbf{x}$是n维模式向量，$\{\mathbf{x}\}$是来自m个模式类的样本集，总样本数为N.

利用K-L变换将$\mathbf{x}$变换为d维$(d < n)$向量$\mathbf{y}$的具体步骤如下：

(1) 平移坐标系，将总体均值向量作为新坐标系的原点；

(2) 求出自相关矩阵$\mathbf{R}_{\mathbf{x}} = E(\mathbf{x}\mathbf{x}^T) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T$；(或协方差矩阵等)

(3)求出$\mathbf{R}_{\mathbf{x}}$的特征值$\lambda_1, \lambda_2, ..., \lambda_n$及其对应的特征向量$\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n$；

(4)将特征值从大到小排序，如：

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$$

取前d个大的特征值所对应的特征向量构成变换矩阵

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_d)$$

(5)将n维的原向量变换成d维的新向量$\mathbf{y} = \mathbf{A}^T\mathbf{x}$

例：已知模式样本数据：

$$\begin{pmatrix}-5\\-5\end{pmatrix},\begin{pmatrix}-5\\-4\end{pmatrix},\begin{pmatrix}-4\\-5\end{pmatrix},\begin{pmatrix}-5\\-6\end{pmatrix},\begin{pmatrix}-6\\-5\end{pmatrix},\begin{pmatrix}5\\5\end{pmatrix},\begin{pmatrix}5\\6\end{pmatrix},\begin{pmatrix}6\\5\end{pmatrix},\begin{pmatrix}5\\4\end{pmatrix},\begin{pmatrix}4\\5\end{pmatrix}$$

(a) 试用PCA/K-L变换作一维数据降维处理；

(b) 编制PCA一维数据降维处理程序(Python/R/MATLAB)。

(a)解： (1)求样本总体均值向量

$$\bar{\mathbf{x}} = \frac{1}{10}\left[\begin{pmatrix}-5\\-5\end{pmatrix}+\begin{pmatrix}-5\\-4\end{pmatrix}+\ldots+\begin{pmatrix}4\\5\end{pmatrix}\right]=\begin{pmatrix}0\\0\end{pmatrix}$$

故无需作坐标系平移。

(2) 求自相关矩阵$\mathbf{R}_x$

$$\mathbf{R}_x = \frac{1}{10}\left[\begin{pmatrix}-5\\-5\end{pmatrix}\begin{pmatrix}-5 & -5\end{pmatrix}+\ldots+\begin{pmatrix}4\\5\end{pmatrix}\begin{pmatrix}4 & 5\end{pmatrix}\right]$$

$$=\begin{pmatrix}25.4 & 25.0\\25.0 & 25.4\end{pmatrix}$$

(3) 求$\mathbf{R}_x$的特征值及其对应的特征向量

$$\left|\lambda\mathbf{I}-\mathbf{R}_x\right|=\begin{vmatrix}25.4-\lambda & 25.0\\ 25.0 & 25.4-\lambda\end{vmatrix}=0$$

即: $(25.4-\lambda)^2-25.0^2=0,$

解得特征值: $\lambda_1=50.4, \qquad \lambda_2=0.4$。

由$\mathbf{R}_x\mathbf{a}_i=\lambda_i\mathbf{a}_i(i=1,2)$, 可解出特征向量为

$$\mathbf{a}_1=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \quad ,\mathbf{a}_2=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}$$

(4) 取$\mathbf{a}_1$作变换矩阵$\mathbf{A}$

$$\mathbf{A}=\mathbf{a}_1$$

(5) 将原二维样本变为一维样本$\mathbf{y}=A^T\mathbf{x}$：

$$\left(-\frac{10}{\sqrt{2}}\right),\left(-\frac{9}{\sqrt{2}}\right),\left(-\frac{9}{\sqrt{2}}\right),\left(-\frac{11}{\sqrt{2}}\right),\left(-\frac{11}{\sqrt{2}}\right),$$

$$\left(\frac{10}{\sqrt{2}}\right),\left(\frac{11}{\sqrt{2}}\right),\left(\frac{11}{\sqrt{2}}\right),\left(\frac{9}{\sqrt{2}}\right),\left(\frac{9}{\sqrt{2}}\right)$$

(b) 本例K-L变换Python程序 (使用scikit-learn)：

```
#Filename: PCA_Example.ipynb
#Import Library
import numpy as np
from sklearn.decomposition import PCA
# Assumed you have training and test data set as train and test
train=np.array([[-5,-5],[-5,-4],[-4,-5],[-5,-6],[-6,-5],[5,5],[5,6],[6,5],[5,4],[4,5]])
train_trans_cor= train - np.mean(train) #把样本数据平移到均值向量为中心的坐标系下
# Create PCA obeject pca=PCA(n_components=k)
# default value of k =min(n_sample, n_features)
# For Factor analysis
# fa=decomposition.FactorAnalysis()
pca=PCA(n_components=1)   #保留数据的第1个主成分
# Reduced the dimension of training dataset using PCA
train_reduced = pca.fit_transform(train_trans_cor)
print(train_reduced)
# Reduced the dimension of test dataset
# test_reduced = pca.transform(test)
```

本例K-L变换MALAB程序(*：了解)：

```
%Filename: PCA_Example.m
X=[-5 -5
   -5 -4
   -4 -5
   -5 -6
   -6 -5
    5  5
    5  6
    6  5
    5  4
    4  5];
[m1 n1]=size(X);
m=mean(X);          %均值向量
```

```
for i=1:m1
   X(i,:)=X(i,:)-m;     %把样本数据平移到m为中心的坐标系下
end
R=zeros(n1);
for i=1:m1
R=X(i,:)'*X(i,:)+R;
end
R=R/10;    %样本X的自相关矩阵R
[A,D]=eig(R) %计算矩阵R的特征值对角阵D及其对应的特征向量矩阵A
%将原二维样本变换成一维样本(这里简化了按特征值排序等操作)
if D(1,1)>D(2,2)   y=A(:,1)'*X'
else y=A(:,2)'*X'
end
```

运行结果：

```
A =
   -0.7071    0.7071
    0.7071    0.7071

D =
    0.4000         0
         0    50.4000

y =

Columns 1 through 9

-7.0711   -6.3640   -6.3640   -7.7782   -7.7782    7.0711
7.7782    7.7782    6.3640
Column 10

6.3640
```

MATLAB中使用函数princomp (Principal Components Analysis)实现了对PCA的封装。

princomp常见调用形式：

[COEFF, SCORE, latent]=princomp(X)

参数说明：

X：原始样本矩阵，其中的1行表示1个样本，其中的1列表示样本特征向量的1维；

COEFF：主成分分量，即样本自相关矩阵的特征向量a_1~a_n；

SCORE：主成分，X的低维表示，即X中的数据在主成分分量上的投影a_1~a_d；

latent：一个包含样本自相关矩阵特征值的向量$[\lambda_1 \sim \lambda_d]^T$。

例：使用princomp作主分量分析。

```
%Filename:ex7_2.m
%使用princomp函数作主成分分析
X=[-5 -5
   -5 -4
   -4 -5
   -5 -6
   -6 -5
    5  5
    5  6
    6  5
    5  4
    4  5];
[A,SCORE,latent]=princomp(X);%主成分分析
A  %A主成分分量，即A中的各列为样本X自相关矩阵的特征向量a1~a2
SCORE  %SCORE主成分,SCORE(:,1)为X在变换空间的一维表示, SCORE为X在变换空间的二维表示
latent %X样本自相关矩阵的特征值
```

运行结果：

```
A =

    0.7071    0.7071
    0.7071   -0.7071
SCORE =

   -7.0711   -0.0000
   -6.3640   -0.7071
   -6.3640    0.7071
   -7.7782    0.7071
   -7.7782   -0.7071
    7.0711   -0.0000
    7.7782   -0.7071
    7.7782    0.7071
    6.3640    0.7071
    6.3640   -0.7071
latent =

   56.0000
    0.4444
```

上面主成分分析采用的是样本自相关矩阵，也可以采用样本协方差矩阵作主分量。

基于样本协方差矩阵的主成分分析在图像分析中使用较广泛。

设有n×d维样本矩阵X(n个样本, 每个样本有d 维特征)。

其协方差矩阵C为:

$$C=cov(X)=E\{(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\}\approx\frac{1}{\mathrm{n}-1}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^{T}$$

其散布矩阵/离散度矩阵S为:

$$S=\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^{T}=\mathrm{cov}(\mathbf{X})*(n-1)$$

例：PCA在人脸识别中的应用。

在人脸识别中，PCA是一种常用的特征提取方法。设一幅$p\times q$大小的人脸图像，可以将它看成是一个矩阵$(f_{ij})_{p\times q}$，f_{ij}为图像在该点的灰度(亮度)。若将该矩阵按列相连构成一个$p\times q$维向量$\mathbf{x}_i=(f_{11}, f_{21}, \cdots, f_{p1}, f_{12}, f_{22}, \cdots, f_{p2}, \cdots, f_{1q}, f_{2q}, \cdots, f_{pq})^{\mathrm{T}}$。设训练样本集为 $X=\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$，包含N幅图像。

N幅图像的协方差矩阵为：

$$\mathbf{R}=\frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T$$

其中，$\overline{\mathbf{x}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$

求出矩阵R的前d个最大特征值$\lambda_1, \lambda_2, \ldots, \lambda_d$及其对应的正交化、归一化特征向量$\alpha_1, \alpha_2, \ldots, \alpha_d$。分别将这$d$个特征向量化为p×q矩阵，得到d幅图像，称为“特征脸”(eigenface)。

下图(主要参考文献[01]P84图5-2)显示的是对应前30个最大特征值的特征向量的图像 (*：了解)。

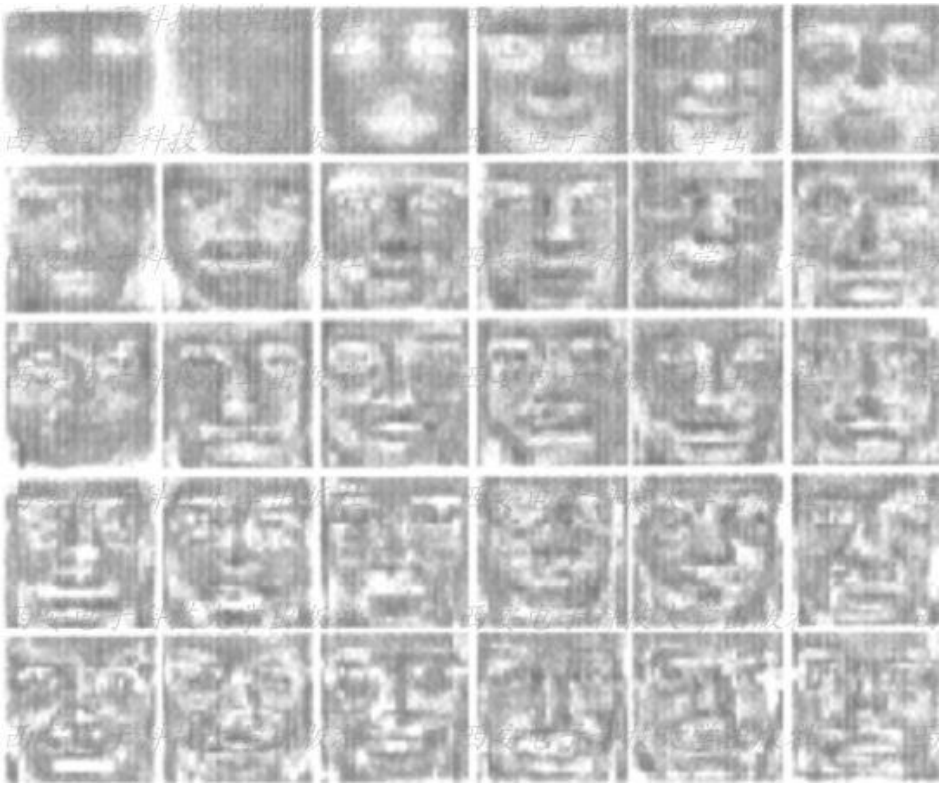


图　“特征脸(eigenface)”图像

将每一幅人脸图像投影到由$a_1, a_2, \ldots, a_d$张成的子空间中，对应于该子空间的一个点，该点的坐标系数对应于图像在子空间的位置，可以作为识别人脸的依据。对于任意待识别人脸图像模式样本$\mathbf{x}$，可通过向“特征脸”子空间投影获得系数向量$\mathbf{y}=(a_1, a_2, \ldots, a_d)^{\mathrm{T}}\mathbf{x}$。

点评：一幅人脸图像往往是由较多的像素构成的，如果以每个像素作为1维特征，将得到一个维数相当高的特征向量，计算将十分困难；而且这些像素之间通常具有相关性。这样，利用PCA技术（或其它特征提取技术）在降低维数的同时在一定程度上去除原始特征各维之间的相关性自然成为一个较理想的解决方案。

人脸图像预处理方法 (*: 选学):

研究和设计人脸识别算法，通常需要引用国际上的一些标准人脸库，ORL人脸库是其中的一种。

1.ORL人脸库(ORL Database of Faces)简介

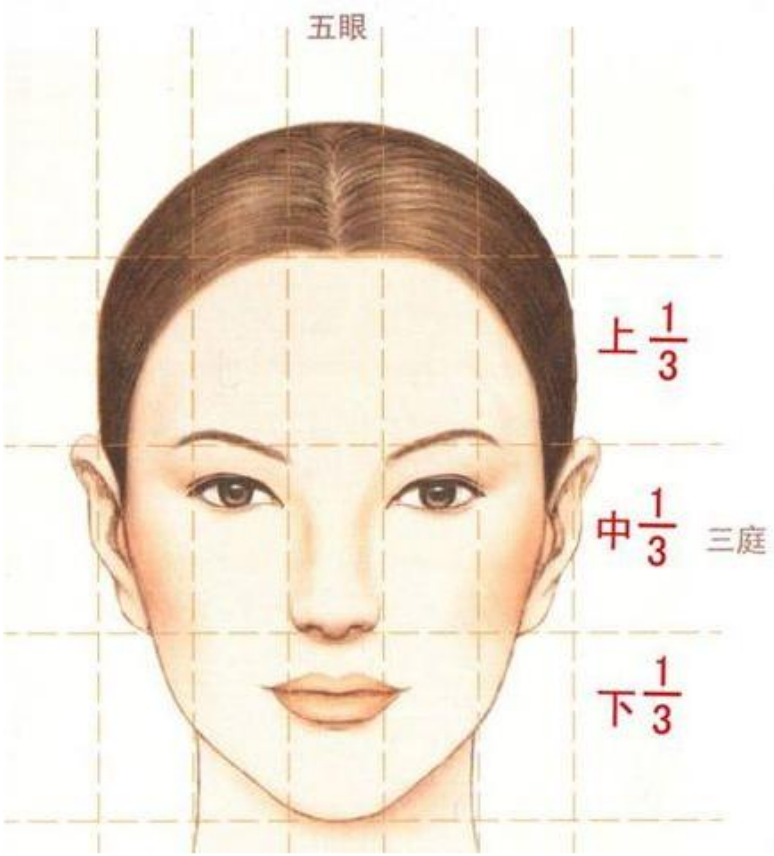
ORL人脸数据库由英国剑桥大学的AT&T实验室采集。

(1)ORL数据库共有400幅人脸图像(40人，每人10幅，大小为112像素×92像素，灰度级为256级)。

(2) 该数据库比较规范，大多数图像的光照方向和强度都差不多。但有少许表情、姿势、伸缩的变化，眼睛对得不是很准，尺度差异在10%左右。

(3)不是每个人都有所有的这些变化的图像，即有些人姿势变化多一点，有些人表情变化多一点，有些还戴有眼睛，但这些变化都不大。

正是基于ORL人脸库图像在光照，以及关键点如眼睛、嘴巴的位置等方面比较规范，一些实验可在该图像集上直接展开（可以省略归一化等图像预处理过程，但归一化后的图像识别率会更高一些）。在用ORL进行人脸识别算法研究时，通常采用一部分图像做训练集(如选用每个人的前5张图像作为训练集，这样40个人共有200幅样本图像)，剩下的另外一部分图像做测试集(如选用每个人的后5张图像作为测试集，这样40个人共有200幅待测试图像)。

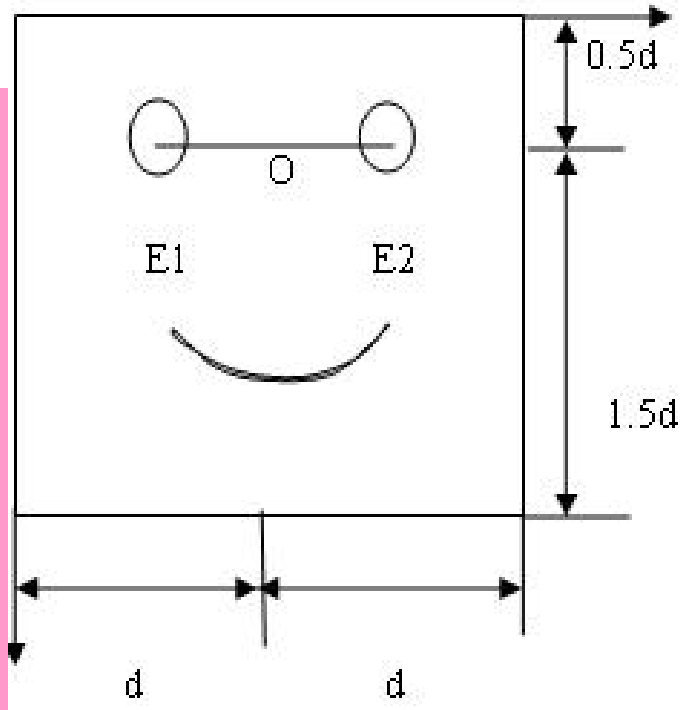


关于五官的比例(见上图)，中国古代有三庭五眼之说，即长三庭，横五眼（此定义仅适用于有亚洲特色的亚洲人）。三庭：发际线至眉毛=眉毛至鼻孔=鼻孔至下巴的距离；五眼：右外耳孔至右眼外角之长=右眼长=眼间距离=左眼长=左外耳孔至左眼外角之长，这种简单的概括，是成人的一般比例关系。

2.人脸图像的预处理

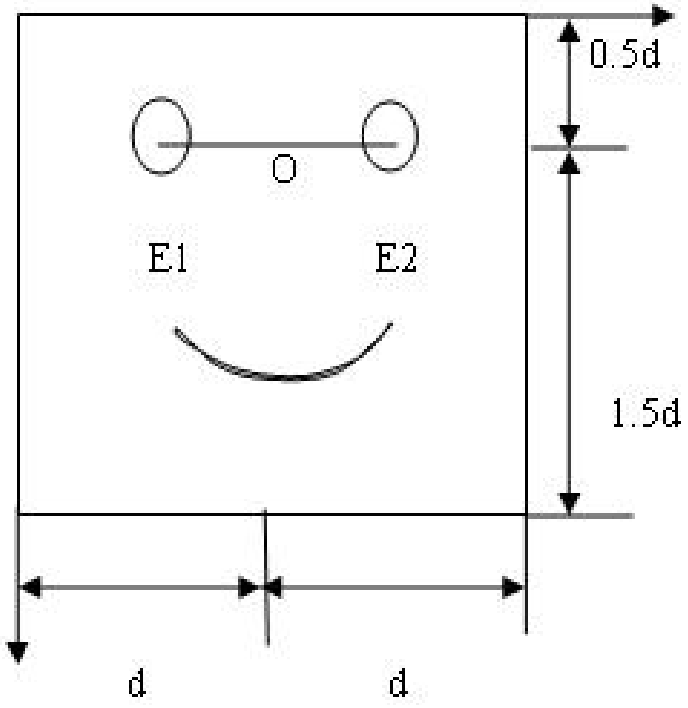
对于不规范的人脸图像，在进行特征提取之前必须对其作预处理(如去噪、几何归一化)。由于两眼之间的距离对大多数人来说都是基本相同的，因此可采用两眼的位置作为几何归一化的依据。

假设人脸图像中两只眼睛的位置分别是**E1**和**E2**(右图所示)。通过如下三个步骤，便能实现人脸的几何归一化。即输入一幅人脸图像，通过旋转、裁剪、缩放，便得到一个双眼水平、间距为**d**，图像比例为**2d×2d**大小，大小为**32×32**的统一规范化图像。



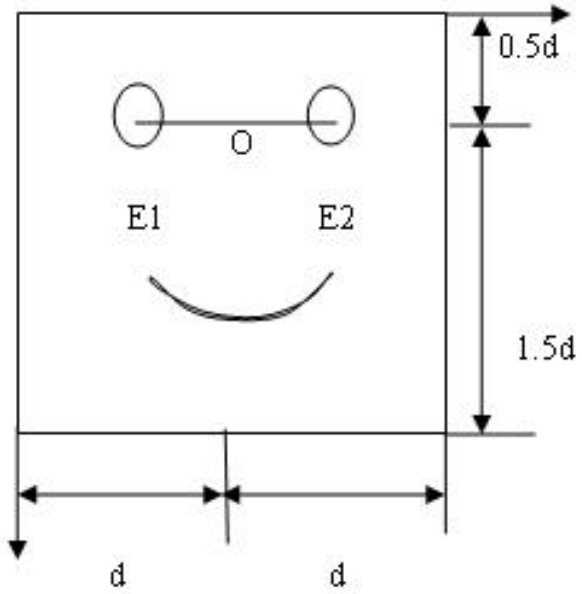
(1)图像旋转

经过图像旋转。使E1和E2的边线 E1E2保持水平，以保证人脸方向的一致性。根据是人脸在平面内的旋转不变性，在这之前需要做的工作是双眼定位，可通过人机交互方式用鼠标点击两只眼的位置实现或采用算法对人眼进行自动定位(如可采用“结合梯度积分投影和灰度积分投影的人眼自动定位方法”等，但人眼自动定位难度较为复杂)。



(2)图像裁减

根据图所示的比例关系，进行图像裁剪，图中O为E1E2的中点，且假定为E1E2=d。经过裁剪后，在2d×2d的图像内，保证O点固定于(d, 0.5d)处，这样就保证了人脸位置的一致性。根据是人脸在图像平面内的平移不变性。



在第(1)步旋转过程中，两眼位置确定之后，就能够求得O点的坐标；但旋转会改变O点的位置，因此先将图像的中心移到O点，然后进行旋转。这样，旋转之后，O点还是图像的中心。

(3)图像缩放

进行图像的缩放变换。得到统一大小的标准图像。统一规定的图像的大小是32×32像素点，即使d=E1E2为定长(16个像素点)，缩放倍数β=2d/32，这样就保证了人脸图像的大小一致性。根据是人脸在图像平面内的尺度不变性。

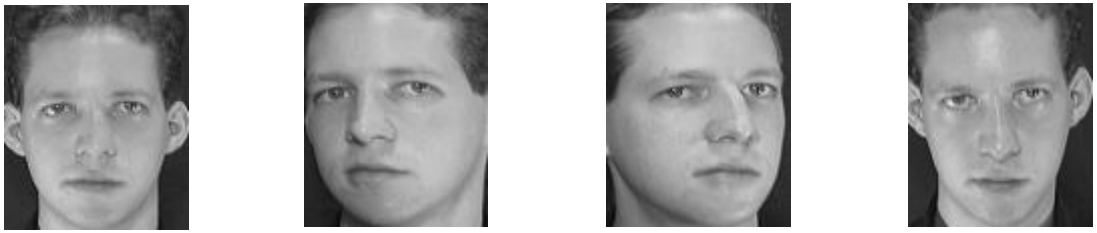


图 a) ORL人脸库中的4幅112×92原始图像

图b) 图a)的32×32几何归一化图像

使用不同散布矩阵(离散度矩阵)进行K-L变换[02]：

根据不同的散布矩阵进行K-L变换，对保留分类鉴别信息的效果不同。

(1) 采用多类类内散布矩阵 $\boldsymbol{S}_w$ 作 K-L 变换

多类类内散布矩阵：

$$\boldsymbol{S}_w = \sum_{i=1}^{m} P(\omega_i) E[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \mid \mathbf{x} \in \omega_i]$$

若要突出各类模式的主要特征分量的分类作用：

选用对应于大特征值的特征向量组成变换矩阵；

若要使同一类模式聚集于最小的特征空间范围：

选用对应于小特征值的特征向量组成变换矩阵。

这一方法也是一种有监督的特征变换 (降维) 方法。

(2) 采用类间散布矩阵 S_b 作 K-L 变换

类间散布矩阵：$$\boldsymbol{S}_b = \sum_{i=1}^{m} P(\omega_i)(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}}$$

适用于**类间距离比类内距离大得多的多类问题**，选择与大特征值对应的特征向量组成变换矩阵。

(3) 采用总体散布矩阵 $\boldsymbol{S}_t$ 作 K-L 变换

把多类模式合并起来看成一个总体分布。

总体散布矩阵：$$\boldsymbol{S}_t = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}] = \boldsymbol{S}_b + \boldsymbol{S}_w$$

适合于**多类模式在总体分布上具有良好的可分性的情况**。采用大特征值对应的特征向量组成变换矩阵，能够保留模式原有分布的主要结构。

利用K-L变换进行特征提取的优点：

(1) 在均方逼近误差最小的意义下使新样本集$\{X^*\}$逼近原样本集$\{X\}$的分布，**既压缩了维数、又保留了数据集的分布信息和类别鉴别信息**。(与样本集变异情况，也即协方差阵有关)

(2) 变换后的新模式向量各分量相对总体均值的方差等于原样本集总体自相关矩阵的大特征值，表明变换加强了模式类之间的差异性。

$$\boldsymbol{C}^* = E\{(\mathbf{x}^* - \boldsymbol{\mu}^*)(\mathbf{x}^* - \boldsymbol{\mu}^*)^{\mathrm{T}}\} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & 0 & \\ 0 & & & \lambda_d \end{bmatrix}$$

(3) $\boldsymbol{C}^*$为对角矩阵说明了**变换后样本各分量特征互不相关**，即消除了**原分量特征间**的相关性，便于进一步进行特征的选择。

K-L变换的不足：

(1) 对两类问题容易得到较满意的结果。**类别愈多，效果愈差**。

(2) 需要通过足够多的样本估计样本集的协方差矩阵或其它类型的散布矩阵。**当样本数不足时，矩阵的估计会变得十分粗略**，变换的优越性也就不能充分的地显示出来。

(3) 矩阵的特征值和特征向量缺乏统一的快速算法，计算较困难。

9.5 灰度共生矩阵及纹理特征提取 (*: 选学)

灰度共生矩阵(Gray Level Co-occurrence Matrix, GLCM)简称共生矩阵，它是统计空间上具有某种位置关系的一对像素灰度对出现的频度，d为灰度共生矩阵的生成步长，θ为生长方向。

灰度共生矩阵常用于图像分析中的纹理特征提取和图像检索中。

The co-occurrence matrix is commonly used for texture analysis . Each matrix element $M_{\theta,d}(i, j)$ represents the joint probability that two neighboring image pixels in the direction θ at a distance d have gray values equal to i and j. Four directions are usually considered in texture analysis, $\theta =$ 0, 45, 90, 135 degrees. The distance d is usually set to 1. Hence, four matrices are built. The width and height of the matrices are equal to the number of possible gray levels. Usually, natural scenes do not have a preferred texture orientation. For this reason, the contributions of the matrices in different directions can be summed, producing

$$Co_d(i, j) = \sum_{\theta \in S} M_{d,\theta}(i, j)$$

Where $S = \{0, 45, 90, 135\}$. The matrix $Co_d(i, j)$ exploits the local distribution of intensity values of the input image. An example of a co-occurrence matrix is shown in Figure 2. In this example, the left is an image of 3 gray levels, and the right is its co-occurrence matrix of the distance $d = 1$ and $\theta = 135$ degree.

取图像(N×N)中任意一点(x, y)及偏离它的另一点(x+a, y+b)，设该点对的灰度值为(g1, g2)。令点(x, y)在整个画面上移动，则会得到各种(g1, g2)值，设灰度值级数为k，则(g1, g2)的组合共有k的平方种。对于整个画面，统计出每一种(g1, g2)值出现的次数，然后排列成一个方阵，再用(g1, g2)出现的总次数将它们归一化为出现的概率P(g1, g2)，这样的方阵称为灰度共生矩阵。距离差分值(a, b)取不同的数值组合，可以得到不同情况下的联合概率矩阵。(a, b)取值要根据纹理周期分布的特性来选择，对于较细的纹理通常选取(1, 0)、(1, 1)、(2, 0)等小的差分值。

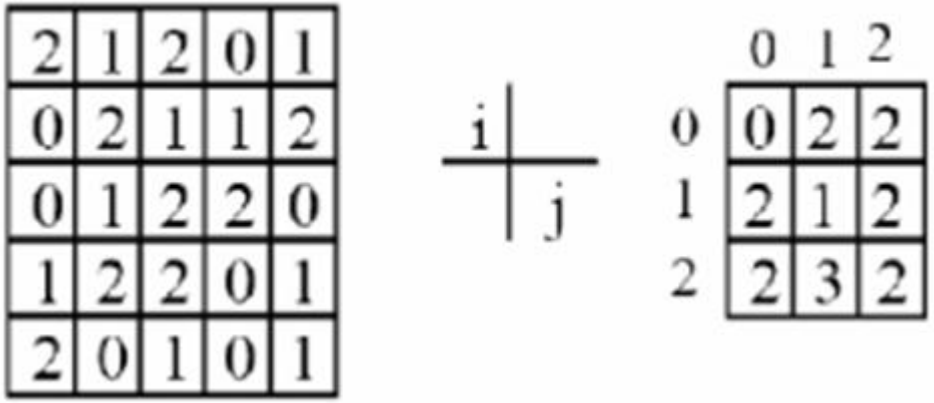


图 灰度共生矩阵示例 (d=1,θ=135°)

由于灰度共生矩阵的维度较大，一般不直接作为区分纹理的特征，而是基于它构建的一些统计量作为纹理分类特征。

Haralick[04]提出的灰度共生矩阵共有14个特征参数，在实际应用中可根据其各自的意义和实验效果选用，常用的有角二阶矩/能量 (Angular Second Moment/Energy，ASM)、逆差矩(Inverse Difference Moment/Homogeneity)、对比度(Contrast)、熵(Entropy)等。

1. 角二阶矩

$$energy = \sum_{i,j} Co_d(i,j)^2$$

能量大则纹理粗糙；能量小则纹理细致。

2. 逆差矩

$$inverse\_difference\_moment = \sum_{i,j} \frac{Co_d(i,j)}{1+(i-j)^2}$$

用于度量图像纹理局部变化的大小。纹理越规则，逆差距越大，反之亦然。

3. 对比度

$$contrast = \sum_{i,j} (i-j)^2 Co_d(i,j)$$

表征纹理的清晰程度；相邻像素对的灰度差别越大，对比度就越大，图像就越清晰。

4. 熵

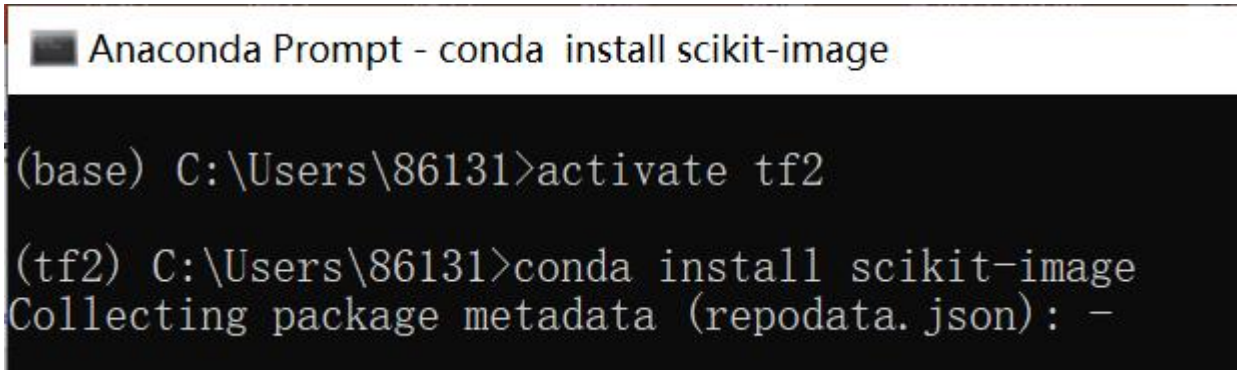
$$entropy = -\sum_{i,j} Co_d(i,j) \log_2 Co_d(i,j)$$

代表图像的信息量，表示纹理的复杂程度，是图像内容随机性度量。无纹理熵为0，纹理越复杂熵越大。

灰度共生矩阵分析法在图像纹理特征提取中有着广泛的应用。文献[05-06]是另外两篇灰度共生矩阵应用的相关论文，感兴趣者可以参考阅读。

示例：使用scikit-image库进行灰度共生矩阵及其纹理特征计算。

安装第三方scikit-image库: conda install scikit-image。



现对前面的Figure2，使用Python调用skimage计算灰度共生矩阵，在此基础上，可进一步计算GLCM的纹理特征。

Scikit-image库提供了两个模块，其中**(1) skimage.feature.greycomatrix(image, …[, …])**用于计算灰度共生矩阵(GLCM)，**(2) skimage.feature.greycoprops(P[, prop])**用于计算GLCM的纹理特征。

In [1]:
```
#Filename: glcm.ipynb
from skimage.feature import greycomatrix, greycoprops
import numpy as np
image = np.array([[2,1,2,0,1],
                  [0,2,1,1,2],
                  [0,1,2,2,0],
                  [1,2,2,0,1],
                  [2,0,1,0,1]], dtype=np.uint8)
glcm = greycomatrix(image,[1],[0, np.pi/4,np.pi/2,np.pi*3/4],levels=3)#调用GLCM函数
```

In [2]:
```
print(glcm.shape)
print(glcm)
```

```
(3, 3, 1, 4)
[[[[0 0 2 1]]

  [[5 2 3 2]]

  [[1 2 0 0]]]


 [[[1 2 0 2]]

  [[1 1 1 3]]

  [[4 2 6 1]]]


 [[[4 2 4 2]]

  [[2 3 3 0]]

  [[2 2 1 5]]]]
```

```
In [3]: glcm[:,:,0,0] #距离d为1时，扫描角度为0度时的GLCM

Out[3]: array([[0, 5, 1],
               [1, 1, 4],
               [4, 2, 2]], dtype=uint32)

In [4]: glcm[:,:,0,1] #距离d为1时，扫描角度为45度时的GLCM

Out[4]: array([[0, 2, 2],
               [2, 1, 2],
               [2, 3, 2]], dtype=uint32)

In [5]: glcm[:,:,0,2] #距离d为1时，扫描角度为90度时的GLCM

Out[5]: array([[2, 3, 0],
               [0, 1, 6],
               [4, 3, 1]], dtype=uint32)

In [6]: glcm[:,:,0,3] #距离d为1时，扫描角度为135度时的GLCM

Out[6]: array([[1, 2, 0],
               [2, 3, 1],
               [2, 0, 5]], dtype=uint32)
```

基于以上GLCM计算结果，可用**skimage.feature.greycoprops**实现GLCM对比度、ASM能量、逆差距等特征的计算 (此略)。

9.6 快速PCA (*: 选学)

PCA的计算中最主要的工作量是计算样本协方差矩阵的特征值和特征向量。设样本矩阵X大小为n×d(n个d维样本特征向量)，则样本协方差矩阵S (离散度矩阵/散布矩阵scatter matrix是样本协方差矩阵的n-1倍) 将是一个d ×d的方阵，故当维数d较大时计算复杂度会极高。如，当维数d=10000，S是一个10000×10000维的矩阵，此时如果采用前面的scikit-learn PCA函数或MATLAB princomp函数计算主成分，Python/MATLAB通常会出现内存耗尽的错误，即使有足够的内存，要得到S的全部特征值也可能要花费数小时的时间。

快速PCA的基础理论：

当样本散布矩阵S的维数d较大时，计算复杂度会极高。有一个非常好的技巧可以用来计算矩阵S非零特征值所对应的特征向量。

设$Z_{n\times d}$为样本矩阵X中的每个样本减去样本均值$\boldsymbol{\mu}$后得到的矩阵，则散布矩阵S为$(Z^TZ)_{d\times d}$. 现在考虑矩阵$R=(ZZ^T)_{n\times n}$，很多情况下，由于样本数目n远远小于样本维数d(如人脸图像)，R 的大小也远远小于散布矩阵S，然而， R与S有着相同的非零特征值.

设n维列向量$\vec{v}$是R的特征向量，则有：

$$(ZZ^T)\vec{v}=\lambda\vec{v} \qquad (5.6.1)$$

将(5.6.1)式两边同时左乘Z^T，并应用矩阵乘法的结合律得：

$$(Z^TZ)(Z^T\vec{v})=\lambda(Z^T\vec{v}) \qquad (5.6.2)$$

$$(Z^T Z)(Z^T \vec{v}) = \lambda (Z^T \vec{v}) \qquad (5.6.2)$$

式(5.6.2)说明$(Z^T \vec{v})$为散布矩阵$S=(Z^T Z)_{d\times d}$的特征向量. 这说明可以先计算小矩阵$R=(ZZ^T)_{n\times n}$的特征向量$\vec{v}$, 然后通过左乘Z^T得到散布矩阵$S=(Z^T Z)_{d\times d}$的特征向量$Z^T \vec{v}$.

本章小结：

1. 特征提取与特征选择的基本概念

特征选择是通过映射 (变换) 的方法把高维的特征向量转为低维的特征向量。特征选择是从原始特征中挑选出一些最有代表性、分类性能好的特征，以达到降低特征空间维数的目的。

2. 类别可分性判据 (准则)

基于距离的可分性判据：各类样本之间的距离越大、类内离散度越小，则类别的可分性越好。

基于概率密度函数的可分性判据、基于熵函数的可分性判据 (*：选学)。

3. K-L变换/主成分分析 (PCA)

K-L变换/主成分分析 (PCA)/主分量分析是一种常用的无监督线性降维特征提取方法。K-L变换适用于任何概率分布，它是以最小均方误差为准则进行数据压缩，是在最小均方误差意义下的最优正交解。

4. 灰度共生矩阵(*: 选学)

灰度共生矩阵是通过研究灰度的空间相关特性来描述纹理特征的一种方法。由于纹理是由灰度分布在空间位置上反复出现而形成的，因而在图像空间中相隔某距离的两像素之间会存在一定的灰度关系，即图像中灰度的空间相关特性。

课外作业题9 (特征提取与选择)：

1. 简述特征选择和特征提取的异同。

2. 简述基于距离可分性判据的特征提取方法。

3. (1)为什么在机器学习与模式识别应用中，需要尽可能进行特征提取？(2) 能用特征提取方法进行有效的数据降维，需要数据有什么样的特点？

4. 已知一组数据的协方差矩阵为 $\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$

试问：

(1) 协方差矩阵中各元素的含义是什么？

(2) K-L变换的最佳准则是什么？

(3) 为什么说经K-L变换后消除了各分量之间的相关性？

5. 假定 ω_i 类的样本集为$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4\}$，它们分别为

$$\boldsymbol{X}_1 = [2, 2]^{\mathrm{T}},\ \boldsymbol{X}_2 = [3, 2]^{\mathrm{T}},\ \boldsymbol{X}_3 = [3, 3]^{\mathrm{T}},\ \boldsymbol{X}_4 = [4, 2]^{\mathrm{T}}$$

(1) 求类内散布矩阵(类内离散度矩阵)；

(2) 求类内散布矩阵的特征值和对应的特征向量；

(3) 求变换矩阵*A*，将二维模式变换为一维模式。

6. 试编写程序(Python/R/MATLAB等)计算以下矩阵的4方向(d=1; θ=0°, 45°, 90°, 135°)灰度共生矩阵(*：选做)。

2	1	2	0	1
0	2	1	1	2
0	1	2	2	0
1	2	2	0	1
2	0	1	0	1

5’. 若有下列两类样本集：

ω1: $x1=(0,0,0)^T$, $x2=(1,0,0)^T$,$x3=(1,0,1)^T$,$x4=(1,1,0)^T$

ω2: $x5=(0,0,1)^T$,$x6=(0,1,0)^T$,$x7=(0,1,1)^T$,$x8=(1,1,1)^T$

要求用K-L变换法/PCA方法，分别把特征空间维数降到d=2和d=1。试编写满足要求的程序(Python/R/MATLAB等语言)。

6. 试编写程序(MATLAB/Python/R等)计算以下矩阵的4方向(d=1; θ=0°, 45°, 90°, 135°)灰度共生矩阵(*：选做)。

2	1	2	0	1
0	2	1	1	2
0	1	2	2	0
1	2	2	0	1
2	0	1	0	1

课外作业9参考答案(特征提取与选择)：

1.特征选择是指,从L个度量值$(x_1, x_2, ..., x_L)^T$中按一定的准则选择出供分类用的子集,作为降维(m维,$m < L$)的分类特征。

特征提取是指, 使一组度量值$(x_1, x_2, ..., x_L)^T$通过某种变换$T_i(.)$产生新的m个特征$(y_1, y_2, ..., y_m)^T$作为降维的分类特征,这里$i = 1, 2, ..., m; m < L.$

注意，特征选择是“挑选”出较少的特征用于分类，特征提取是通过“数学变换”产生较少的特征。它们都是为了在尽可能保留识别信息的前提下，降低特征空间的维数，以实现有效的分类。

特征选择和特征提取有时并不是截然分开的，如，可以先进行特征选择，从原始测量数据中去掉那些明显没有分类信息的特征，然后再进行特征提取，进一步降低维数。

2.假设有n个原始特征：$\mathbf{x}=[x_1,x_2,...,x_n]^T$，希望通过线性映射压缩为$d$个特征$\mathbf{y}=[y_1,y_2,...,y_d]^T$，其变换关系为$\mathbf{y}=W^T\mathbf{x}$，$W$为$n\times d$矩阵。

令S_w，S_b为原空间(即X的)类内/类间离散度矩阵，S_w^*，S_b^*为映射后(即Y的)离散度矩阵：$S_b^*=W^T S_b W$，$S_w^*=W^T S_w W$，经变换后的J_2变为

$$J_2(W)=tr[(S_w^*)^{-1}S_b^*]=tr[(W^T S_w W)^{-1}(W^T S_b W)]$$

将上式对W的各个分量求偏导数并令其为零即可以确定一个W值。

3.答:

(1)减少特征数量，有利于降低模型复杂度；对特征之间存在共线性的数据，通过有效的特征提取，能降低模型求解的困难 (改善求解的病态条件)。

(2)特征之间存在较大相关性 (多重共线性)，是特征提取得以有效进行的基础。

4. 答：已知协方差矩阵 $\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$ 则：

(1) 其对角元素是各分量的方差，非对角元素是各分量之间的协方差。

(2) K-L变换的最佳准则为：对一组数据按一组正交基进行分解，在只取相同数量分量的条件下，以均方误差计算截尾误差最小。

(3) 在经过K-L变换后，协方差矩阵成为对角矩阵，因而各分量间的相关被消除。

5.解:

（1） $$\boldsymbol{M}=\frac{1}{4}\sum_{i=1}^{4}\boldsymbol{X}_i=\frac{1}{4}\left\{\begin{bmatrix}2\\2\end{bmatrix}+\begin{bmatrix}3\\2\end{bmatrix}+\begin{bmatrix}3\\3\end{bmatrix}+\begin{bmatrix}4\\2\end{bmatrix}\right\}=\begin{bmatrix}3\\9/4\end{bmatrix}$$

类内散布矩阵：

$$\boldsymbol{C}=\frac{1}{4}\sum_{i=1}^{4}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}-\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}$$

$$=\frac{1}{4}\left\{\begin{bmatrix}2\\2\end{bmatrix}[2,\ \ 2]+\begin{bmatrix}3\\2\end{bmatrix}[3,\ \ 2]+\begin{bmatrix}3\\3\end{bmatrix}[3,\ \ 3]+\begin{bmatrix}4\\2\end{bmatrix}[4,\ \ 2]\right\}-\begin{bmatrix}3\\9/4\end{bmatrix}[3,\ \ 9/4]$$

$$=\frac{1}{4}\begin{bmatrix}38 & 27\\27 & 21\end{bmatrix}-\begin{bmatrix}9 & 27/4\\27/4 & 81/16\end{bmatrix}=\begin{bmatrix}1/2 & 0\\0 & 3/16\end{bmatrix}$$

（2）① 由$|\lambda \boldsymbol{I}-\boldsymbol{C}|=0$求特征值。

$$\begin{vmatrix} \lambda-1/2 & 0 \\ 0 & \lambda-3/16 \end{vmatrix}=0$$

$$(\lambda-1/2)(\lambda-3/16)=0$$

$$\lambda_1=1/2,\quad \lambda_2=3/16$$

② 由$(\lambda \boldsymbol{I}-\boldsymbol{C})\boldsymbol{u}_1=0$解得$\lambda_1$对应的特征向量$\boldsymbol{u}_1$为$\boldsymbol{u}_1=[1,\ \ 0\ ]^{\mathrm{T}}$。

由$(\lambda \boldsymbol{I}-\boldsymbol{C})\boldsymbol{u}_2=0$解得$\lambda_2$对应的特征向量$\boldsymbol{u}_2$为$\boldsymbol{u}_2=[0,\ \ 1]^{\mathrm{T}}$。

（3）① 选择较小特征值 $\lambda_2 = 3/16$ 对应的特征向量 $\boldsymbol{u}_2 = [0, \quad 1]^{\mathrm{T}}$ 构成变换矩阵。

$\boldsymbol{u}_2$ 已为归一化特征向量，直接构成变换矩阵：

$$\boldsymbol{A} = [\boldsymbol{u}_2^{\mathrm{T}}] = [0, \quad 1]$$

② 变换：

$$\boldsymbol{X}_1^* = \boldsymbol{A}\boldsymbol{X}_1 = [0, \quad 1]\begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2$$

$$\boldsymbol{X}_2^* = \boldsymbol{A}\boldsymbol{X}_2 = [0, \quad 1]\begin{bmatrix} 3 \\ 2 \end{bmatrix} = 2$$

$$\boldsymbol{X}_3^* = \boldsymbol{A}\boldsymbol{X}_3 = [0, \quad 1]\begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3$$

$$\boldsymbol{X}_3^* = \boldsymbol{A}\boldsymbol{X}_3 = [0, \quad 1]\begin{bmatrix} 4 \\ 2 \end{bmatrix} = 2$$

6.灰度共生矩阵计算MATLAB参考程序:

(1) 程序清单

```
%%Filename:Co_Matrixa.m
%灰度共生矩阵程序：在MATLAB中，三维数组(3阶张量)是按(row,column,pages)方式寻址
%程序Example三维数组按(pages,row,column)寻址
%待处理图像矩阵的行数m，列数n；
%这里以[2 1 2 0 1;0 2 1 1 2;0 1 2 2 0;1 2 2 0 1;2 0 1 0 1]为例,共生矩阵的大小为3*3,
%程序验证正确;
clear all;clc;
Example=[2 1 2 0 1;0 2 1 1 2;0 1 2 2 0;1 2 2 0 1;2 0 1 0 1];
m=5;
n=5;
Matrix_Co=zeros(3,3,4);
%分别计算0，45，90，135度四个方向的共生矩阵
%Matrix_Co(1,:,:)为0度的同现矩阵，Matrix_Co(2,:,:)为90度的共生矩阵；
%Matrix_Co(3,:,:)为45度的同现矩阵，Matrix_Co(4,:,:)为135度的共生矩阵；
for r=1:4
   for i=1:m
      for j=1:n
         %0度的共生矩阵
         if r==1
           if j<n
              p=Example(i,j)+1;
              q=Example(i,j+1)+1;
              Matrix_Co(p,q,r)=Matrix_Co(p,q,r)+1;
           end
```

```
        %90度的共生矩阵
        elseif r==2
          if i<m
            p=Example(i,j)+1;
            q=Example(i+1,j)+1;
            Matrix_Co(p,q,r)=Matrix_Co(p,q,r)+1;
          end
        %45度的共生矩阵
        elseif r==3
          if i<m&&j<n
            p=Example(i,j)+1;
            q=Example(i+1,j+1)+1;
            Matrix_Co(p,q,r)=Matrix_Co(p,q,r)+1;
          end
        %135度的共生矩阵
        elseif r==4
          if i<m&&j>1
            p=Example(i,j)+1;
            q=Example(i+1,j-1)+1;
            Matrix_Co(p,q,r)=Matrix_Co(p,q,r)+1;
          end

        end
      end
    end
end
Matrix_Co    %显示
```

(2)程序运行结果

```
Matrix_Co(:,:,1) =
    0    5    1
    1    1    4
    4    2    2
Matrix_Co(:,:,2) =
    2    3    0
    0    1    6
    4    3    1
Matrix_Co(:,:,3) =
    0    2    2
    2    1    2
    2    3    2
Matrix_Co(:,:,4) =
    1    2    0
    2    3    1
    2    0    5
```

本章主要参考文献：

[01] 李弼程等. **模式识别原理与应用**: 第5章，西电版，2008

[02] 齐敏等. **模式识别导论**: 第5章，清华版，2009

[03] 王耀明. **图像的矩函数-原理、算法及应用**. 华东理工大学出版社, 2002年.

[04] Haralick. Statistical and structural Approaches to Texture. Proceeding of IEEE, 1979, 67 (5): 786-804.

[05] 白雪冰等. 基于灰度共生矩阵的木材纹理分类方法的研究. 哈尔滨工业大学学报，2005年第37卷第12期, Pages: 1667-1670.

[06] 童隆正等. 肝纤维化图像的灰度共生矩阵分析. 首都医科大学学报. 2003年第24卷第3期, Pages: 240-242.

End of this lecture.

Thanks!