**特征提取**

1. 引言
2. 识别分类特征的条件
3. **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**

**补充：**

①特征可获取。观察对象可以通过数据采集设备输入到计算机。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。

②类内聚集。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样本构成的，因此它们同属一类，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类间离散。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**

1. 特征提取和特征选择（在保证性能的前提下减少特征数量、压缩特征空间）（构造模式识别系统最困难的任务：没有统一的有效方法）

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

作用：（1）简化计算：维数与计算机资源 （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

原始特征：直接测量得到的特征

特征提取：通过映射的方法把高维特征向量转化为低维特征向量，得到有效分类的特征（二次特征：原始特征的某种组合，最常用线性组合）。同时实现特征维数的降低，当模式在空间内移动旋转缩放，特征值应保持不变，保证有相同的识别效果。

特征选择：在原始特征中选择有代表性，分类性能好，对分类最有利的特征以降低维数

**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。

1. 基本概念
2. 特征的类别

①物理特征（观测型特征）：直接，人们容易感知，未必能有效表征分类对象

②结构特征。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**。**结构特征的表达能力一般要高于物理特征。**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图形信息。**结构信息对对象的尺寸往往不太敏感**。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

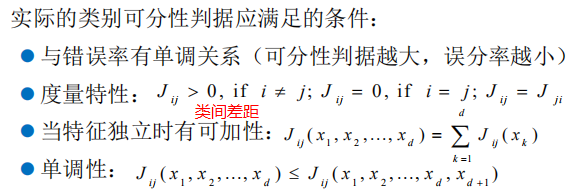
③数字特征。数字特征是为了表征观察对象而设立的特征，这种特征是抽象的，不容易被人感知。**数字特征有时和观察对象的固有特性没有任何联系，有时则是物理或结构特征的变换结果**

2.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**类内聚集**、**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**

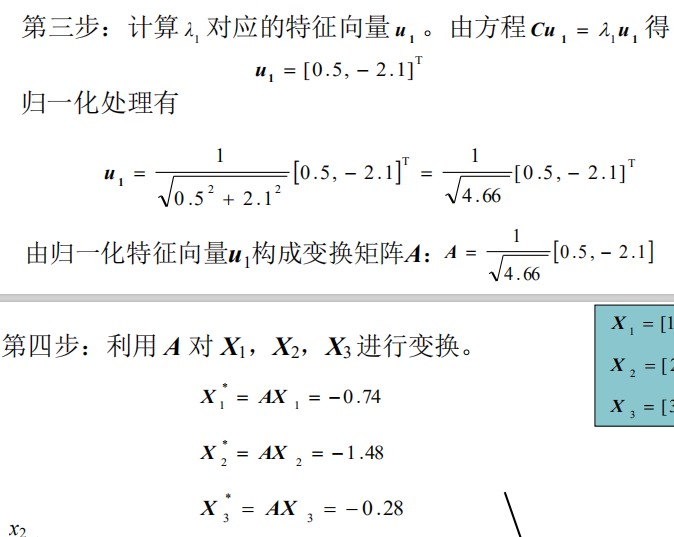
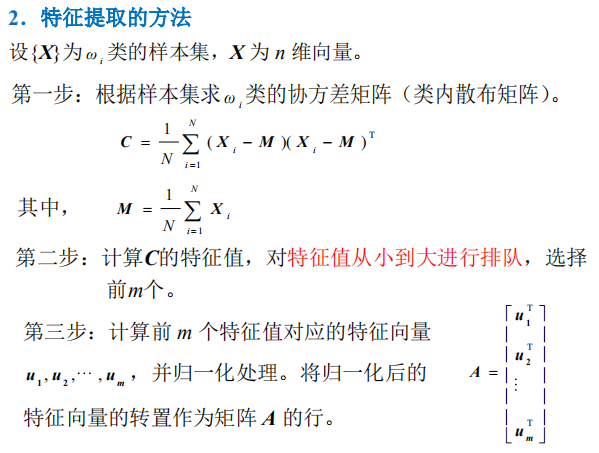
1. 类别可分性依据（特征评判标准）：衡量不同特征及其组合对分类是否有效的定量准则
2. 概念

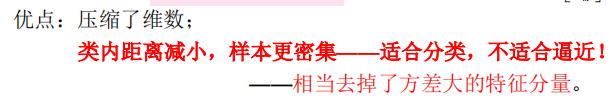
理想准则：某组特征使分类器错误概率最小

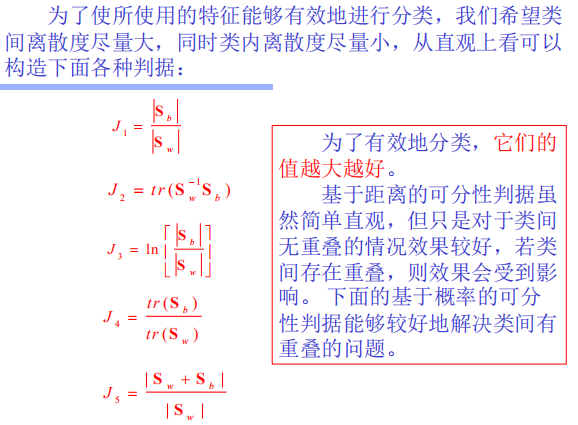
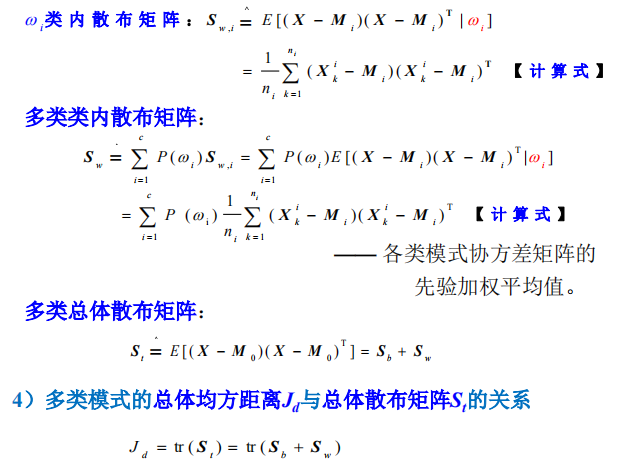
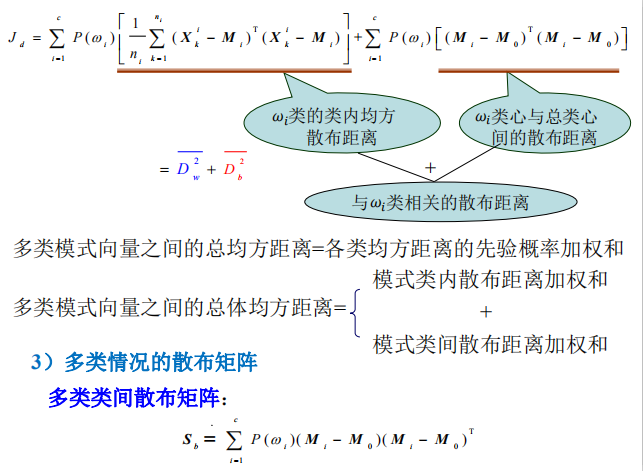
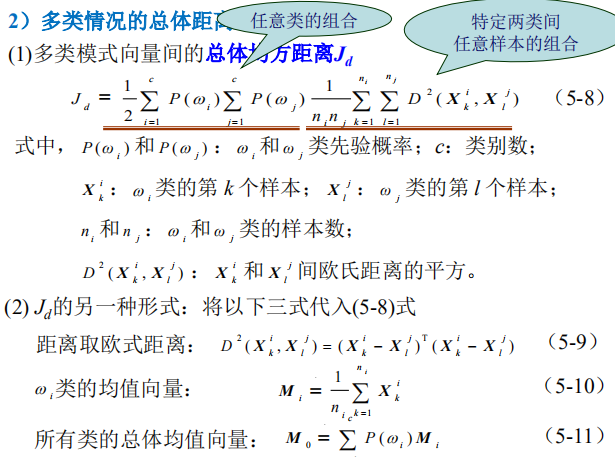


1. 类别可分性依据（p229）

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**。常用的点间距离有：欧氏距离、马氏距离、绝对距离(城市距离、 Hamming距离)、 Minkowsky距离等。

1. 多类模式向量之间的总体距离和总体散布矩阵（附录ppt）
2. LDA（类内散布矩阵）——有监督降维

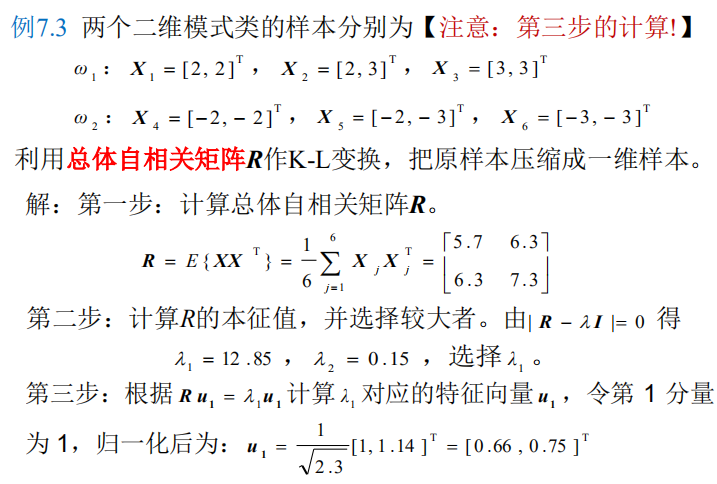


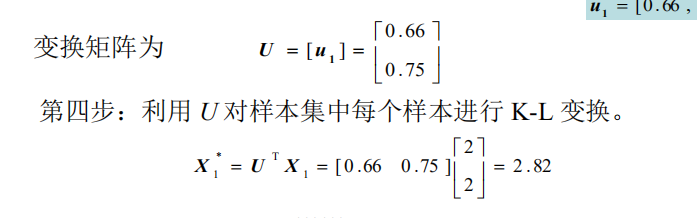


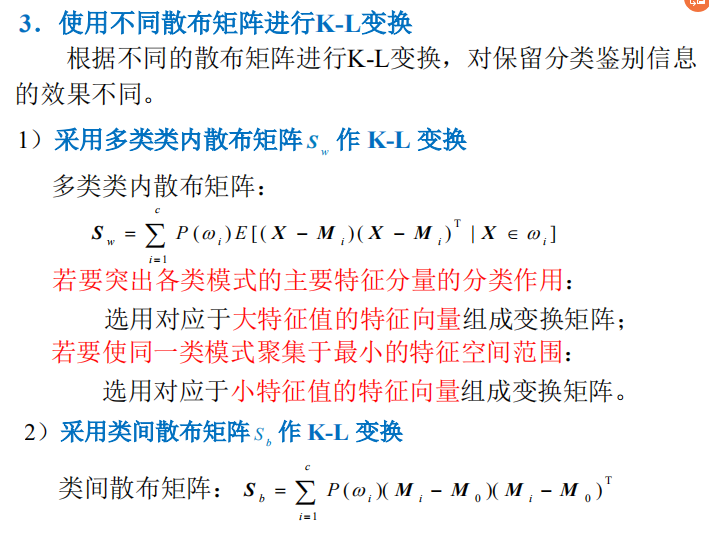
**PCA（离散KL变换）**

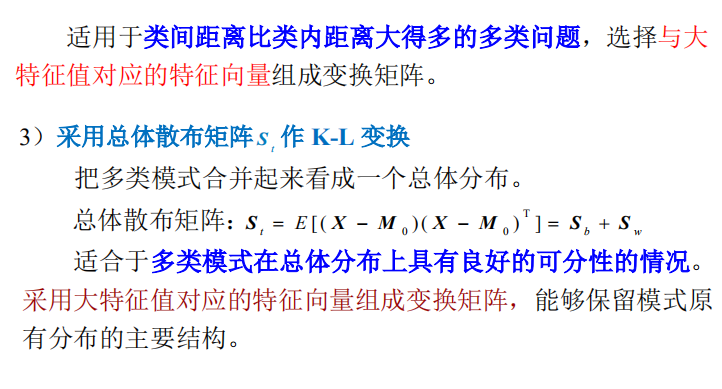
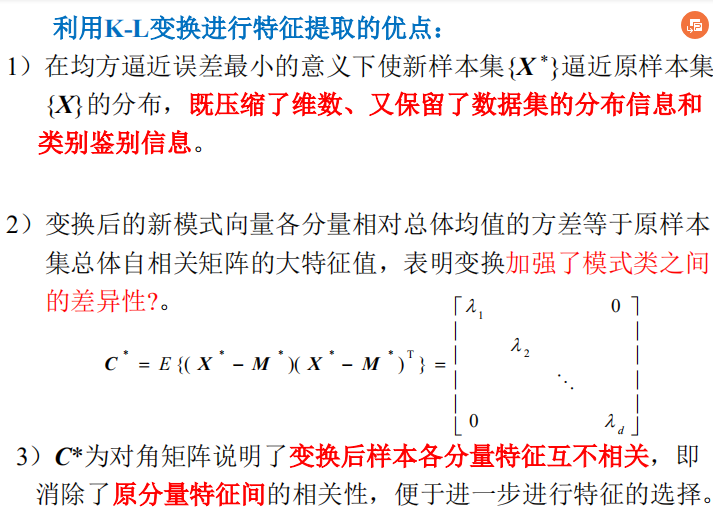
***1.K*-*L*变换**是一种基于目标统计特性的最佳正交变换，**它的最佳性体现在变换后产生的新 的分量正交或不相关**。

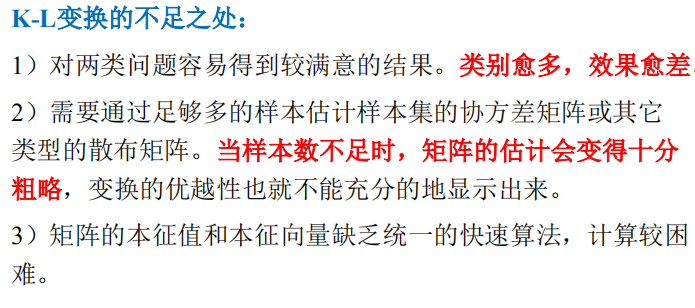
变换后与之前的误差最小









加回变换的均值