题型：

1.填空题 5 题
2.名词解释 4 题
3.问答题 4 题
4.计算作图题 3 题
5.综合计算题 1 题

填空题

备注 1：没有整理第一章和第六章，老师说不考的
备注 2：非线性判别函数相关概念 P69
概率相关定义、性质、公式 P83 以后
最小错误率贝叶斯决策公式 P85
最小风险贝叶斯 P86
正态贝叶斯 P90
综合计算有可能是第六次作业

一、填空题

物以类聚 人以群分体现的是聚类分析的基本思想。

模式识别分类： **1.**从实现方法来分模式识别分为监督分类和非监督分类； **2.**从理论上来分，有统计模式识别，统计模式识别，模糊模式识别，神经网络模式识别法

聚类分析 是按照不同对象之间的差异，根据距离函数的规律做模式分类的。

模式的特性 ：可观察性、可区分性、相似性

模式识别的任务 ：一是研究生物体（包括人）是如何感知对象的，二是如何用计算机实现模式识别的理论和方法。

计算机的发展方向： 1.神经网络计算机－－模拟人的大脑思维； 2.生物计算机－－运用生物工程技术、蛋白分子作芯片；3.光计算机－－用光作为信息载体，通过对光的处理来完成对信息的处理。

训练学习方法： 监督学习、无监督学习（无先验知识，甚至类别数也未知）。

统计模式识别有： 1.聚类分析法（ 非监督）； 2.判决函数法 /几何分类法 (监督)；3.基于统计决策的概率分类法 - 以模式集在特征空间中分布的类概率密度函数为基础，对总体特征进行研究，以取得分类的方法

数据的标准化目的： 消除各个分量之间数值范围大小对算法的影响

模式识别系统的基本构成： 书 P7

聚类过程遵循的基本步骤： 特征选择；近邻测度；聚类准则；聚类算法；结果验证；结果判定。

相似测度基础： 以两矢量的方向是否相近作为考虑的基础 ,矢量长度并不重要。

确定聚类准则的两种方式： 阈值准则，函数准则

基于距离阈值的聚类算法 ——分解聚类：近邻聚类法；最大最小距离聚类法

类间距离计算准则 **:**1）最短距离法 2）最长距离法 3）中间距离法 4）重心法 5）类平均距离法 6）离差平方和法 P24

系统聚类法 ——合并的思想

用于随机模式分类识别的方法 ，通常称为贝叶斯判决。

BAYES 决策常用的准则： 最小错误率；最小风险

错误率的计算或估计方法： ①按理论公式计算；②计算错误率上界；③实验估计。

名词解释

1.名词解释

相似性测度： 衡量模式之间相似性的一种尺度

明氏距离： P17 当 **m=2** 时，明氏距离为欧氏距离。当 **m=1** 时：绝对距离（ 曼哈顿距离 ）称为“ 街坊 ”距离

感知器算法： 就是通过训练样本模式的迭代和学习，产生线性（或广义线性）可分的模式判别函数。

梯度： P59

感知器 **P227**

模糊度 **P182**

清晰性 **P182**

含混性

近似性

随机性

》》》》》》》

模式： 对客体（研究对象）特征的描述（定量的或结构的），是取自客观世界的某一样本的测量值的集合（或综合）。

模式所指的不是事物本身，而是从事物获得的信息。

模式识别： 确定一个样本的类别属性（模式类）的过程，即把某一样本归属于多个类型中的某个类型。

模式类： 具有某些共同特性的模式的集合。

特征选择 ：在原始特征基础上选择一些主要特征作为判别用的特征。

特征提取： 采用某种变换技术，得出数目上比原来少的综合特征作为分类用。

特征抽取： 通过各种手段从原始数据中得出反映分类问题的若干特征（有时需进行数据标准化）

特征空间： 进行模式分类的空间。

特征向量： 用 n 维列向量来表示 一个（模式）样本，说明该样本具有 n 个数字特征

x= (x1 , x2 ， ， , xn)T

常称之为特征向量。

人工智能： 是研究如何将人的智能转化为机器智能，或者是用机器来模拟或实现人的智能。

聚类分析： 根据模式之间的相似性（相邻性）对模式进行分类，是一种非监督分类方法。

聚类准则： 根据相似性测度确定的，衡量模式之间是否相似的标准。即把不同模式聚为一类还是归为不同类的准则——同一类模式相似程度的标准或不同类模式差异程度的标准。

聚类准则函数： $J=\sum_{j=1}^{c}\sum_{X\in S_j}\left|X-M_j\right|^2$

聚类准则函数： 在聚类分析中，表示模式类内相似或类间差异性的函数。

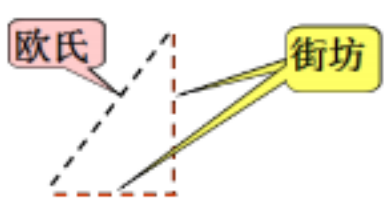
相似度： 衡量模式之间相似程度的尺度。

相似性测度： 衡量模式之间相似性的一种尺度。

欧氏距离 （简称距离）： P15

马氏距离： P15

明氏距离： P17 当 **m=2** 时，明氏距离为欧氏距离。当 **m=1** 时：绝对距离（ 曼哈顿距离 ）称为“ 街坊 ”距离



汉明 **(Hamming)** 距离 **:**P17

判别函数： 直接用来对模式进行分类的准则函数。

感知器算法： 就是通过训练样本模式的迭代和学习，产生线性（或广义线性）可分的模式判别函数。

梯度： P59

分类器的正确率： 指分类器正确分类的项目占所有被分类项目的比率。

过拟合： 高维空间训练形成的分类器，相当于在低维空间的一个复杂的非线性分类器，这种分类器过多的强调了训练集

的准确率甚至于对一些错误 /异常的数据也进行了学习， 而正确的数据却无法覆盖整个特征空间。 为此， 这样得到的分类器在对新数据进行预测时将会出现错误。这种现象称之为过拟合，同时也是维数灾难的直接体现。

问答题

2.问答题

统计模式识别的优缺点：

主要优点：

1）比较成熟

2）能考虑干扰噪声等影响

3）识别模式基元能力强

主要缺点：

1）对结构复杂的模式抽取特征困难

2）不能反映模式的结构特征，难以描述模式的性质

3）难以从整体角度考虑识别问题

句法模式识别优缺点：

主要优点：

1）识别方便，可以从简单的基元开始，由简至繁。

2）能反映模式的结构特征，能描述模式的性质。

3）对图象畸变的抗干扰能力较强。

主要缺点：

当存在干扰及噪声时，抽取特征基元困难，且易失误。

模糊模式识别优缺点：

主要优点：

由于隶属度函数作为样本与模板间相似程度的度量，故往往能反映整体的与主体的特征，从而允许样本有相当程度的干扰与畸变。

主要缺点：

准确合理的隶属度函数往往难以建立，故限制了它的应用。

神经网络模式识别法优缺点：

主要优点：

可处理一些环境信息十分复杂，背景知识不清楚，推理规则不明确的问题。允许样本有较大的缺损、畸变。

主要缺点：

模型在不断丰富与完善中，目前能识别的模式类还不够多。

分类与聚类的区别：

分类：用已知类别的样本训练集来设计分类器（监督学习），由学习过程和识别过程两部分组成，且用于学习的样本类别是已知的。

聚类（集群）：事先不知样本的类别，而利用样本的先验知识来构造分类器（无监督学习）。

马氏距离的优缺点：

优点：

它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关；

由标准化数据和中心化数据（即原始数据与均值之差）计算出的二点之间的马氏距离相同；

马氏距离还可以排除变量之间的相关性的干扰；

满足距离的四个基本公理：非负性、自反性、对称性和三角不等式。

缺点：

有可能夸大变化微小的变量的作用；

协方差不易计算

近邻聚类法优缺点：

优点：

计算简单（一种虽粗糙但快速的方法）。

局限性：

聚类过程中，类的中心一旦确定将不会改变，模式一旦指定类后也不再改变。

聚类结果很大程度上依赖于第一个聚类中心的位置选择、 待分类模式样本的排列次序、 距离阈值 T 的大小以及样本分布的几何性质等。

最大最小距离算法（小中取大距离算法 ） ：

算法思想：

在模式特征矢量集中以最大距离原则选取新的聚类中心。以最小距离原则进行模式归类，通常使用欧式距离。

层次聚类法（系统聚类法、分级聚类法） ：

思路：

每个样本先自成一类，然后按距离准则逐步合并，减少类数。

动态聚类的基本步骤：

建立初始聚类中心，进行初始聚类；

计算模式和类的距离，调整模式的类别；

计算各聚类的参数，删除、合并或分裂一些聚类；

从初始聚类开始， 运用迭代算法动态地改变模式的类别和聚类的中心使准则函数取得极值或设定的参数达到设计要求时停止。

ISODATA 与 **K-** 均值算法比较：

相似：聚类中心的位置均通过样本均值的迭代运算决定。

相异： K- 均值算法的聚类中心个数不变；

ISODATA 的聚类中心个数变化。

ISODATA 基本思路：

（1）选择初始值——包括若干聚类中心及一些指标。可在迭代运算过程中人为修改，据此将 N 个模式样本分配到各个聚类中心去。

（2）按最近邻规则进行分类。

（3）聚类后的处理：计算各类中的距离函数等指标，按照给定的要求，将前次获得的聚类集进行分裂或合并处理，以获得新的聚类中心，即调整聚类中心的个数。

（4）判断结果是否符合要求：

符合，结束；

否则，回到（ 2）。

不同聚类算法比较：

算法		基本思想	聚类中心个数	样本归类	聚类结果对初始中心选择	类中心	类间距离	其他特点
分解聚类	近邻	分裂	单调变化，阈值确定	不变	敏感	不变	否	模式样本的几何分布性质影响均存在！ 排列次序或读入次序的影响不可忽视。
	最大最小距离			不变	不敏感	不变	否	
系统聚类	层级聚类	合并	同上	变化	不敏感	变化	需要	
动态聚类	K-均值	兼顾	指定，不变	变化	敏感	变化	否	
	ISODATA		变化	变化	不敏感	变化	需要	

线性判别函数的特点：

形式简单，容易学习；用于线性可分的模式类。

分段线性判别函数特点：

相对简单；

能逼近各种形状的超曲面。

一维正态曲线的性质：

（1）曲线在 x 轴的上方，与 x 轴不相交。

（2）曲线关于直线 x = μ 对称。

（3）当 x = μ 时，曲线位于最高点。

（4）当 x＜ μ 时，曲线上升；当 x＞ μ 时，曲线下降 .并且当曲线向左、右两边无限延伸时，以 x 轴为渐近线，向它无限靠近。

（5） μ 一定时，曲线的形状由 σ 确定。 σ 越大，曲线越“ 矮胖 ”，表示总体的分布越分散； σ 越小。曲线越“ 瘦高 ”。表示总体的分布越集中。

特征选择和提取的目的：

经过选择或变换，组成识别特征，尽可能保留分类信息，在保证一定分类精度的前提下，减少特征维数，使分类器的工作既快又准确。

K-L 变换进行特征提取的优缺点：

优点：

变换在均方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类别鉴别信息。

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，表明变换突出了模式类之间的差异性。

C* 为对角矩阵说明了变换后样本各分量互不相关，亦即消除了原来特征之间的相关性，便于进一步进行特征的选择。

缺点：

对两类问题容易得到较满意的结果。类别愈多，效果愈差。

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。

计算矩阵的本征值和本征向量缺乏统一的快速算法，给计算带来困难。

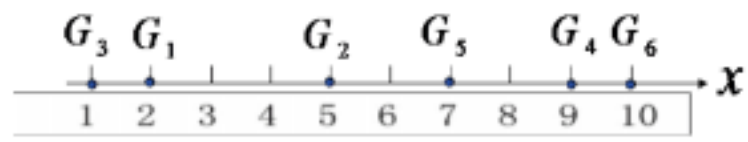
计算作图题

3.计算作图题

最大最小距离算法（小中取大距离算法　　）　：计算按照　P22 的例　2.1 来

最短距离准则进行系统聚类分类：　　P25 例 2.2

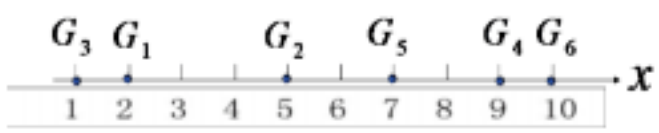
层次聚类分析：　（此处为例题）



(1) 设全部样本分为　6 类

(2) 作距离矩阵　D(0)

(3) 求最小元素



(4) 把 ω 1，ω 3 合并 ω 7=(1,3);

ω 4，ω 6 合并 ω 8=(4,6)

(5) 作距离矩阵　D(1)

(6) 若合并的类数没有达到要求，转　　(3)，否则停止

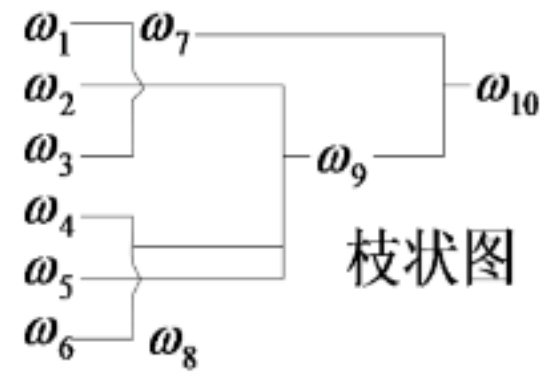
(7) 求最小元素

(8) ω 8，ω 5，ω 2 合并，ω 9=（2,5,4,6）

(9) 如无阈值要求，直至最后归为一类　　ω 10

	ω 1	ω 2	ω 3	ω 4	ω 5
ω 2	**3**				
ω 3	**1**	**4**			
ω 4	**7**	**4**	**8**		
ω 5	**5**	**2**	**6**	**2**	
ω 6	**8**	**5**	**9**	**1**	**3**

	ω 7	ω 2	ω 8
ω 2	**3**		
ω 8	**7**	**4**	
ω 5	**5**	**2**	**2**



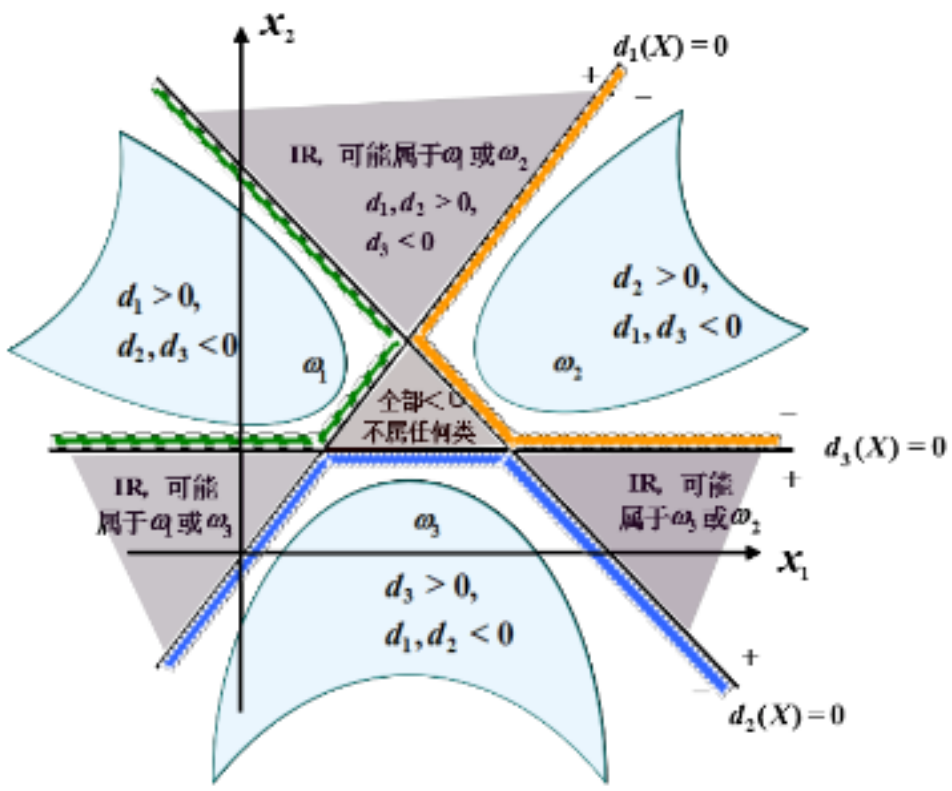
枝状图

K 均值算法　P28 例　2.3

判别函数：　P40

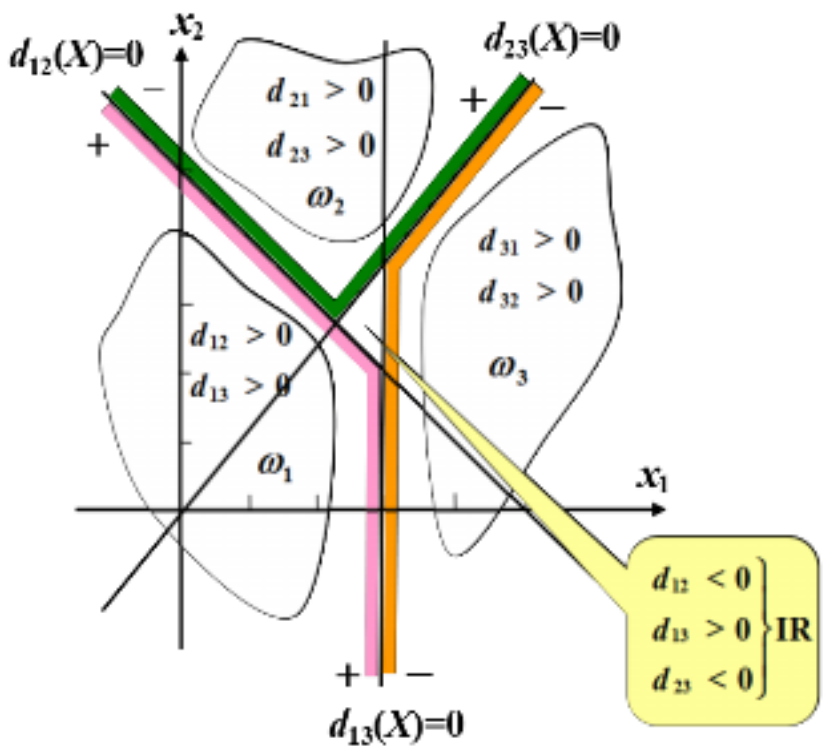
》第一种：例　**3.1&3.2P42**

此法将　**M**　个多类问题分成　**M** 个两类问题，识别每一类均需　　**M** 个判别函数。识别出所有的　　**M** 类仍是这　**M** 个函数。



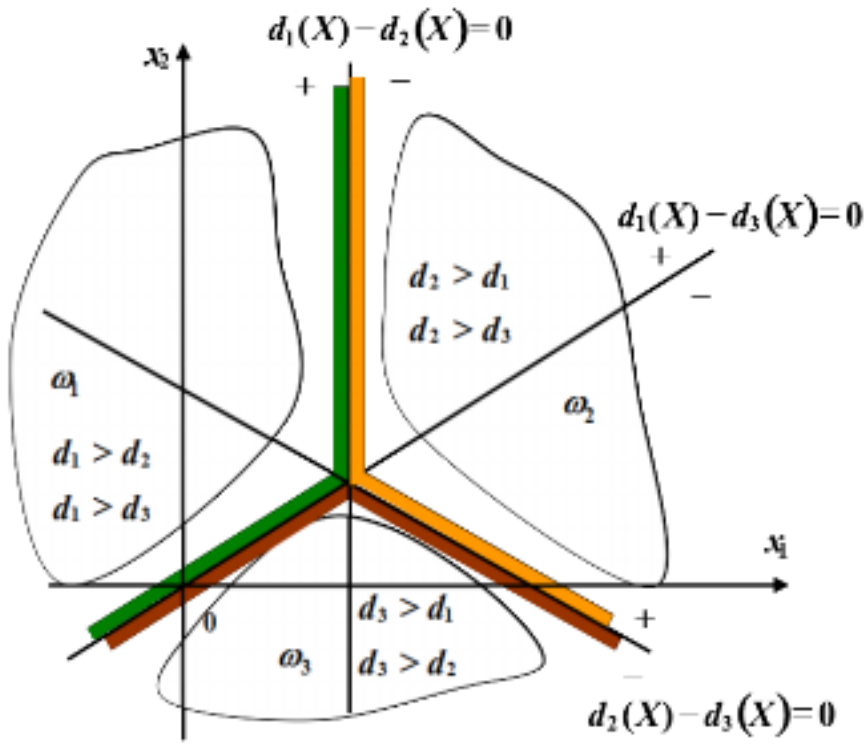
》第二种：　**P44** 例　**3.3&3.4**

每分离出一类，需要与　　**I**　有关的　**M-1**　个判决函数；要分开　　**M**　类模式，共需　　**M(M-1)/2**　个判决函数。对三类问题需要　**3(3-1)/2=3**　个判决函数。即：每次从　　**M**　类中取出两类的组合：



》第三种： **P45** 例 **3.5&3.6**

把 **M** 类情况分成了 **(M -1)** 个两类问题。



感知器算法： P54 例 3.8；P57 例 3.9

最小风险贝叶斯决策分类： P88 例 4.2

二维样本变换成一维样本： P138 例 5.2

样本压缩： P143 例 5.3