**第一章 模式识别概述**

➢一、模式识别的基本概念

(1)模式：指可测量且需要分类的对象的描述。

(2)模式类：把模式所属的类别或同一类别中模式的总体称为模式类（或简称为类）。

(3)样本：所研究对象的一个个体。(4)样本集：若干样本的集合。

(5)特征：指用于表征样本的观测，通常是数值表示的某些量化特征。

模式的基本属性：可观测性、可区分性、相似性。

模式识别：用计算的方法根据样本的特征将样本划分到一定的类别中去。

➢二、模式识别系统

1）模式采集：模式的数字化表达

2）预处理：提高输入模式的质量，去噪声、去模糊、模式结构转换（常用方法：滤波、变换、图像增强、图像恢复）

3）特征提取与选择：是数据降维的两种途径。特征的分类：度量/属性特征-待识别对象的度量或属性参数；基元特征-有效描述的基本子模式（轮廓、边缘、纹理）特征表达可以使用特征向量方法、结构方法描述。

4）训练：根据特征空间分布决定分类器参数。分为监督学习、无监督学习、半监督学习。

5）分类决策：对样本进行分类识别的过程。

6）后处理：根据分类器输出确定动作。

➢三、模式识别的基本方法分类

根据特征表达：统计模式识别（包含几何方法、概率方法）、结构模式识别

根据是否有监督：有监督（非参数方法、参数方法）、无监督（聚类）

➢四、有关模式识别的若干问题

1.紧致性：临界点的数量与总的点数相比很少；内点间可以用光滑线连接，连接线上的点仍属于原集合；每个内点有足够大邻域，邻域中只包含同一集合的点。

2.特征选取

3.相似性度量：特征空间中描述样本属性即使用**距离函数**描述相似性度量。常用距离函数：欧氏距离、马氏距离、Minkowsky距离，Manhatten距离，City Block距离。

距离函数应该满足的四个条件：





4.分类：不存在纯客观分类标准；分类问题不是纯数学问题。

5.性能评价：正确识别率=正确分类数/总数；错误识别率=错误分类数/总数；拒绝识别率=拒绝分类数/总数。

**第2章 贝叶斯决策理论**

➢贝叶斯决策基础知识

模式分类：根据待识别对象的观测值，确定其类别；目标：实现没有错误的分类，或使分类错误率最小；混叠现象：特征选择不合适，给分类带来不确定性

样本由n维实数特征组成：；c个类别（类别数已知）：；决策：是从样本空间S，到决策空间Θ的一个映射，表

示为

D: S --> Θ；Bayes决策常用的准则：✓最小错误率准则✓最小风险准则

全概率公式；

贝叶斯公式

➢基于最小错误率的贝叶斯决策**（证明题：错误率定义-两类条件错误概率讨论 -画图说明最小错误率的分界线）**

已知类别，先验概率，类条件概率密度；目标：设计一个具有最小误分概率的两类分类器；设观测样本为X，则X的全概率密度为；

最小错误率的贝叶斯准则：or

优势：最大后验概率判决的错误概率最小；缺点：没有考虑判决的风险

➢基于最小风险的贝叶斯决策

将样本x看作d维随机向量; 状态空间由c个可能状态(c类)组成; "决策空间由k个决策" α\_i，i=1，2，...，k组成。k可与c不同，例如可以采取“拒绝”等其他策略

条件风险

，当风险为0-1时等价于最小错误率准则**（证明）条件风险：**

➢正态分布时的统计决策

单变量正态分布：(1) 概率密度函数

(2) 均值、方差及标准差

多变量正态分布：(1) 概率密度函数

各分量之间的不相关性等价于独立性，边缘分布和条件分布仍服从正态分布，服从多变量正态分布的随机向量经线性变换后仍服从多变量正态分布，多变量正态随机向量的各分量的线性组合服从正态分布

➢贝叶斯分类器设计

选择具有最大后验概率的类作为该对象所属的类

1.判别函数

2.基于最小错误率的判别函数决策面为

、

多类情况：定义一组判别函数，用于表示多类决策规则，可定义为or

or，决策面为，后加上最大值选择器

多元正态分布下的最小错误率贝叶斯判别函数和决策面：多类判别函数选择ln形式，g最大的划分到对应类

为

第一种情况：判别函数可进一步简化

，

，

若类别i和j相邻，则相应的决策面为，

，

第二种情况：，

，

**第三章 概率密度函数估计**

参数估计的分类：监督参数估计-样本类别、类条件概率密度已知，表征概率密度的某些参数未知，估计这些参数；非监督参数估计-概率密度函数形式已知，样本类别未知；非参数估计-样本类别已知，概率密度函数形式未知，估计概率密度函数数值。

➢一、极大似然估计

1. 基本假设：1）参数是未知确定量；2）样本集按类别分为c个样本子集，均为从类条件概率密度为的总体中独立抽取；3）具有确定函数形式；4）不同类别参数在函数上独立

2. 似然函数：已知某一类样本集包含N个样本，，待估计未知参数，则在概率密度函数为时获得样本集的联合概率密度为，称为参数相对于样本集的似然函数。

3. 极大似然估计量：为参数相对于样本集的似然函数，使似然函数最大的称为的极大似然估计量。

4. 一元参数：的解。定义对数似然函数，由于对数函数单调递增，因此极大似然估计量也是方程的解。

5. 多元参数：待估计参数向量，对数似然函数，在样本独立抽样的条件下可以写成，求解关于梯度的线性方程组即可得到极大似然估计量。

➢二、贝叶斯估计（针对参数，贝叶斯分类针对数据和选择）

**（原理证明）**R-加E减E-展开证明一项是0

1. 目的：找出估计量使得其带来的贝叶斯风险最小。

2. 步骤：1）确定的先验分布；2）由样本集求出联合概率分布；3）利用贝叶斯公式求出后验分布；4）利用**求解定理**得贝叶斯估计量，如果损失为，则

➢三、贝叶斯学习

通过样本集推断总体分布，与前两种方法的不同是估计后验分布后直接求类条件概率密度设样本集，代入贝叶斯公式可得 ，其中为无样本条件下的概率密度，反复使用上式可以得到一系列对概率密度函数参数的估计，随样本数增加，概率密度逐渐趋向于以的真实值为中心的尖峰，当样本无穷多时收敛于在参数真实值上的脉冲函数，就把这一过程称为贝叶斯学习。

➢四、正态分布的参数估计

1. 极大似然估计

单变量正态分布参数：，极大似然估计方程：，其中。

极大似然估计量，。

多元正态分布的估计量，

2. 贝叶斯估计

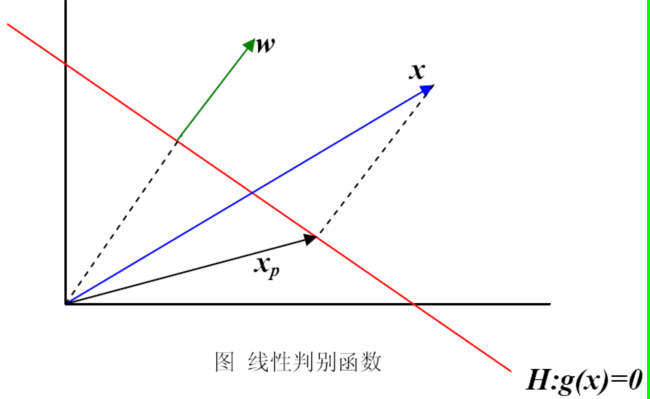
样本集取自分布，均值为未知参数，且有先验分布，估计量，

➢五、非参数估计

目标：当给定样本集类别属性已知并且经过预分类，能够得到相应类条件概率密度的估计；当给定样本集是类别属性未知的混合样本集时，能够给出相应混合概率密度的估计。

基本思想：令是包含样本点的一个区域，体积为，共个样本其中个落在中，可对概率密度做出估计。构造一系列包含的区域，则有一系列估计，当满足下列条件时，收敛于：，，

1. 近邻估计

基本思想：使体积为数据的函数，预先确定的某个函数，在周围选择一个体积，不断增长直到捕获个样本为止，这些样本为的个近邻。

仍使用估计公式

**第4章 线性判别函数**

➢线性判别函数的基本概念

定义：对于c 类问题，可以定义c 类判别函数、，现在的问题是如何利用样本去估计 w和 b。xp为投影

任意一点，带入，==，，这说明了判别函数正比于代数距离。

➢Fisher线性判别分析

Fisher线性判别的基本思想是希望投影后的**一维数据**满足：

1两类之间的距离尽可能远；2每一类自身尽可能紧凑。形式y=wTx

均值向量，类内离散度矩阵

，类间离散度矩阵

在一维Y样本空间：，类内

，类间离散度矩阵，Fisher准则函数

， ， 取极大值，拉格朗日法，，对应矩阵的特征值问题，由于只关心方向，取，否则

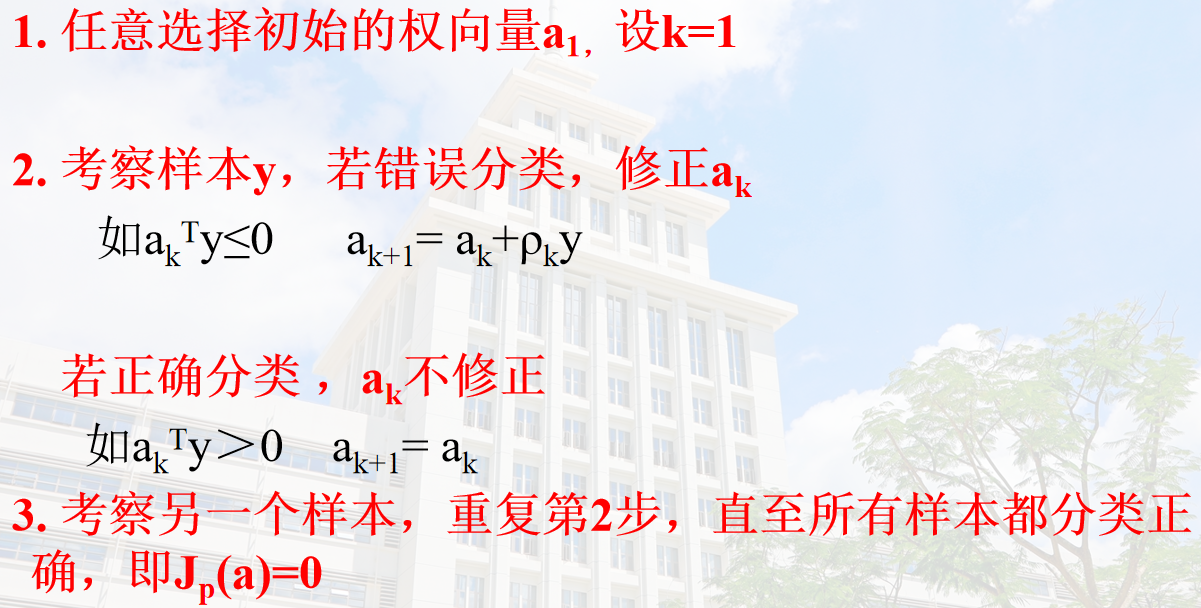
**➢感知器（计算大题）**

，，

称y为增广样本向量，只对线性可分样本有收敛解

**规范化（开头增1取负值）-迭代修正ak（如akTyi≤0，ak+1= ak+ρkyi）-权向量相同-判别函数ak**

感知器算法：

➢最小平方误差判别

根据前面提到的感知准则函数的目的是找一个权向量a使得到满足的不等式a^Ty\_n>0，n=1，2.…N的数目最大，从而使错分的样本数目最少。MSE考虑引入任取正常数，Ya=b，不能保证所有a使得a^Ty\_n>0，但使得误差e=Ya-b最小

准则函数为：，

如果对应同一类样本的bn选择为相同的值，那么最小平方误差方法的解等价于Fisher线性判别的解。如果对所有样本都取bn=1，那么当N趋近于无穷大时，最小平方误差方法的解是贝叶斯判别函数的最小平方误差逼近。

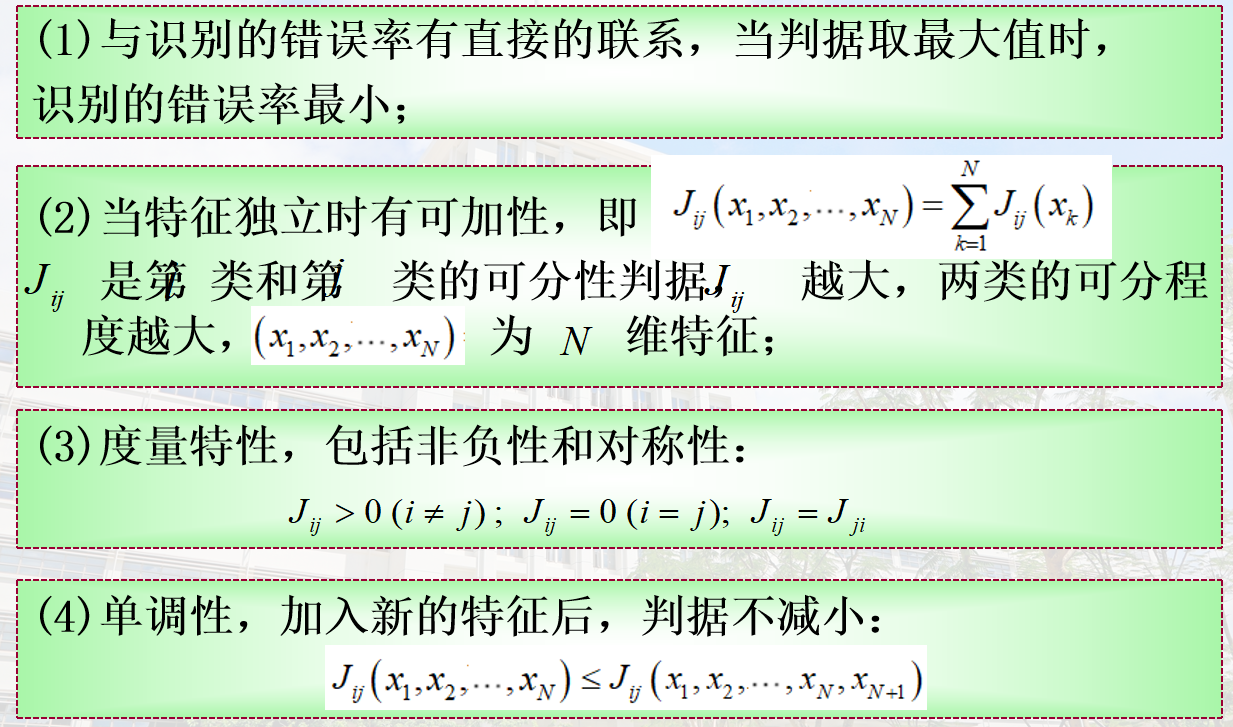
➢多分类问题

，形式简单、意义明确但不可识别区域过多；，较少不可识别区域

**第五章 特征选择**

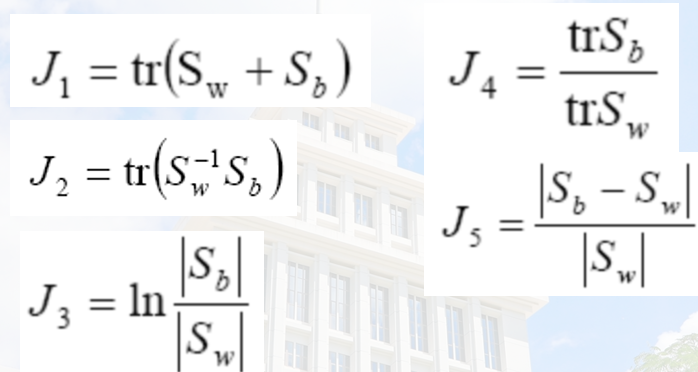
➢1.引言 数据降维的两种途径：1)删掉一些次要的特，即特征选择；2)通过变换来实现降维、优化，即特征提取

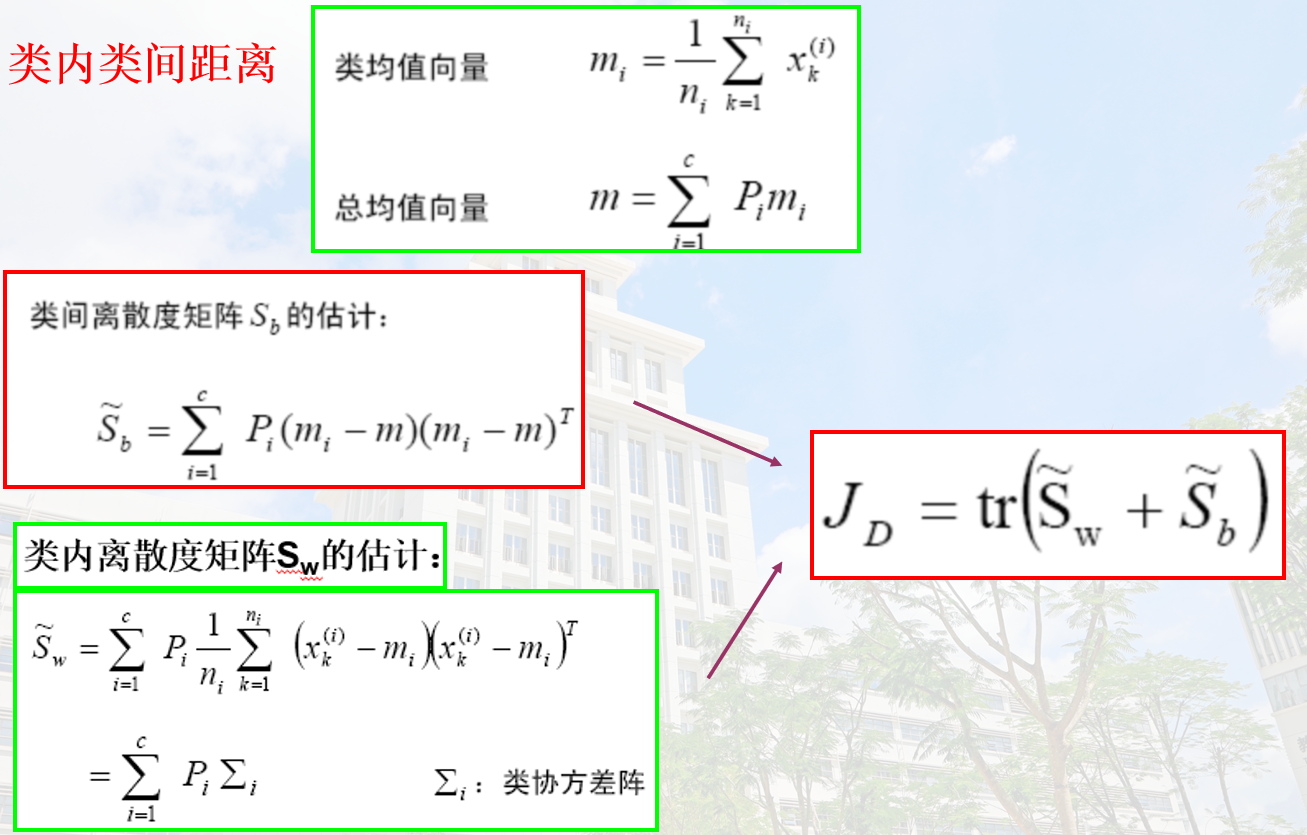
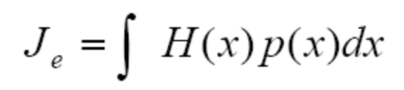
**➢2.特征的评价准则**

可分性判据应满足的条件：

**1）基于类内类间距离**

各类特征平均距离

常用的基于类内类间距离的可分性判据：

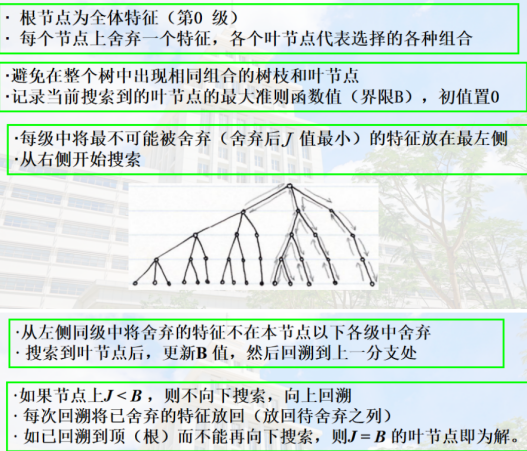
1. **基于概率分布的可分性判据：**Jp为两个概密函数距离，应满足三个条件：Jp非负；Jp=Jmax，完全不重叠；Jp=0完全重叠
2. **熵可分性判据：**据x能完全确定，则就没有不确定性，此时熵为0，x有利于分类Je小，可分性好；Je大则重叠性大，可分性不好

**➢3.特征选择 从D维特征中选取d维（d < D），使分类性能最佳（判据J最优）。**

穷举算法:d或D-d很小；

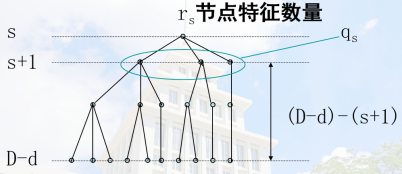
分支定界算法：从顶向下，有回溯。条件；准则函数单调

最优搜索方法：树的每个节点表示一种特征组合，树的每一级各节点表示从其父节点的特征组合中再去掉一个特征后的特征组合，其标号k表示去掉的特征是xk。

算法要点：

由于每一级只舍弃一个特征，整个搜索树除根节点的0级外，还要D-d 级，即全树有D-d 级。6个特征中选2个，故整个搜索树需4级，D-d 级是叶节点，有CDd个叶节点。rs表示集合中元素数目，qs表示当前节点子节点数

后续子节点数qs=rs-(D-d-s-1)，q0=d+1，r0=D

非最优搜索方法 1)单独最优的特征选择，计算各特征单独使用时的判据值并以递减排序，选取前d个分类效果最好的特征（J可分）2)增添特征法 每次从未选入特征中选择一个，使其与现有特征组合一起J值最大，直至到达维数d，缺点：某特征一旦选入，后边的D-k特征中的某组合比它好，无法剔除。

3)剔减特征法一种自上而下的搜索方法，从全部特征开始每次剔除一个特征，所剔除的特征应使尚保留的特征组合的J值最大4)增l减r法(l-r法)，对已选入的k征2方法到k+l个，再3剔除r个

**第六章 特征提取**

y=W(x)，线性情况下y=WTx。目的：消除特征之间可能存在的相关性和/或减少计算量；目标：根据训练样本求适当的W，使得某种特征变换准则

➢一、基于类内类间距离可分性判据的特征提取

目标：求特征变换使得可分性判据

举例：

Lagrange函数：其中为对角阵，I为单位阵

得由迹的性质可知

最优变换矩阵为的前个本征值（降序排列）对应的本征向量组成的矩阵。

➢二、基于K-L变换的特征提取**（大题）**

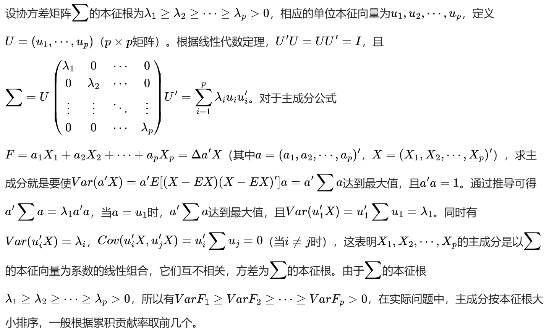
1. 利用训练样本集合估计出相关矩阵：Rx=E[XXT]

2. 计算相关矩阵Rx特征值、特征向量（归一化）并按特征向量由大到小排列

3. 选择前M个特征向量构成变换矩阵，，Y为变换后特征

➢三、主成分分析

概念：基于K-L变换的提取特征的一种最优正交线性变换，可以有效去掉一个随机向量中各元素间的相关性。PCA 是一种非监督的算法， 未必是最有利的

****目的：寻找能够表示采样数据的最优投影子空间。**将许多相关性很高的变量转化成个数较少、能解释大部分原始数据方差且彼此互相独立的几个新变量，也就是所谓的主成分（证明求主成分就是寻找X的线性函数*a’ X*使相应的方差尽可能地大）**

**求解：**本征向量被称为“主分量”，每个样本被它在前几个主分量上的投影近似表示。主成分之间不仅不相关，而且它们的方差依次递减；**步骤：**K-L变换中将相关矩阵换为协方差矩阵，m为均值，其他步骤相同。

人脸识别算法：将待识别的人脸图像 与平均脸的差值脸投影到特征空间，得到其特征向量表示；找出距离待识别的人脸图像最近的训练样本，将训练样本的类别作为待识别图像的类别

1. **无监督学习**

监督模式识别：（已知）样本集 → 训练（学习）→ 识别（分类）

无监督模式识别：（未知）样本集 → 无监督学习（聚类分析）→ 后处理

➢单峰子集（类）的分离方法

假定：各类样本分布是单峰，根据单峰来划分子集

基本思路：把样本按某准则投影到某一维坐标上，在这维上估计样本概率密度函数，根据概率密度函数的单峰划分子集。如果这一维只有一个峰，则寻找下个投影方向；

基本步骤：1、计算样本协方差矩阵最大本征特征值对应本征向量uj，把样本投影到uj，（PCA）;2、估计投影后样本概率密度；3、求p(vj)的极小值点，作垂直uj的超平面作为分类面;4、如果没有这种极小值点，则用下一个本征值对应本征向量;5、对划分的每一个子集重复上述过程，直到不能再分（所有方向上都是单峰）。

➢类别分离的间接方法（动态聚类法） 同一类样本的特征向量应是靠近的。前提：特征选取合理

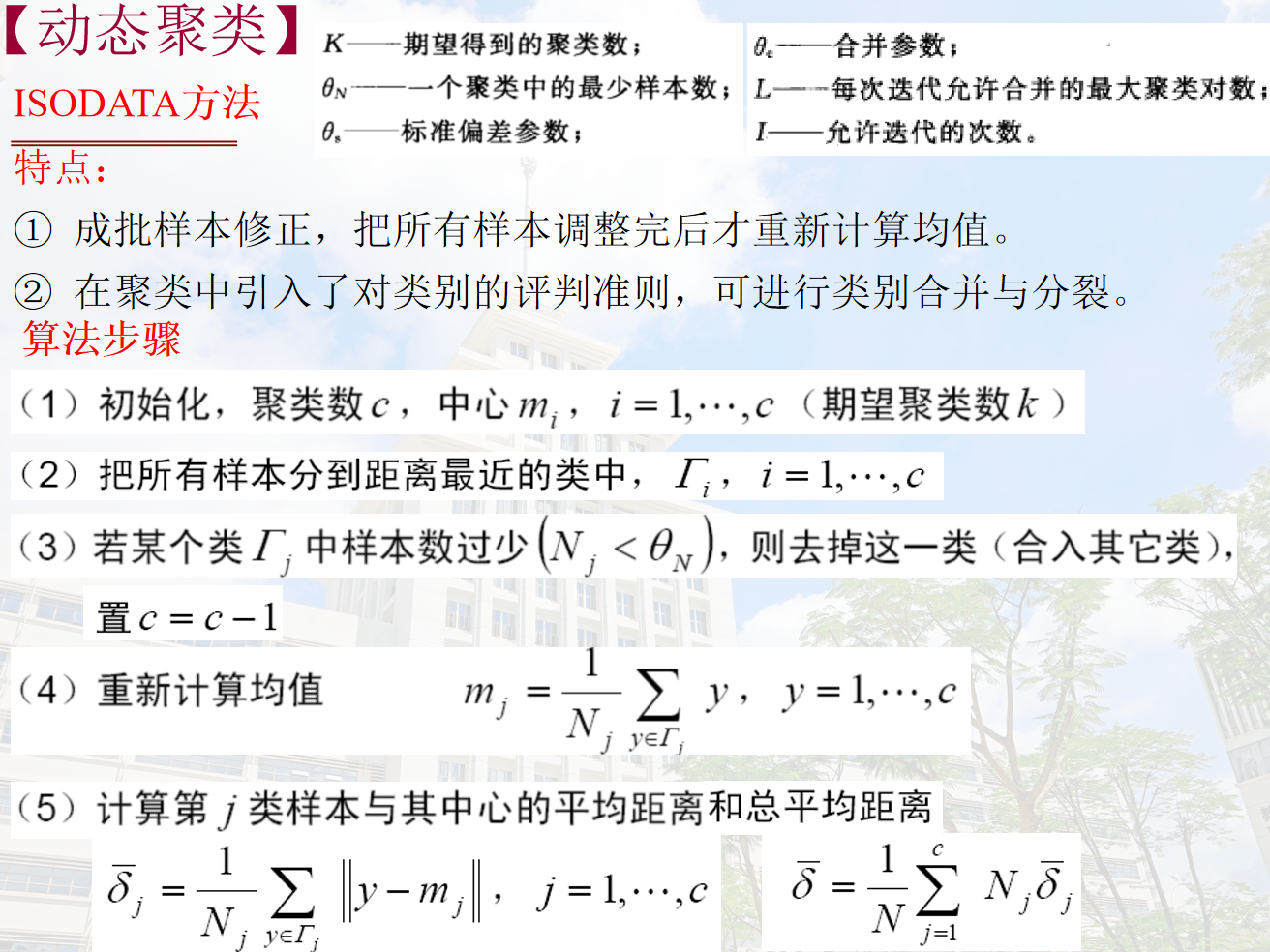
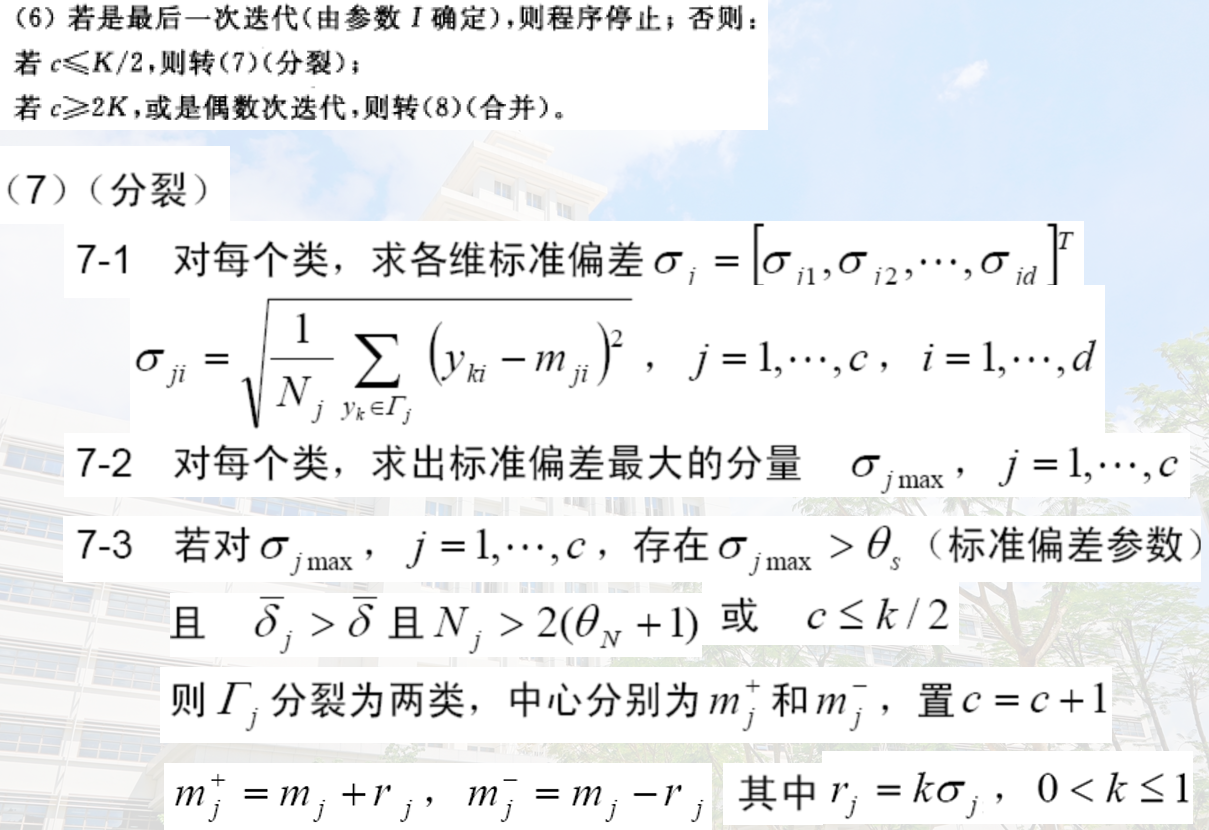
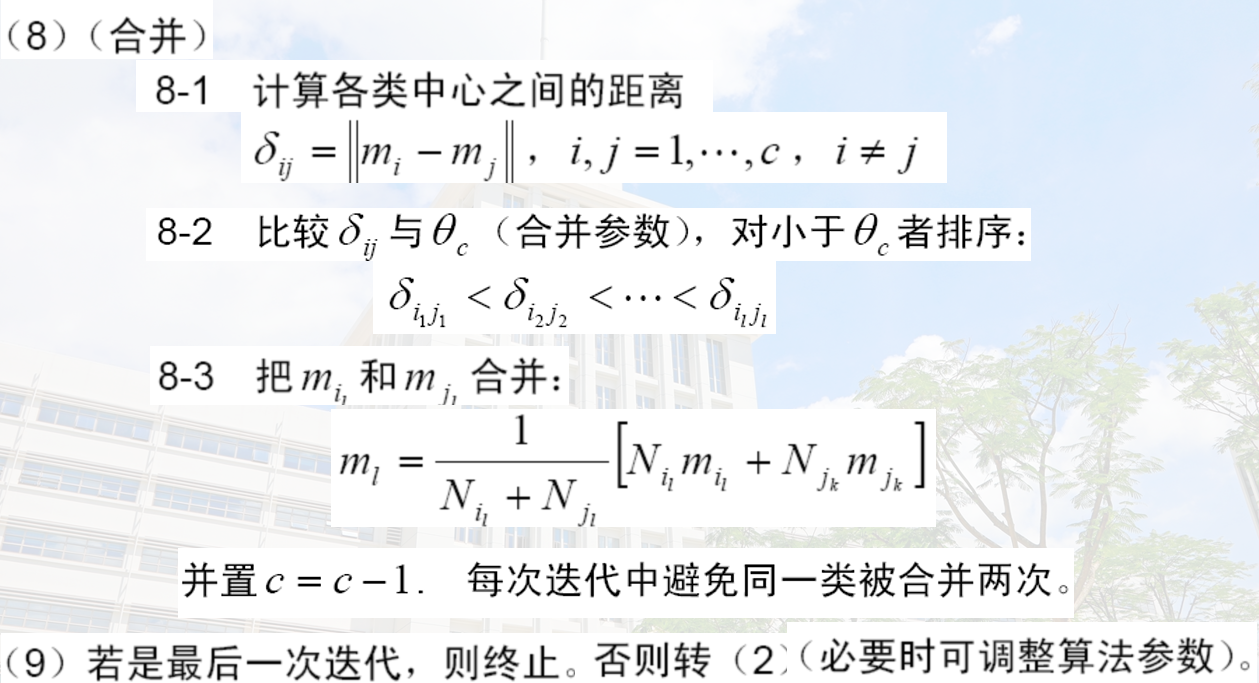
动态聚类的三个要点：1. 选某种距离作为样本相似性度量 2. 定义某个准则函数，用于评价聚类质量。 3. 给定某个初始分类，然后用迭代算法找出使准则函数取极值的最好聚类结果。

1. **均值聚类：**（也叫K-均值）

第j类的准则函数为：，对于所有c个模式类，。

上式表明Γj类聚类中心应为该类样本均值。C均值聚类算法描述：1任选c个初始聚类中心，作为代表点：;2按最小距离原则将其余样本分配到c个聚类中心的某一个; 3计算各个聚类中心的新向量值;重复直至。结果受到所选聚类中心个数初始位置读入次序影响。

**ISODATA方法：**

➢分级聚类方法算法（从底向上）：（1）初始化，每个样本形成一类（2）把相似性最大（距离最小）的两类合并（3）重复（2），直到所有样本合并为两类。

常用的类间相似性度量：最近距离、最远距离、均值距离。

➢**模糊学习** 每元素都以一定程度属于某集合也可以同时以不同程度属于几集合

隶属度μA(x)：x属于集合A的程度；自变量所有可能属于A的对象；值域[0，1]，

μA(x)=1即x属于A，μA(x)=0即x不属于A

**模糊c均值算法（FCM）**：由于可能对某些野值隶属度可能较大，放松归一化条件得到改进模糊c均值算法（AFC）：，特点：有更好的鲁棒性，对聚类数目不十分敏感。

**AFC（改进）**：较前面提到的模糊C 均值算法具有更好的鲁棒性，它不但可以在有孤立样本存在的情况下得到较好的聚类效果，而且可以放松隶属度条件，而且因为放松了隶属度条件，使最终聚类结果对预先确定的聚类数目不十分敏感。与确定性C均值算法和模糊C均值算法一样，改进的模糊C均值算法仍然对聚类中心初值十分敏感，为得到较好的结果可以用确定性C均值算法或模糊C均值算法的结果作为初值。

k近邻的问题：当样本稀疏时，只考虑样本的近邻顺序而不考虑距离远近不合适。

**第九章 前馈神经网络**

➢一、神经网络

1. 人工神经元

线性模型：激活函数：激活函数的性质：连续可导（允许少数点不可导）的非线性函数；尽可能简单；值域合适区间内；单调递增。

2. 常见激活函数（见右图）

➢二、前馈神经网络

（全连接神经网络，多层感知机）

1. 网络结构：各神经元分属不同层，层内无连接；相邻两层之间全部两两连接；无反馈，有向无环图。

2. 信息传递过程，l表示层数 ，x=a(0)迭代

通用近似定理：对于具有线性输出层和至少一个使用“挤压”性质的激活函数的隐藏层组成的前馈神经网络，只要其隐藏层神经元的数量足够，它可以以任意的精度来近似任何从一个定义在实数空间中的有界闭集函数。

➢三、参数学习

训练：交叉熵损失函数：loss=-yTlog(hat\_y)

训练过程：1、前向计算各层状态、激活值；2、反向计算各层对参数的偏导数；3、梯度下降更新参数(链式法则和反向传播见ppt)

**第十章卷积神经网络**

全连接前馈神经网络(1)权重矩阵的参数非常多(2)很难提取局部不变特征

卷积神经网络（CNN）：前馈网络，结构特性：局部连接，权重共享，空间或时间上的次采样；经常用在信号处理中，用于计算信号的延迟累积；

一维卷积：

逆序相乘，也称为卷积翻转，wk为卷积核/滤波器；

卷积核大小K，滑动步长S，零填充P，输出序列长度M’=

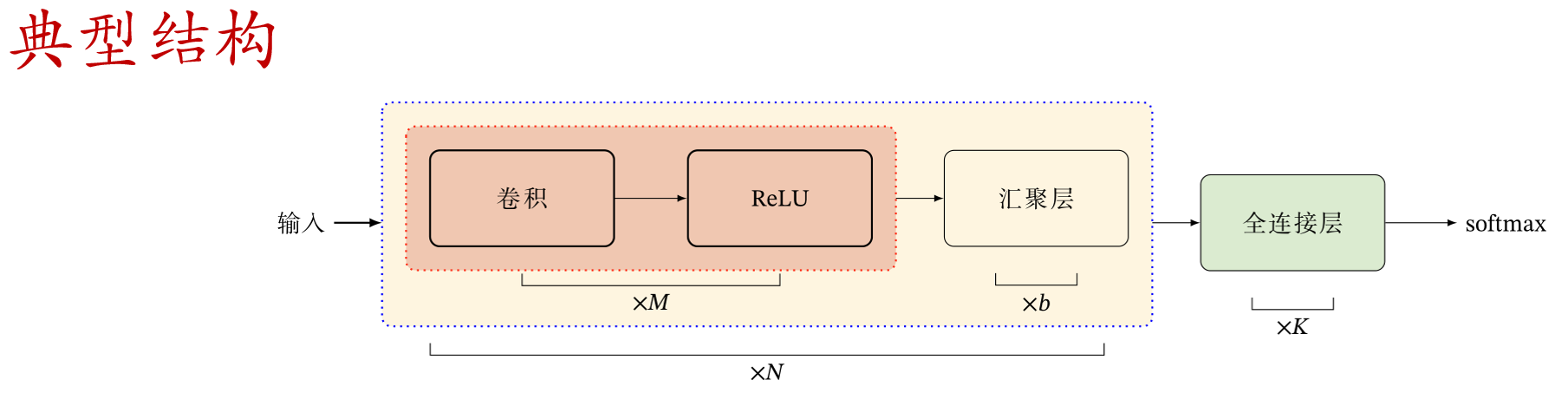
1、窄卷积：输入M，卷积核K，步长T=1；两端补零P=0；输出M-K+1；

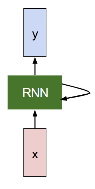
2、宽卷积：步长T= 1，两端补零P=K-1，卷积后输出长度M+K-1；

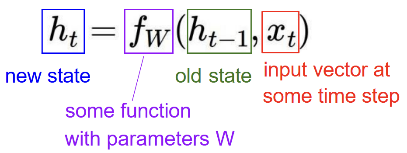
3、等宽卷积：步长T= 1，两端补零P=(K-1)/2，卷积后输出长度M，K为奇数

典型的卷积层为3维结构：输入 ， 输出 y∈ ，卷积核 UXV， 需要参数P\*(D\*(U\*V)+1)个

互相关：不需要翻转，简化计算卷积网络是由卷积层、汇聚层、全连接层交叉堆叠而成：趋向于小卷积、大深度、趋向于全卷积；

其它卷积种类：（1）低维特征映射到高维特征：空洞卷积（通过给卷积核插入“空洞”来变相地增加其大小。）Inception模块 v3用多层小卷积核替换大卷积核，以减少计算量和参数量，残差卷积

****卷积的应用：AlphaGo、目标检测（Object Detection）、Mask RCNN、OCR、图像生成etc

**十一、Recurrent Neutral Networks**

**处理序列** 也叫做序列到序列问题，用编码器-解码器模型解决

**基本思想:**RNN中含有一个随着序列的产生而更新的内部状态 RNN的**优点**：1.可以处理任何长度的输入2.步骤t的计算可以（理论上）使用许多步骤后的信息3.更长的输入不会增加模型大小4.每个时间步骤都应用相同的权重，因此处理输入的方式是对称的 **缺点**：1.递归计算很慢2.在实践中很难从许多步骤后访问信息

**LSTM架构**：使RNN更容易在许多时间步长上保存信息，例如，如果f＝1和i＝0，则该cell的信息被无限期地保存。相比之下，普通RNN更难学习在隐藏状态下保存信息的递归权重矩阵Wh。隐藏状态ht: 短期记忆:ct : 长短期记忆网络参数W 长期内存。LSTM不能保证没有消失/爆炸的梯度，但它确实为模型学习长距离依赖关系提供了一种更容易的方法 **总结**：RNN在架构设计中具有很大的灵活性。原始RNN很简单，但效果不太好。LSTM或GRU很常见：它们的加性相互作用改善了梯度流。RNN中梯度的反向流可能会爆炸或消失。分解由渐变剪裁控制。消失是由加性相互作用（LSTM）控制的。更好/更简单的架构是当前研究的热点，也是对序列进行推理的新范式。需要更好的理解（理论上和实验上）。

**十二、Attention and Transformers**

RNN与Transformer比较：

**RNN：** （+）LSTM可以很好地用于长序列。 （-）期望输入的有序序列 （-）顺序计算：只有在完成前一个隐藏状态之后才能计算后续隐藏状态。

**Transformers** （+）擅长长序列。每个注意力计算都会考虑所有输入。 （+）可以对无序集或具有位置编码的有序序列进行操作。 +）并行计算：所有输入的所有对齐和注意力得分都可以并行完成。 （-）需要大量内存：需要为单个自注意力头计算和存储N x M个对齐和注意力定标器。（但GPU越来越大、越来越好）

**总结：**向RNN添加注意力使其能够在每个时间步长“关注”输入的不同部分；一般注意力层是一种新型的层，可用于设计新的神经网络架构；Transformers是一种使用自我注意力和层规范的层。

