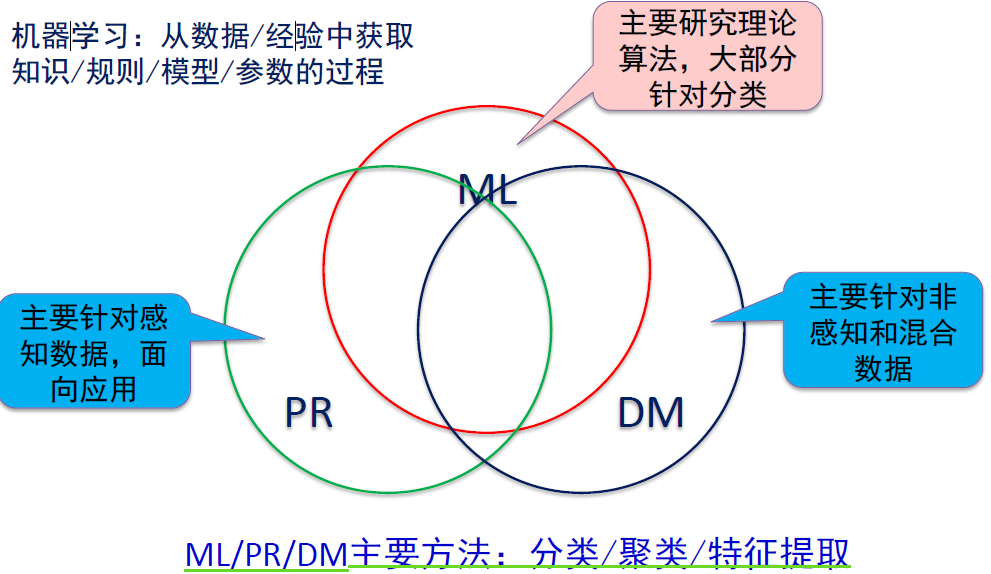
# 第一章 绪论

1、模式的两个层次：样本和类别

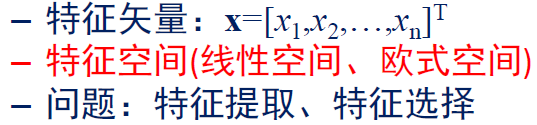
《模式识别》的核心内容：模式分类

- 特征提取/选择、聚类分析、分类器设计（机器学习）

2、模式识别-机器学习-数据挖掘，三者之间的特点和差异：



3、识别对象表示：**特征**

• 线性空间：又称为向量空间，定义了数乘和加法，一组基底（Basis）的线性组合可以表示空间中所有的点；

* 范数空间：定义距离，向量的长度；
* 欧式空间：又称为内积空间，4维或N维的理论无穷大的空间，对现实空间的规则抽象和推广，涉及向量的长度、角度和投影运算。

4、二维特征：

线性分类器-多个特征组合比单个特征具有更好的可分性；

非线性分类器-二次分类器，决策边界为平滑曲面。

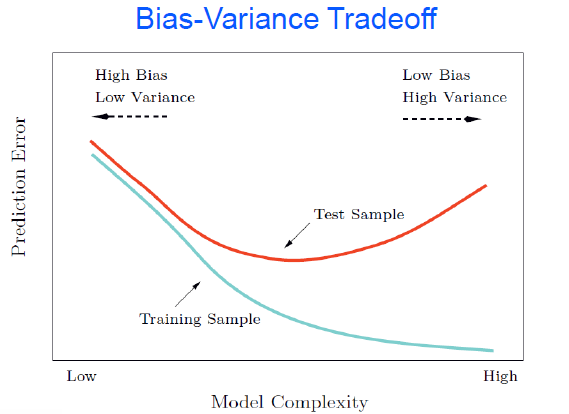
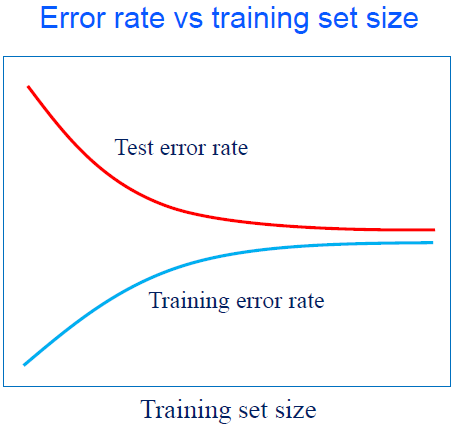
5、识别-训练过程：

物理环境→数据获取→分割→预处理→特征提取→特征→分类→后处理

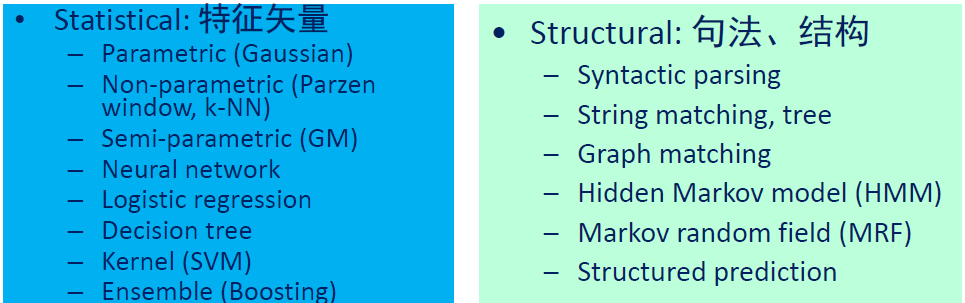
深度学习：特征表示与分类器同时学习（end-to-end）

6、泛化性能(Generalization Performance)：测试数据上的分类性能；

过拟合/过学习：用复杂分类器能将训练数据分类错误率降到极低，泛化性能降低。



7、根据**表示方式**分类：



混合统计-结构模型：Attributed graphs, HMM and MRF/CRF



8、根据**学习方法**分类：

• 监督(Supervised)学习：训练样本有类别标号；

* 无监督(Unsupervised)学习：训练样本无类别标号，得到数据结构表示或分布；
* 半监督(Semi-supervised)学习：训练样本一部分有类别标号，一部分没有；
* 强化学习（Reinforcement Learning）：学习过程中给出奖惩信号；
* 域自适应（Domain Adaptation）：测试样本分布发生变化，分类器参数自适应；
* 在线（增量）学习（Online (Incremental) Learning）：对已经学习的知识具有记忆功能，学习新信息中的有用信息。（学新不忘旧）

9、**生成/判别模型**：

• 生成(Generative)模型：表示各个类别内部结构或特征分布p(x|c)；

生成学习：得到每个类别的结构描述或分布函数，不同类别分别学习。

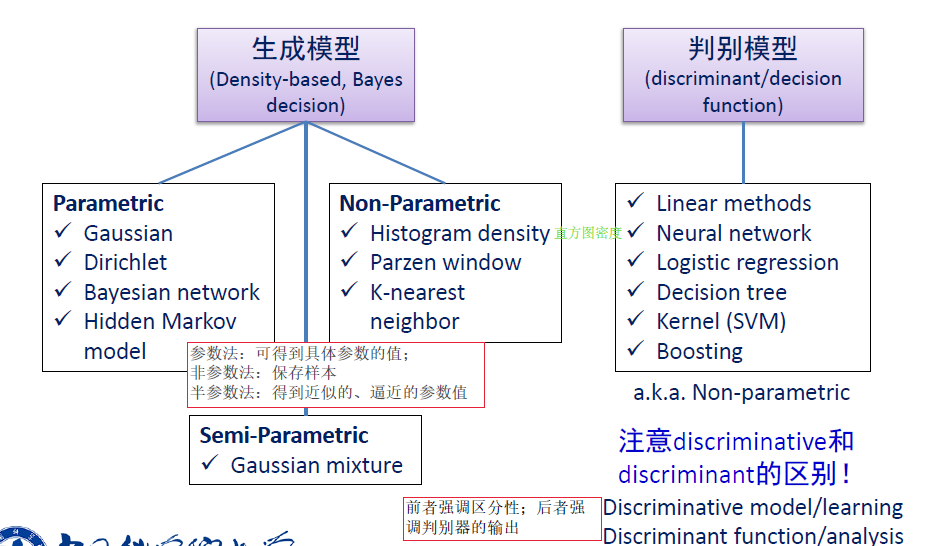
* 判别(Discriminative)模型：表示不同类别之间的区别，判别函数(Discriminant function)、边界函数或后验概率P(c|x)；

判别学习：得到判别函数或边界函数的参数，所有类别样本同时学习。

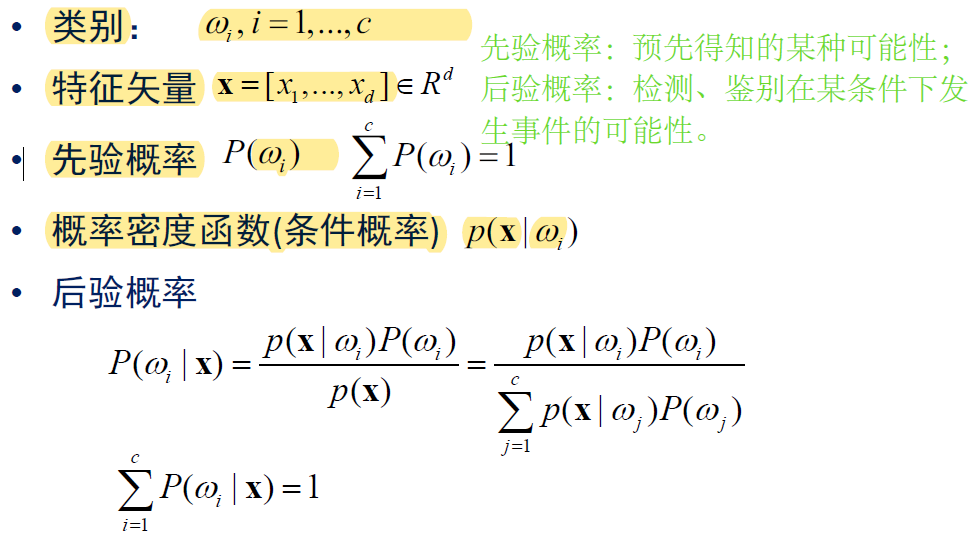
10、非参数方法：线性分类器设计、神经网络与深度学习、SVM与核方法、决策树等。

# 第二章 贝叶斯决策理论

1、统计模式识别方法：

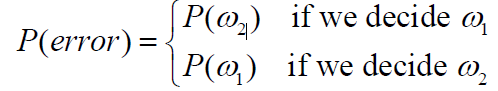


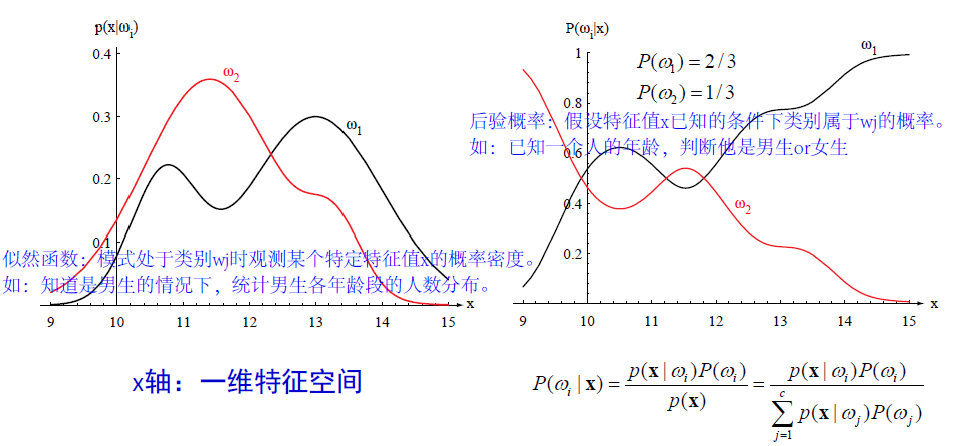
2、**先验概率、概率密度函数**与**后验概率**：



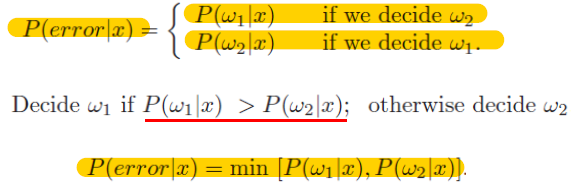
没有传感器（特征），根据以往经验得到先验概率；有传感器（特征X），统计后验概率。

**3、最小错误率决策**：





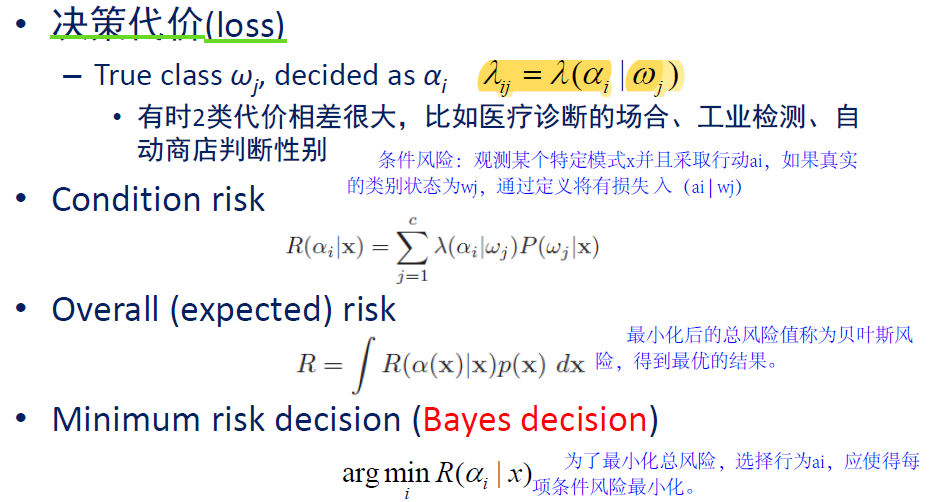
* 基于后验概率的最小错误率决策（两分类）



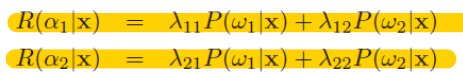
* 单个特征，多个特征组合
* 基于联合概率密度的最小错误率决策：（贝叶斯公式的等价变形，p(x)标量因子，可忽略）



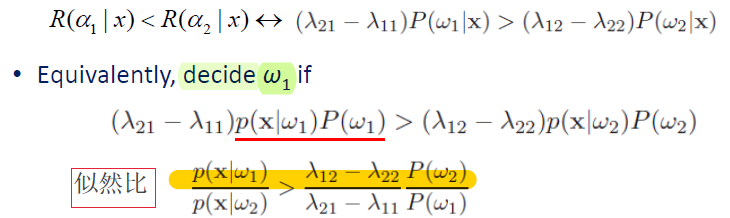
**4、最小风险决策（**贝叶斯决策的一般形式**）**：

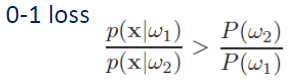


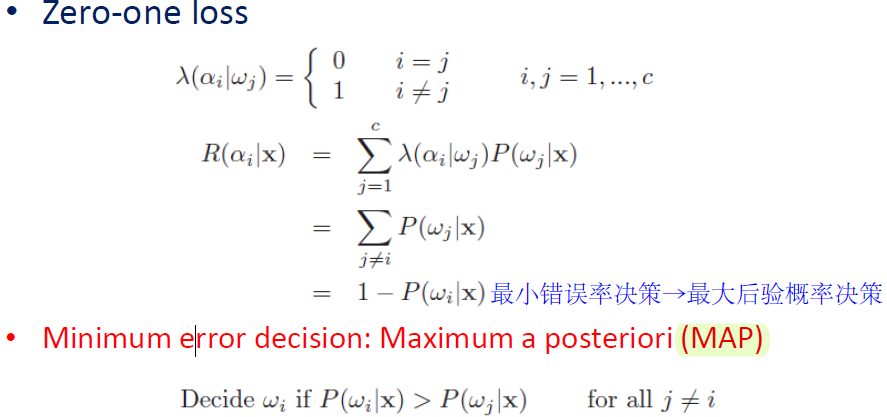
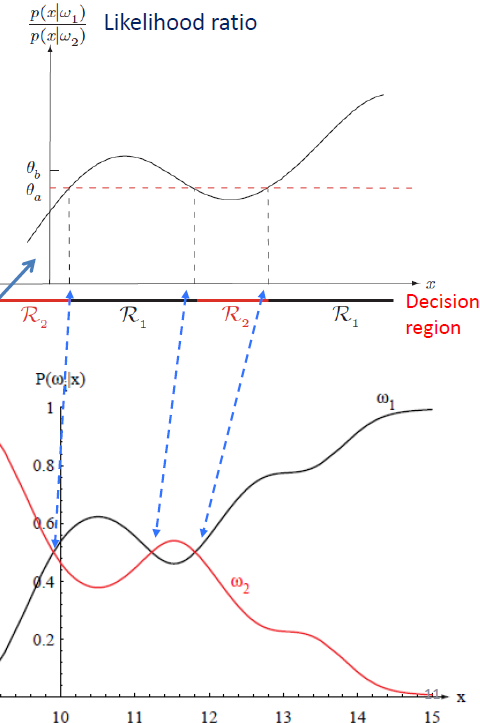
* 两分类下的最小风险决策：



* **决策规则：**

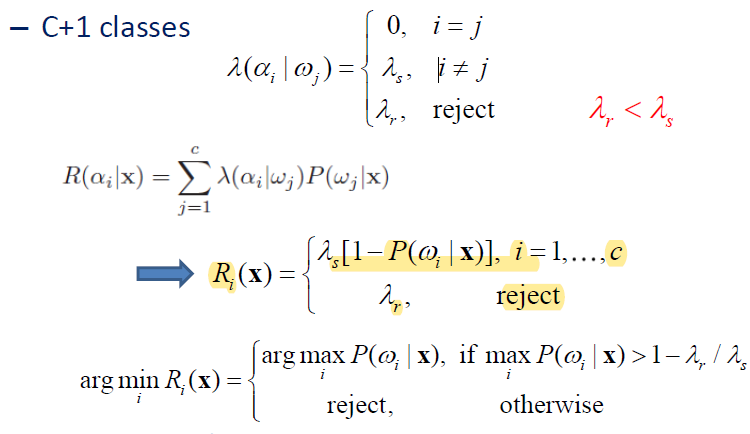


**▶ 多类别下的最小错误率分类**

对应后验概率函数，谁高，判为谁

对右图的解释：似然比超过某个不依赖于x的阈值seta，则可判决为w1。

▶ **带拒识的决策**（考点！！！，第二章问题13）



▶**极小化极大准则**（第一章习题3）

最小化最大可能的总风险，使先验概率取任何一种值时所引起的总风险尽可能小。即：R关于先验概率P(w1)的线性函数，比例常量为0；那么当取到贝叶斯风险最大的先验概率时，得到的R是极小化后的最坏贝叶斯风险。

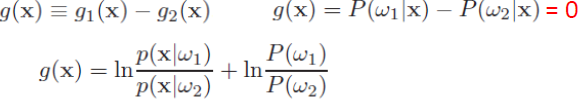
**5、判别函数**和**决策面**

• 分类决策：

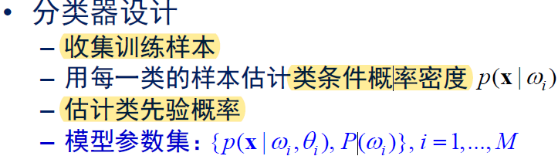
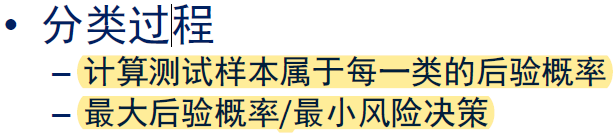
* 常见的判别函数：

，，，

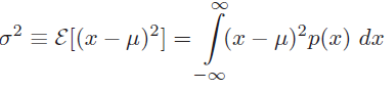
* 决策面：特征空间中二类判别函数相等的点集合。加深对特征空间和分类器性质的理解。

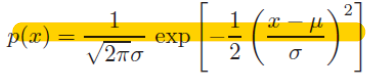
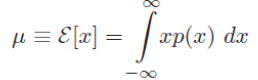


▶贝叶斯决策下的模式分类：

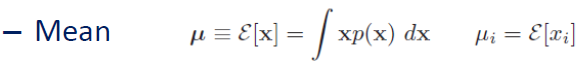


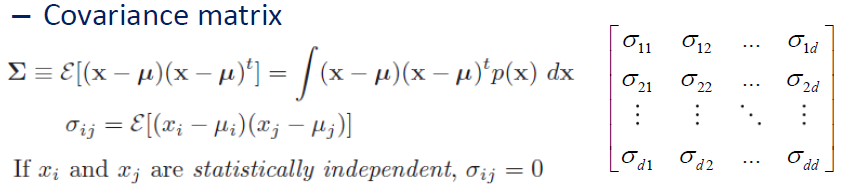
**6、高斯概率密度**（考点!!!）

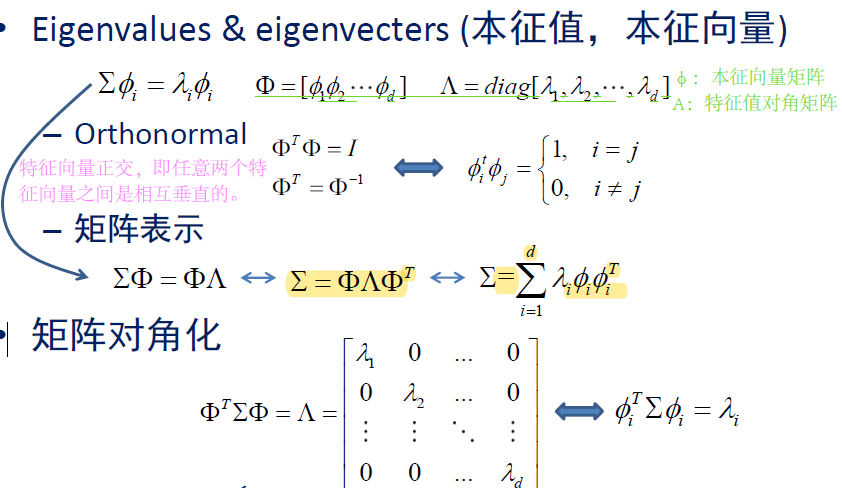
• 选择高斯密度函数的原因：熵最大；大量独立随机变量之和趋近正态分布；符合实际数据。



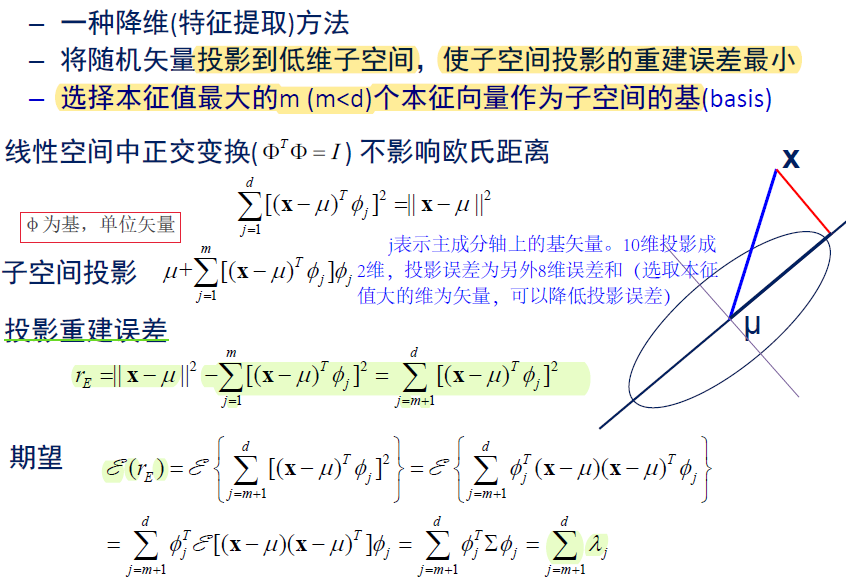
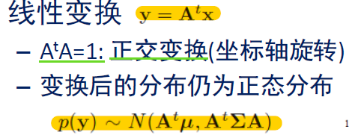
* 多元正态分布：（∑表示d×d的协方差矩阵）





* **协方差矩阵的性质：**

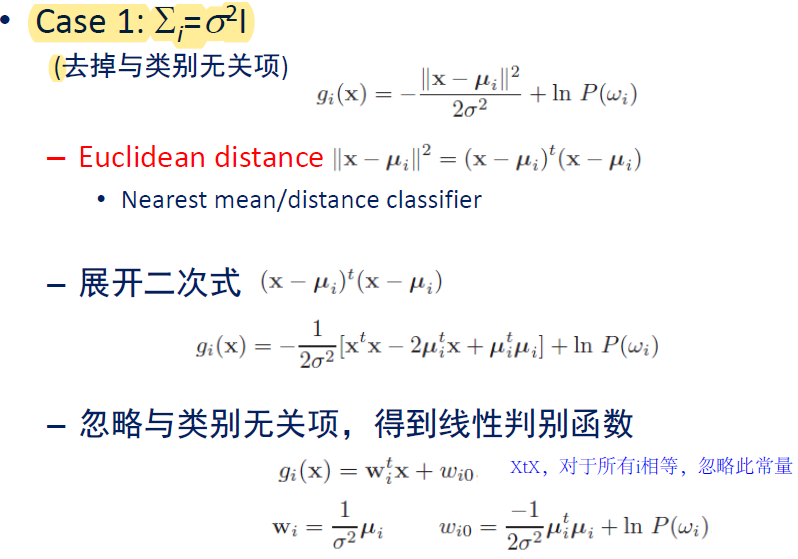
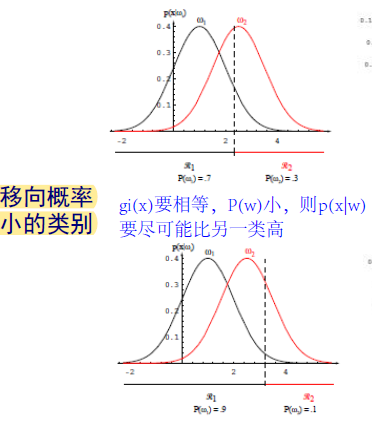
**▶应用：主成分分析（PCA）**



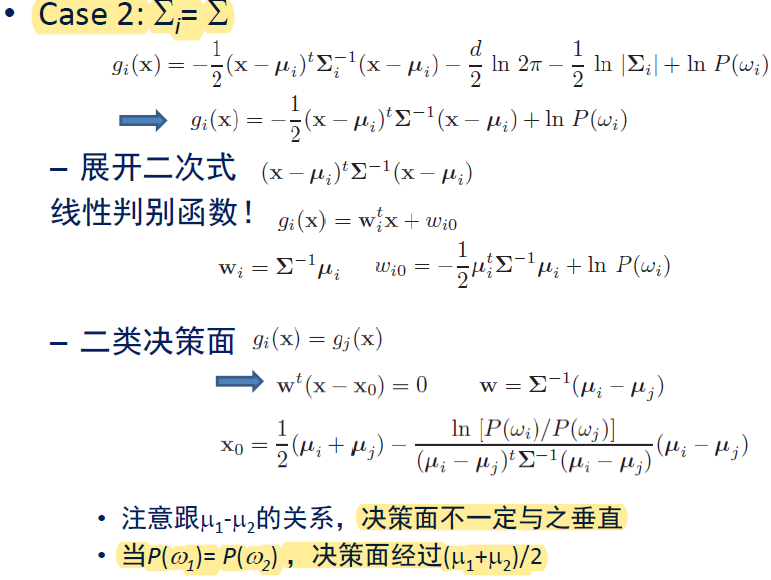
* **线性变换的高斯分布：**

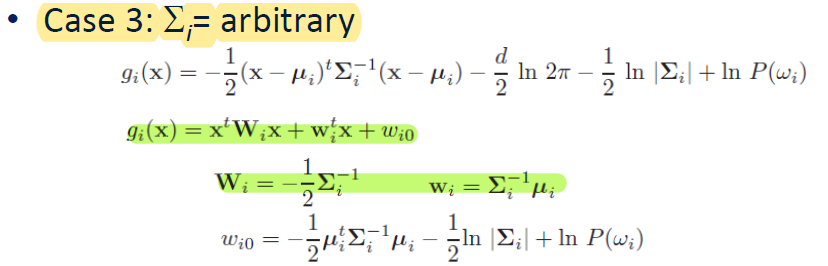
**7、高斯密度下的判别函数**

判别函数：

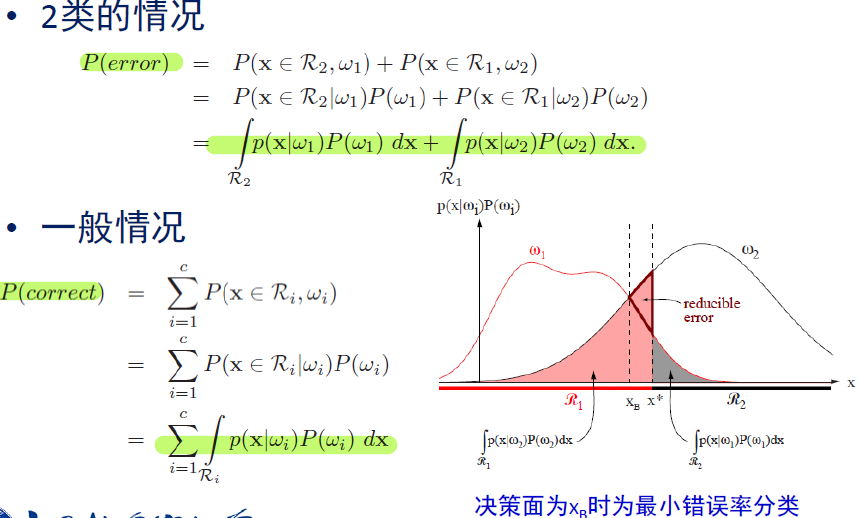
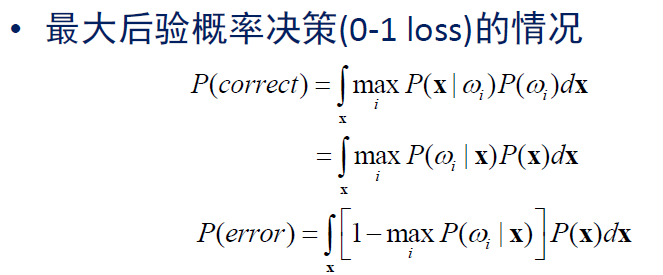




（习题见Question6）

（习题见考试1（2））

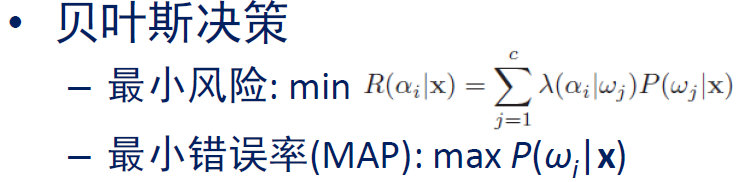
**8、贝叶斯决策的分类错误率（PDF连续的情况）**



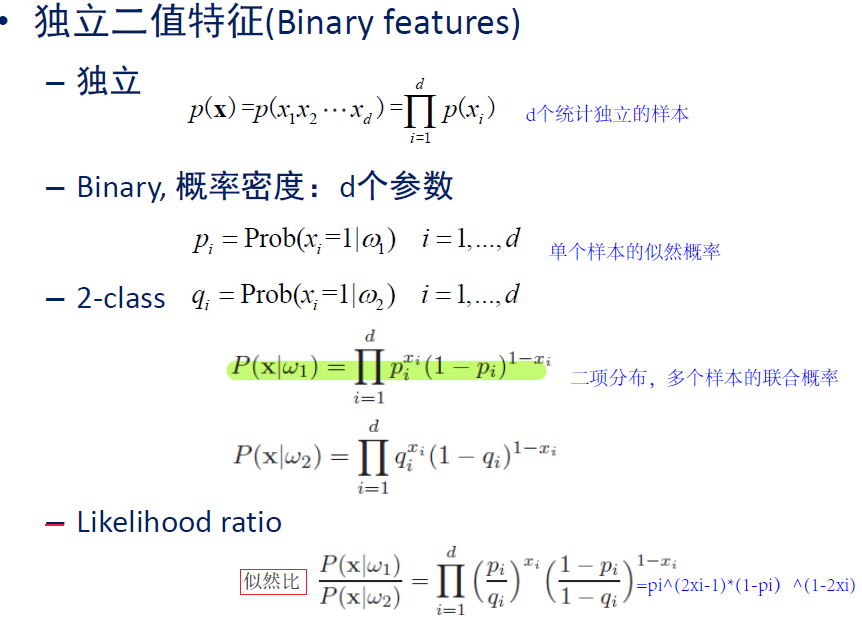
# 第三章 贝叶斯决策和参数估计

## 3.1 贝叶斯决策

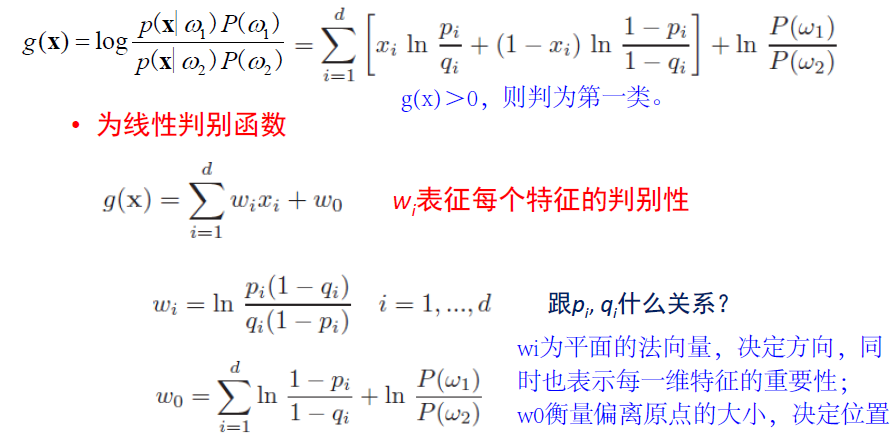
## 3.1.1 离散变量的贝叶斯决策



两分类的情况下，d个样本的联合概率分布：



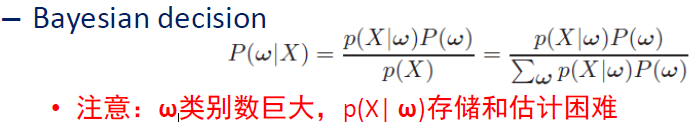
**判别函数**：

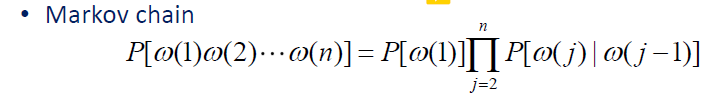


（习题见作业二Q2 & 期末复习P2右栏上）

## 3.1.2 复合模式分类

1.





2. 与复合模式识别类似的问题：**多分类器融合**

多个分类器的决策当作多维特征。k个分类器，共c类，c^k级复杂度，需要大量样本。

实现：1）各分类器相互独立（计算量少，但忽略了相关性）；

2）各分类器之间有关联，运算量略大，但能接受。

## 3.2 参数估计

## 3.2.1 最大似然参数估计

1. 分类器设计：给定分类器函数形式，如：，通过样本估计生成器参数。

2. 统计生成模型的参数估计：

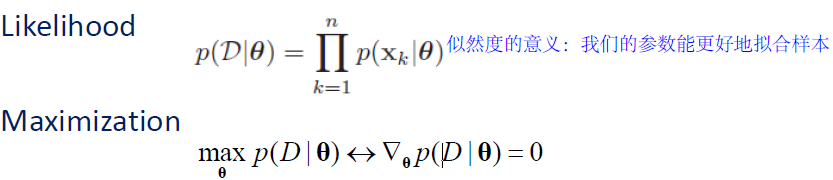
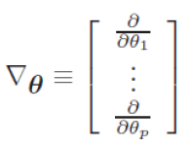
• 最大似然估计（Maximum likelihood ，ML）假设**参数为确定值**，最优估计：似然度最大；

• 贝叶斯估计（Bayesian estimation (Bayesian learning)）假设**参数为随机变量**，估计其分布。

**3. 最大似然估计**

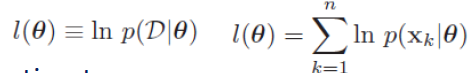
• 基本原理：c个样本独立同分布(i.i.d)，各自单独决策

**似然函数：**



注意：参数空间（p维）和特征空间（d维）不同

求导得不到解析解时，需要迭代求解（如梯度下降）

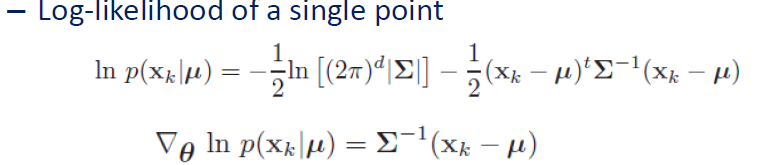


**Log-似然函数：**



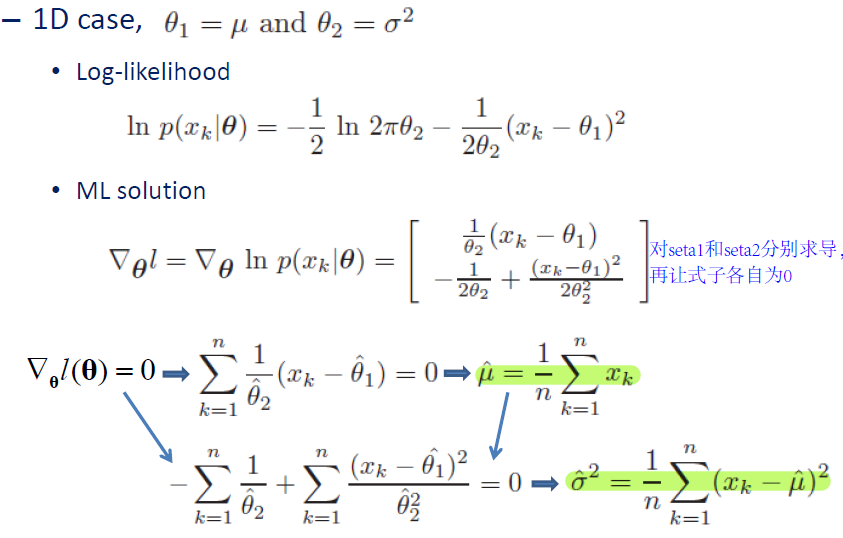
最大后验概率估计(MAP)：

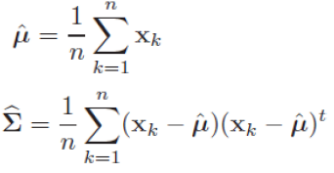
▶ **Case 1:均值未知**



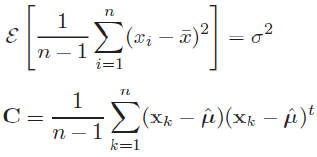
▶ **Case 2:均值和方差都未知**

1. 一维情况：

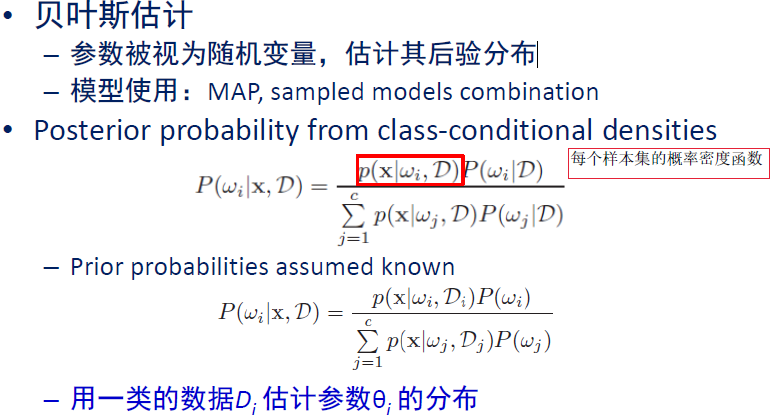


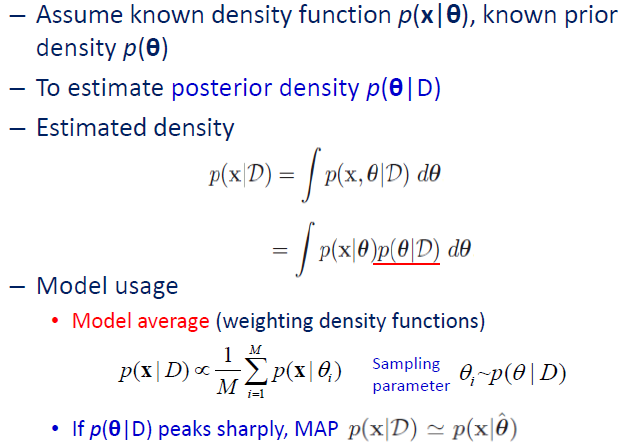
1. 多维情况：

有偏估计的情况：



## 3.2.2 贝叶斯估计



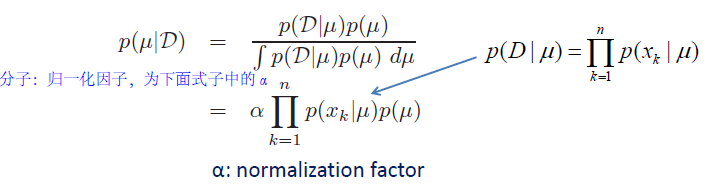
参数分布：

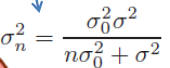
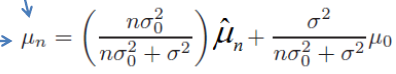
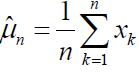
## 3.2.2.1 贝叶斯估计：高斯密度的情况

1.假设x和u都服从一维高斯，假设x的方差已知，估计x的均值，

，

得到后验概率分布如下：

结论：计算结果仍为高斯分布





## 3.2.2.1 贝叶斯估计：一般情况

1. 步骤同于高斯的情况。

2. 递归(增量)的贝叶斯学习，后一个样本在前一次估计的基础上进行更新。

## Summary

1. **Maximum-likelihood (ML) vs Bayesian estimation (BL)：**

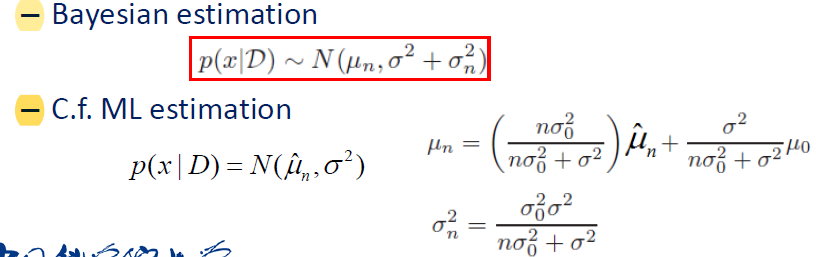
• 当样本个数→n时，ML和BL相等；

* ML：计算简单
* BL：合并先验知识（有时非常有用的），理论上是增量的，从而给参数带来不确定性。

1. **BL下的多变量估计：**

* 通常假设参数的高斯先验和后验；
* 非参数贝叶斯学习；
* 计算中的许多问题采用近似方法。

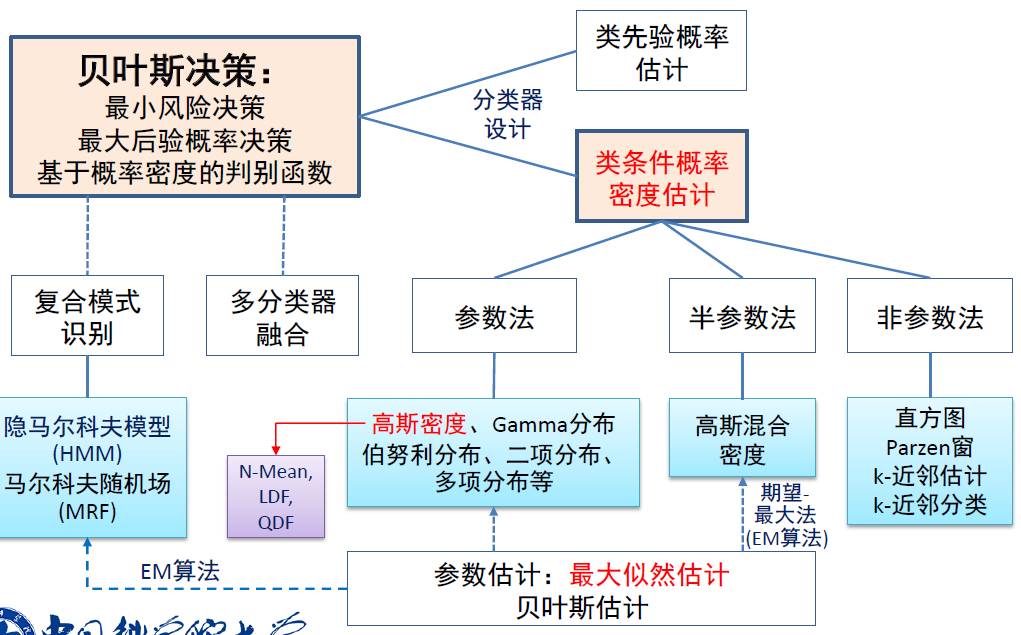
1. **类条件概率密度：**



分析：由于BL中参数是随机变量，所以最终类条件概率一部分受参数的均值和方差影响。

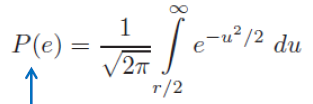
# 第三章 参数估计（续）

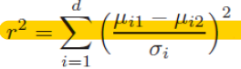
基于**贝叶斯决策**的**模式分类框架**：



## 3. 1 特征维度问题

1. 增加特征可以增大类间差异性，有助于分类，但会引发计算、存储和泛化性能等问题。

2. 分类错误率与特征的关系：



当协方差矩阵是对角阵时：

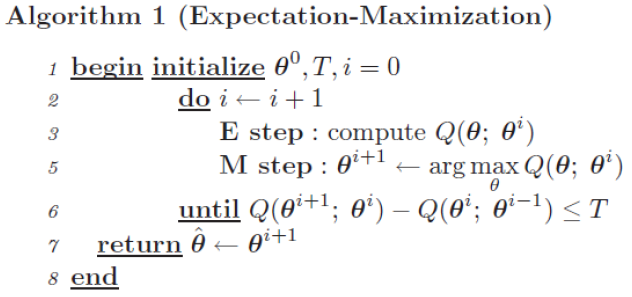
二类均值的距离反映区分度，决定错误率：**特征增加(高维空间，r2增大)有助于减小错误率。**

1. 计算复杂度：参数存储复杂度~O(d2)，分类复杂度~O(d3)，d为特征维数。
2. 过拟合(Overfitting)：特征维数高、训练样本少导致模型参数估计不准确
3. 克服方法：特征降维：特征提取(变换)、特征选择；参数共享

## 3.2 期望最大法（EM）（半参数法）

数据缺失情况下的参数估计：good，bad

EM算法保证good数据的对数似然性单调增加。

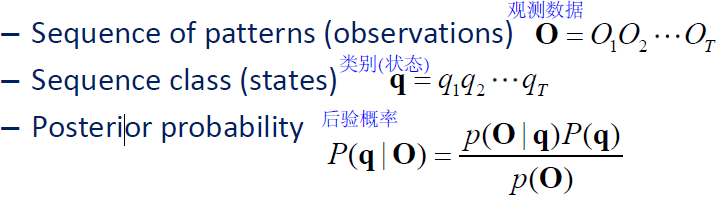


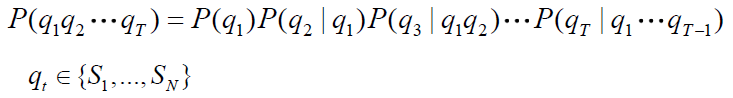
写出以上期望的表达式(E-step)，分别对未知参数求导，导数为0处即为期望最大值点(M-step)，更新得到新参数。

(习题见作业二，Q5)

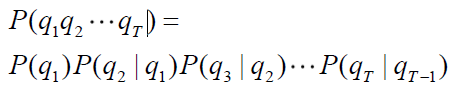
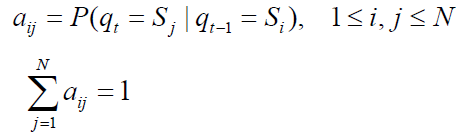
## 3.3 隐马尔科夫模型(Hidden Markov Model ,HMM)

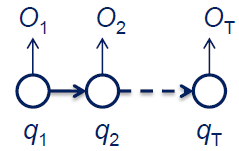
1. 贝叶斯决策下的分类:



2. 马尔科夫链：

• 序列状态(分类)：

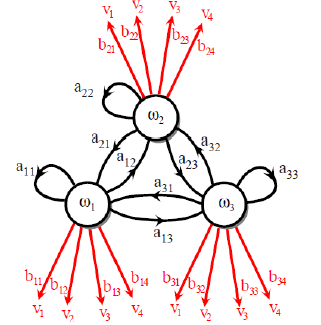
* 一阶马尔科夫链：
* 状态转换概率：



3. HMM模型：

• 实例：假设你在屋子里猜测外面的天气，观测：温度、湿度；隐藏的状态：天气。

• 从状态中估计观测，双重嵌入的随机过程。

N种状态：

M个观测数据：

状态转移矩阵（概率）：

t时刻我们观测到的vk，内部的状态为qj，可见状态矩阵（概率）：



初始状态分布：

由此HMM描述为：

1. HMM的三个核心问题：

•**估值问题：**已有HMM，aij和bjk已知，计算这个模型产生一个特定观测序列VT的概率；

•**解码问题**：已有HMM，决定最有可能产生观测序列的隐状态序列wT；

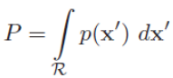
• **学习问题**：只知道HMM的大致结构，aij和bjk未知，如何从一组可见符号的训练序列中，决定这些参数。

运用前向-后向算法：广义EM算法，核心思想：通过递归方式更新权重，得到能够更好地解释训练样本序列的模型参数。

（见第二次作业Q6，复杂度的计算）

# 第四章 非参数估计

## 4. 1 密度估计

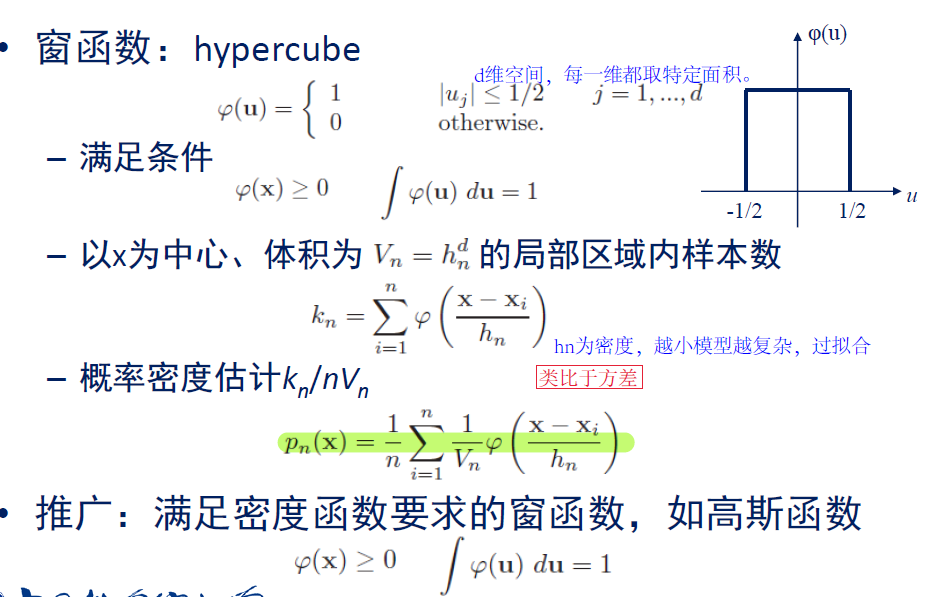
1. 概率密度函数的定义：局部区域体积为V，样本数为k，总样本数为n，

，一定区域内的概率：

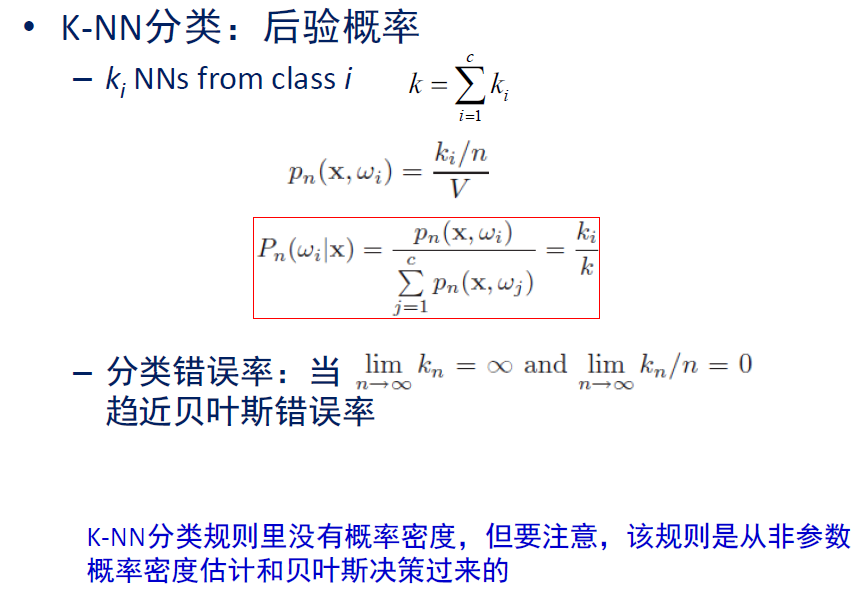
两种非参数方法的**区别**：



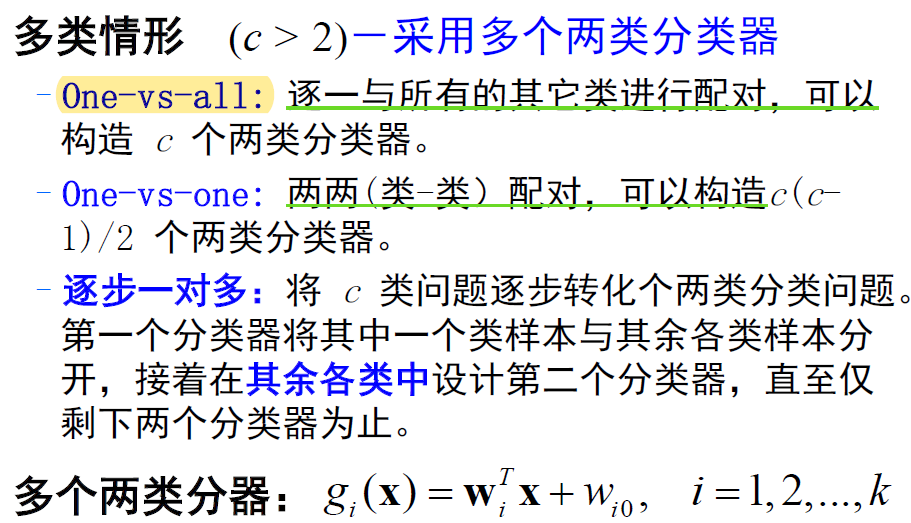
## 4. 2 Parzen窗方法密度估计

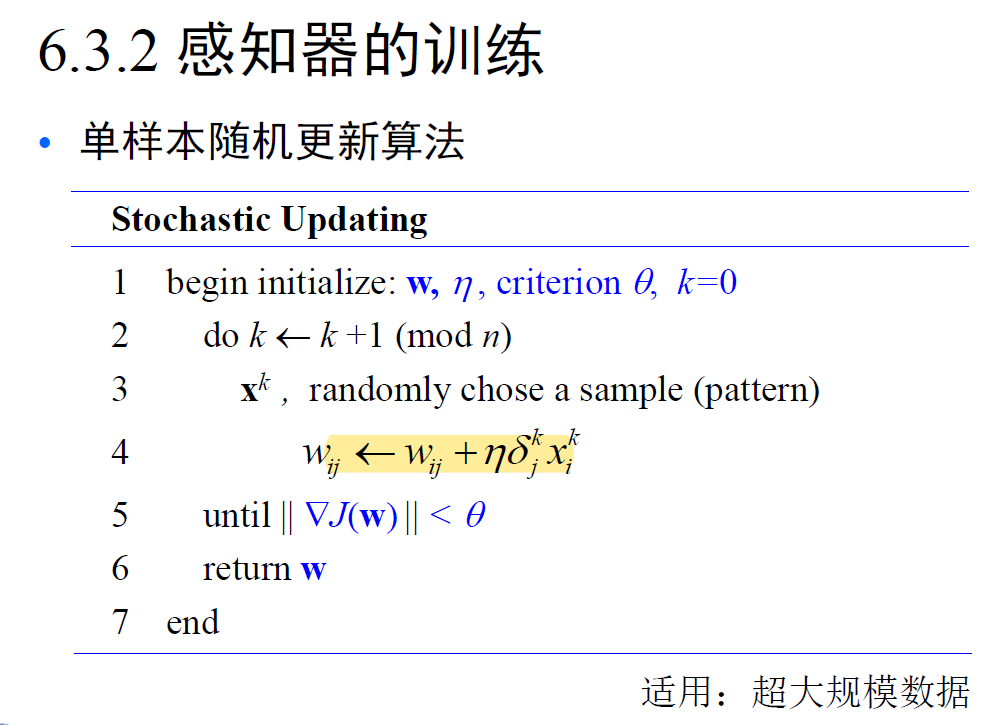


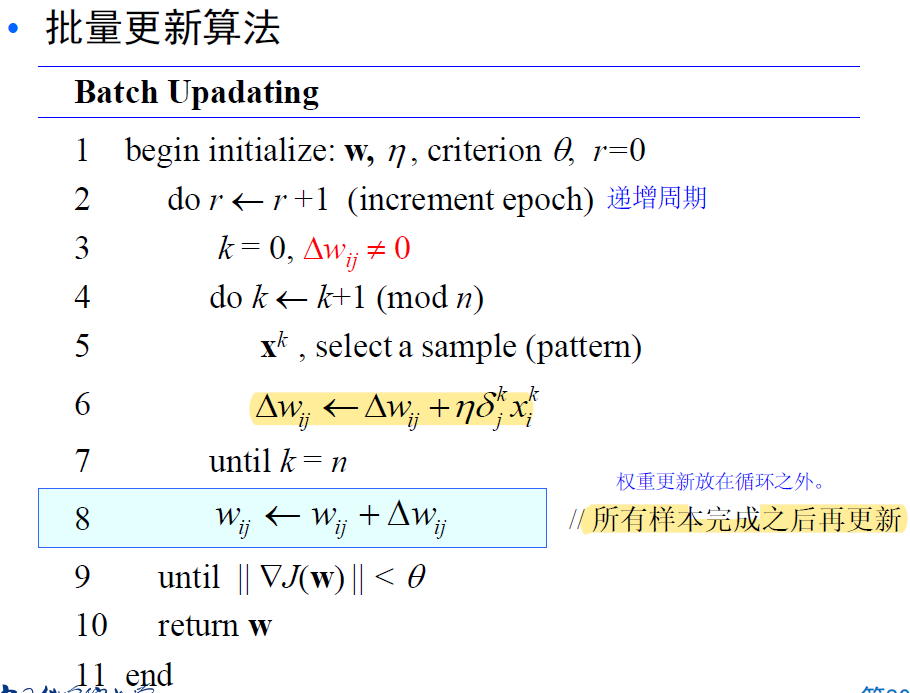
## 4. 3 K近邻估计



# 第6 章 线性判别函数



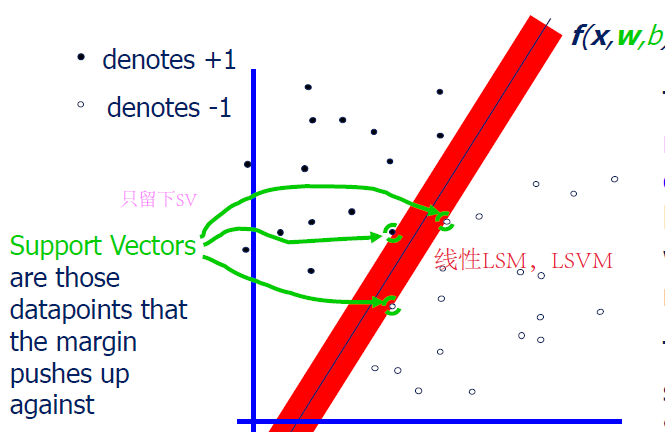


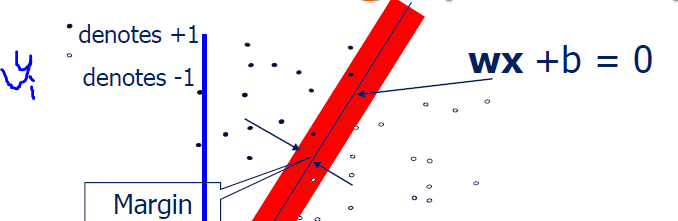


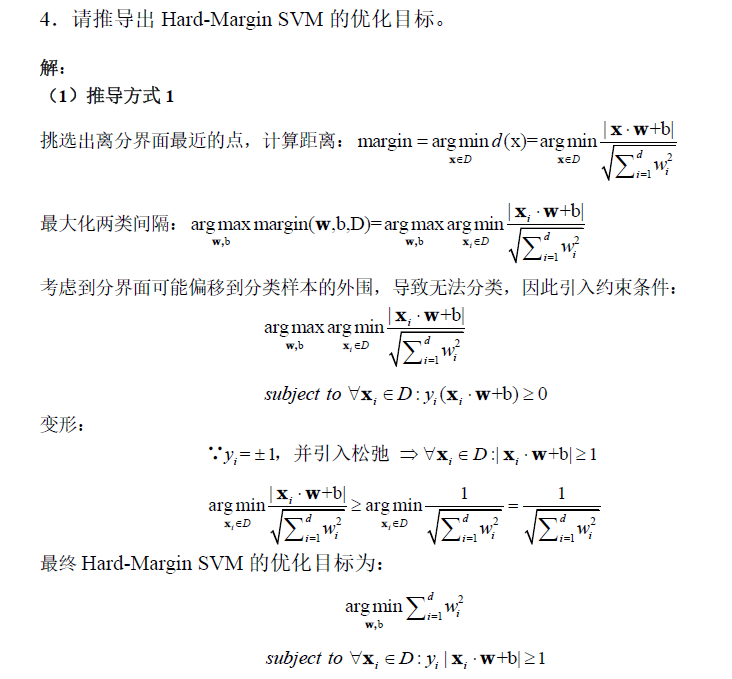
# 第13-14章 SVM

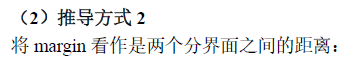
**一、Hard-Margin SVM**

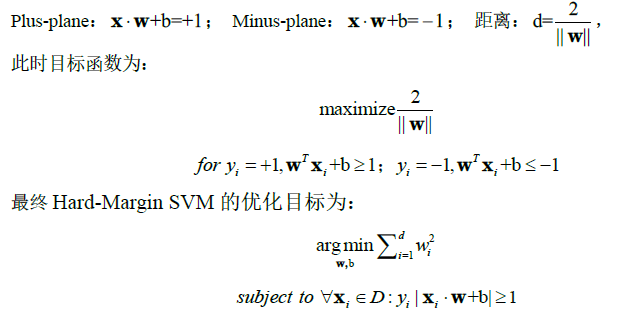
1. 将线性分类器的边距（margin）定义为：到达数据点之前可以增加的边界宽度。



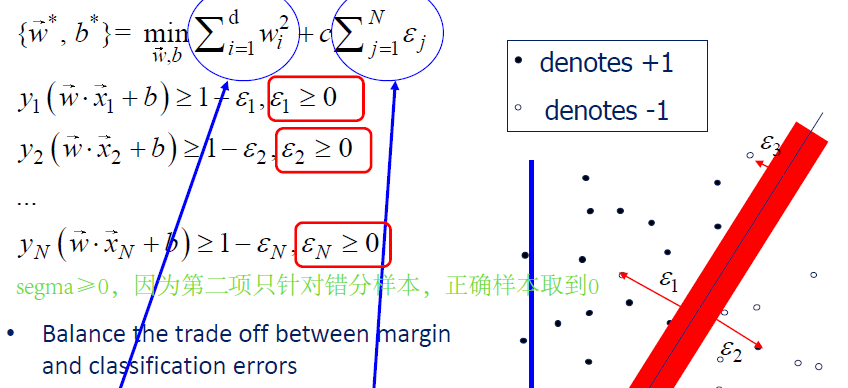
优势：允许样本点的轻微扰动，而不会造成分类错误。

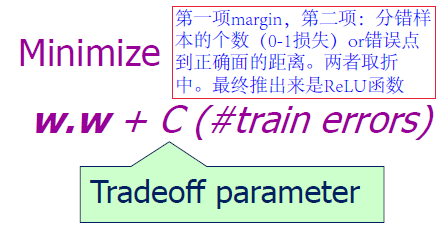


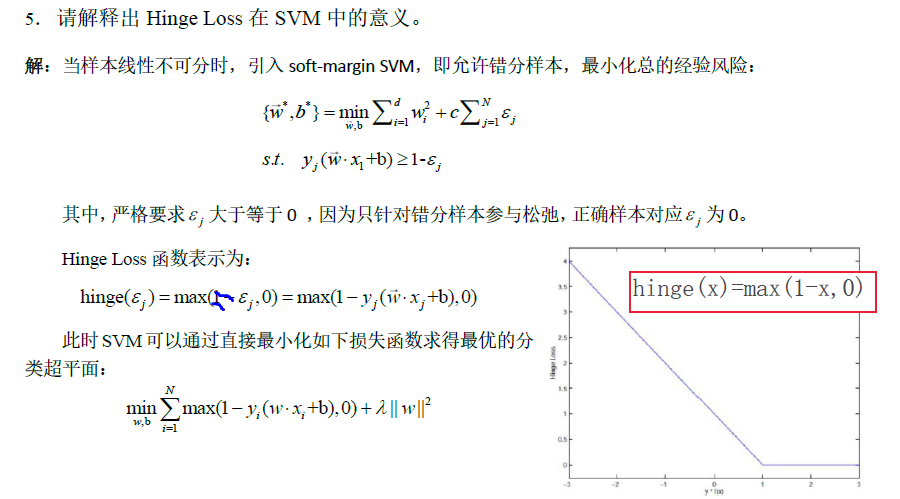


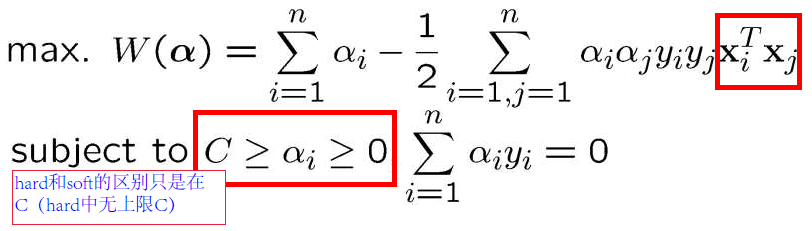


**二、Soft-Margin SVM**

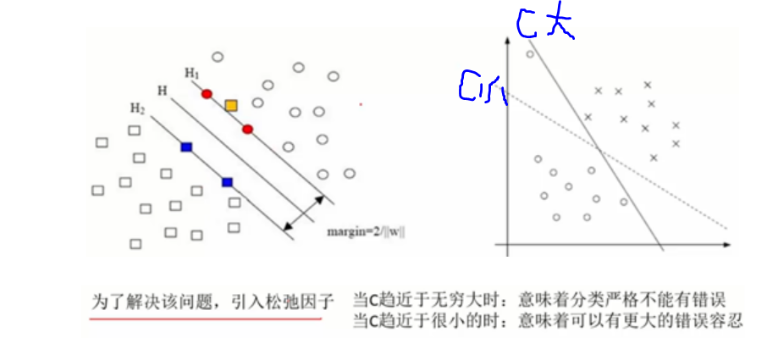
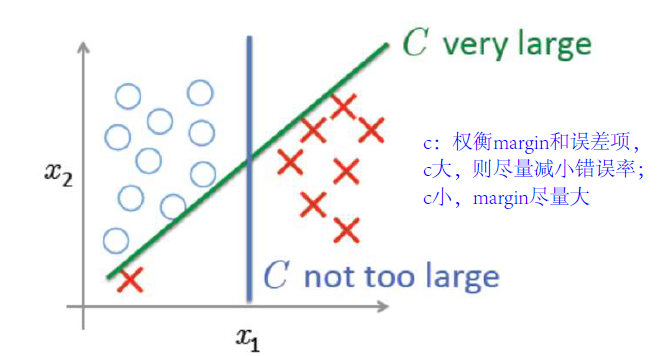
 对于不能线性可分的情况，思路：最小化训练误差。

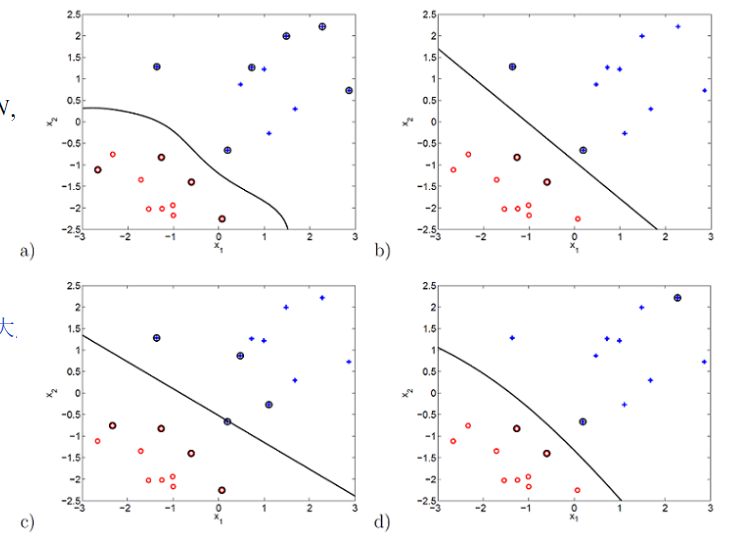
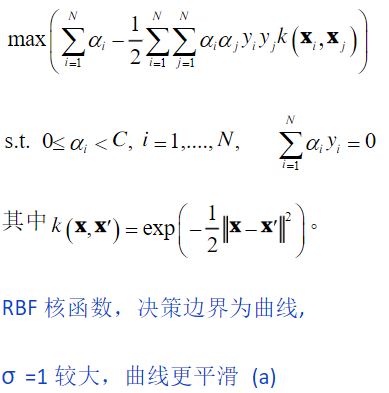


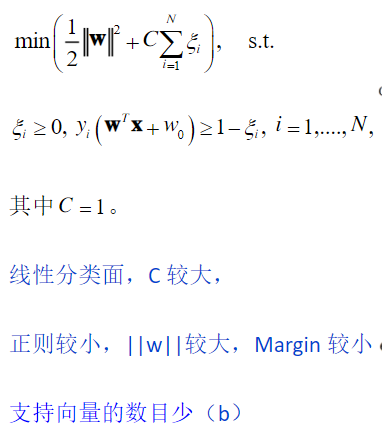
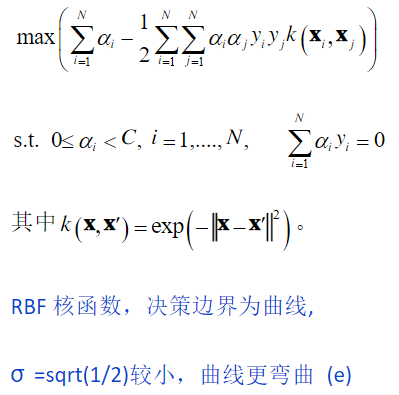
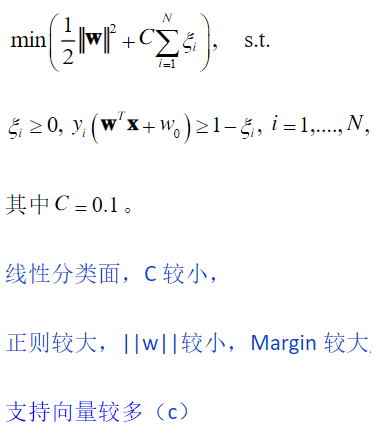
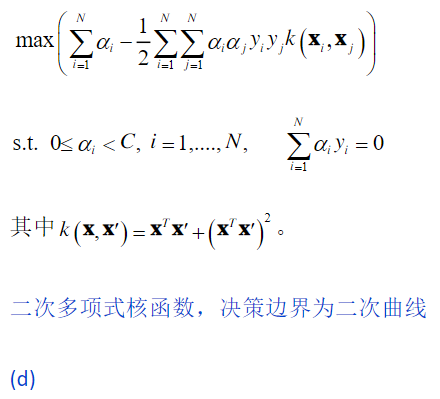
soft-margin SVM**对偶问题：**

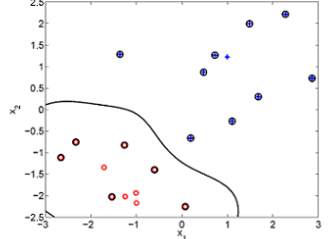


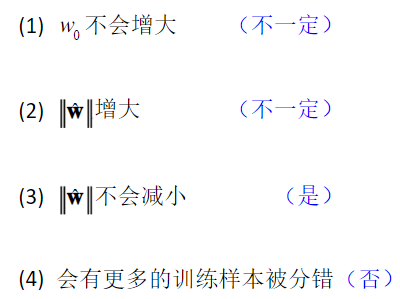
（1）惩罚因子C为无穷大，则SVM退化为硬间隔分类器，此时使用线性核只能处理线性可分的样本（因为对于线性不可分的样本，无法找到一个超平面可以正确划分所有样本，于是模型参数不存在可行解）；

（2）惩罚因子C为适当值，则使用线性核的SVM可以处理线性或**近线性**的样本（这是因为此时允许分类器在某些样本上出错，所以对于近线性的样本，可以找到一个超平面将大部分样本正确划分）。



  
惩罚参数C增大，以下变量的变化：



**3、 Kernel Methods**

原始数据的输入空间始终可以映射到某些高维特征空间中。

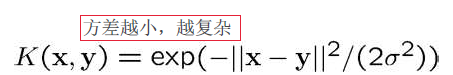
**核方法的基本思想（如何将线性模型转化为非线性）？？？**

 低维空间中线性不可分的点，通过转换为高维空间中的点，很有可能变为线性可分的。输入数据以内积（或其它核函数）的形式出现，而无需进行显式的投影。

1. 多项式核：阶数越高，图像越弯曲，越低，越平滑



（2）RBF核（高斯核）：



（3）Sigmoid核函数：

   
总结：

**SVM的步骤**：

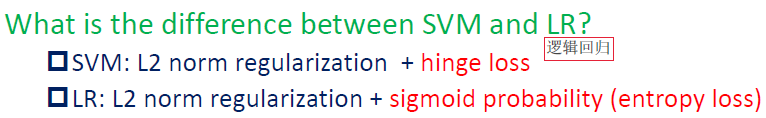
1）准备数据矩阵 2）选择核函数； 3）设置惩罚因子C的大小；

4）训练得到参数的值： 5）通过α和支持向量能对新数据实现分类。

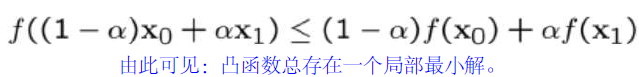
**优势：**1）能找到全局最优解； 2）由于最大化margin的约束，不会引发维度灾难；

3）很难过拟合； 4）几何意义，解释简单

**缺陷**：1）训练和测试慢； 2）本质上是二进制分类器； 3）核选择的不确定性



**4、 模型的选择：**

1.凸函数：

2. SVM实质是一个凸函数，解决的方法：

（1）“内点”法

–从可能违反约束的初始解决方案开始，通过优化目标函数和/或减少约束不满足的数量来改进此方案。

（2）常用的是：顺序最小优化**（SMO）**

-每次只解两个α（两层循环），轮替优化，直至收敛

3.加快SVM训练的方法：

-组块策略：保留支撑向量，非支撑向量放一边和新数据再训练

4. SVM的多类扩展：样本不均衡不算是问题，因为＞margin的样本不起作用。

# 第15章 决策树

1. 决策树的优势：可解释性；可处理离散变量。

缺点：极易过拟合。剪枝-提高泛化能力；随机森林

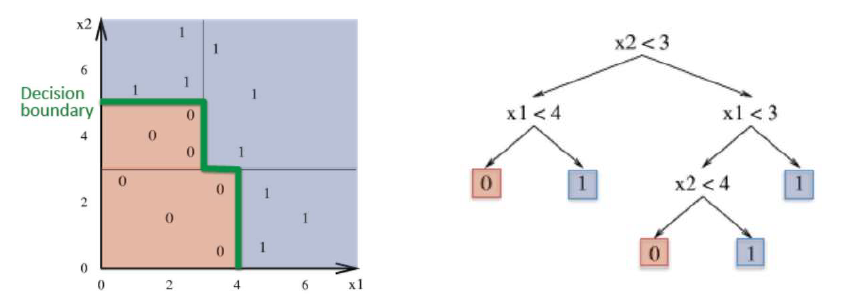
1. 关键：选特征，定阈值
2. 设置合适的特征选择顺序，可以防止树变深，防止过拟合。
3. 如何分割连续性？如：年龄、收入等
   1. 离散化以形成序数分类属性；

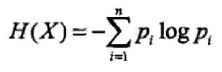
静态-一开始就离散化；动态：通过等间隔时段，等频率时段，百分数，聚类等作为类别

* 1. 二值决策（＞or＜A）

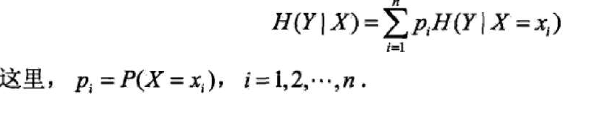
考虑所有可能的分割并找到最佳分割；通常是计算密集型。

1. 决策树的特性：每个内部节点代表许多选择中的某个选择；每个叶子节点代表一个分类或决策。
2. 二值化的决策树：将特征空间划分为轴平行（超）矩形



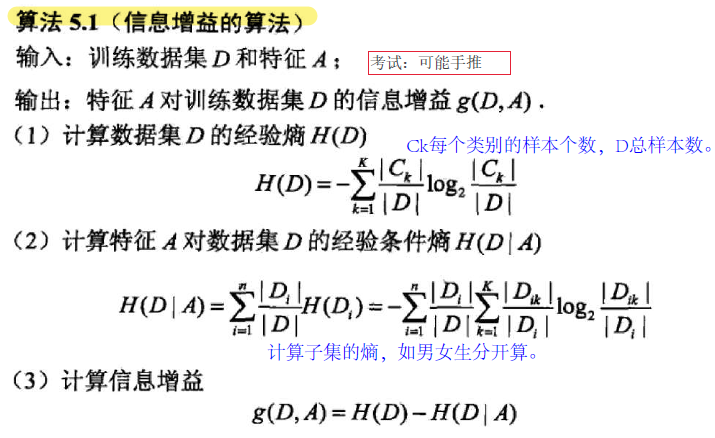
1. **熵**的计算公式：

**条件熵**：随机变量X给定的条件下随机变量Y 的不确定性：



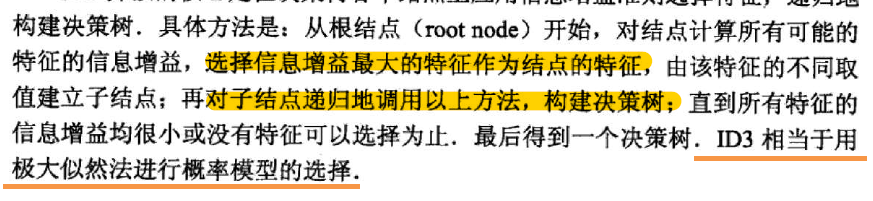
**信息增益（互信息）**：得知X信息后使得Y信息不确定性减少的程度。互信息=熵-条件熵

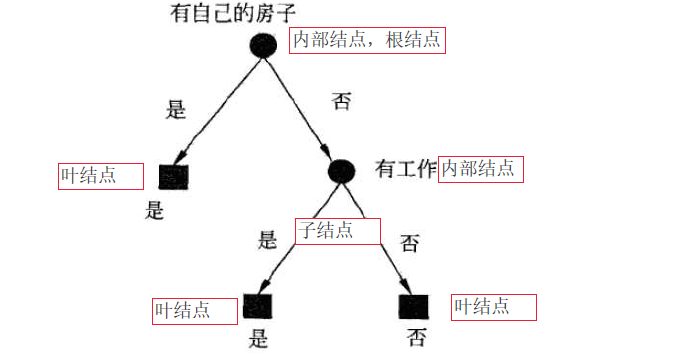
决策树中的信息增益等价于，训练数据集中**类**与**特征**的互信息。



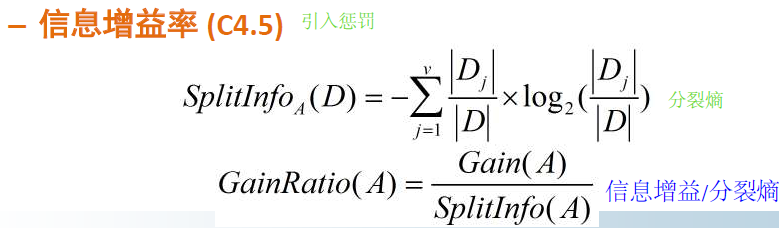
公式（2）：计算特征中子集的熵，单独看第i个特征，其中共有k种属性

1. 第一个分数：单个属性在特征中概率；（如青年1/5，中年1/5,老年3/5）
2. 第二个分数：单个属性对最终结果的熵（如：青年中，是1/3,否2/3）
3. 条件熵=1/5\*S1+1/5\*S2+3/5\*S3=1/5\*[-1/3\*log(1/3)-2/3\*log(2/3)]+ 1/5\*S2+3/5\*S3
4. **ID3算法**





1. **C4.5算法**



**ID3算法**选择特征依赖于**信息增益**, 在面对类别较少的离散数据时效果较好;面对连续的数据（如体重、身高、年龄、距离等）, 或者每列数据没有明显的类别之分（如学号,身份证等）,ID3算法会挑选此类特征作为根节点，把每个数据自成一类,泛化能力弱，划分效果极差。

**C4.5算法**依赖于**信息增益率**, 引入了分裂信息来惩罚更多的属性，能减轻划分行为本身的影响。取值数目多的属性分裂信息也会变大，将增益除以分裂信息，再加上一些额外操作，可以有效控制信息增益过大的问题。

区别:

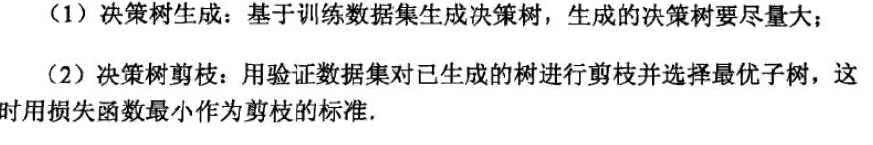
**ID3算法**没有剪枝处理，面对类别较少的离散数据时效果较好;，但处理连续数据或无明显类别之分的数据泛化能力弱，划分效果差。C4.5从以下方面改进了ID3：

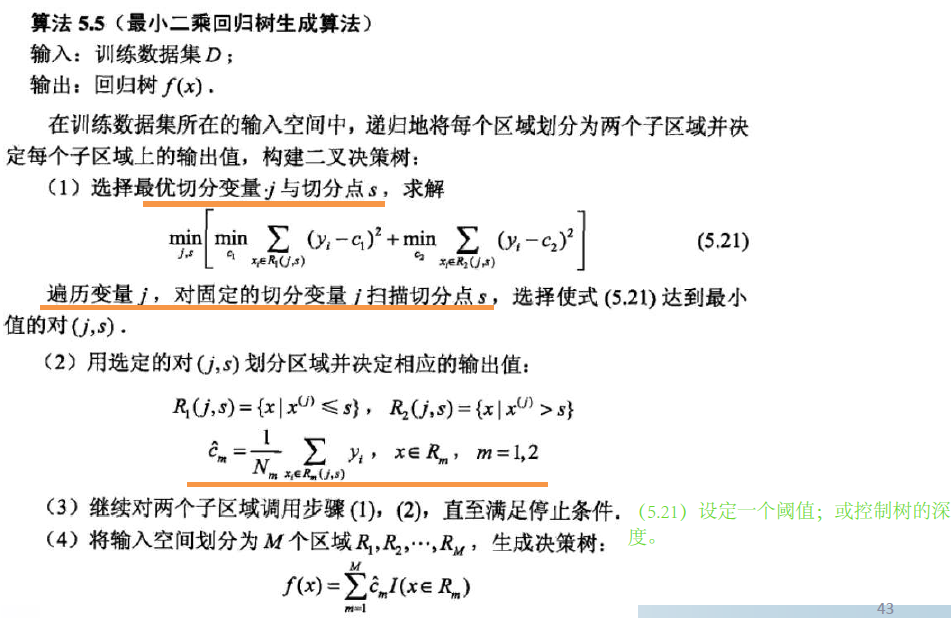
* 1. 用信息增益率选择属性，克服了用信息增益选择属性时偏向选择取值多属性的不足；
  2. 在树构造过程中进行剪枝；
  3. 能够完成对连续属性的离散化处理；
  4. 能够对不完整数据进行处理。

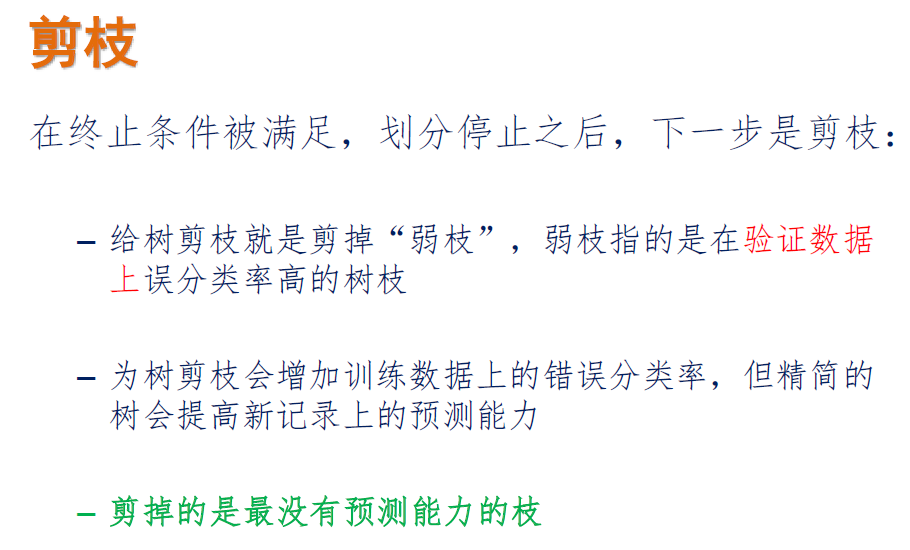
1. 分类与回归树（CART）

ID3和C4.5中没有规定是否为二叉树，CART假设了决策树是二叉树，递归地二分每个特征（是、否），在输入（特征）空间给定的条件下求输出的条件概率分布。

回归树用平方误差最小准则，分类树用基尼指数最小化准则。



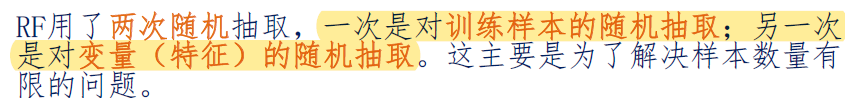




1. 防止决策树过拟合；获取更多数据，or删除不相关的属性（剪枝）。



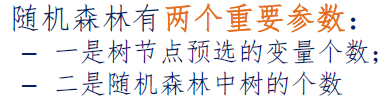
1. 随机森林：用**随机抽取**的方法训练出一群决策树来完成分类任务。

很难再发生过拟合

核心思想：由弱变强。单个决策树只用了部分变量、部分样本训练而成，可能单个的分类准确率并不是很高。但是当一群这样的决策树组合起来分别对输入数据作出判断时，可以带来较高的准确率。

每个决策树独立分类，按**多数原则**投票决定随机森林的最优分类。

回归预测-预测结果取平均。



随机森林是以决策树为基本分类器的一个集成学习模型，它包含多个由Bagging集成学习技术训练得到的决策树。随机森林受以下两个算法的启发：

1）Bootstrap：当样本数量不大，分布情况未知时，可以从原始样本中随机抽取（有放回）的多个样本情况（弱学习）来估计原样本真实的分布情况。

2）Bagging: 用Bootstrap采样选出n个样本, 在所有属性上，对这n个样本建立分类器。重复以上步骤m次，数据在得到的m个分类器上测试。

▶影响随机森林分类性能的因素：单棵树的分类强度（越大，越好）；树之间的相关度（越大，越差，故而增加分类器之间的差异性）

▶随机森林的优势：

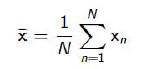
1. 随机筛选特征，方便处理高维数据；
2. 容易进行分布式处理，实现高效；
3. 可以利用out of bag的数据评价算法的误差率，而无需验证集；
4. 能计算每个特征的重要性；
5. 受噪声干扰小，鲁棒性好

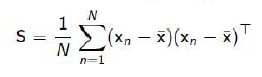
# 第16章 特征提取与选择

1. 手工设计的特征适合：小样本学习；连续学习，能够不间断地添加新数据。
2. 特征生成：SIFT（尺度不变特征变换）：生成的特征不受尺度、旋转等变换影响。
3. **线性维度削减：**

▶PCA（主成分分析）: 使原始数据在投影子空间中各个维度的方差最大（更多地保留数据间的差异性）。将N维的数据投影到K维的空间上（N<K）。首先求出这N维数据的协方差矩阵，然后求出其前K个最大特征值所对应的特征向量，这M个特征向量即为所求的投影空间的基。

能捕获到最大向量的方向；重构误差最小的方法。

Step1：计算数据的均值：



Step2：计算协方差矩阵：

Step3：N×N矩阵进行特征值分解

Step4：选择特征值最大的前K个特征向量；



Step5：把每个样本投影到主成分上（特征向量的线性组合）：

经过投影后得到的数据Zn是K×1的向量。

▶CCA（典型相关分析）:

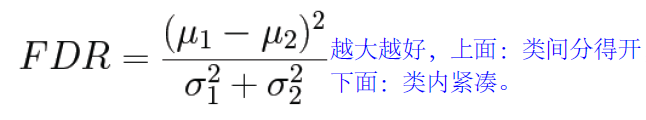
两个多维变量找线性相关，找到两个变量在变换空间中的最大相关性。

在分类问题中：特征向量与类别标签（物体的不同描述方式）

▶LDA（线性判别分析）: 投影后类内方差最小，类间方差最大。

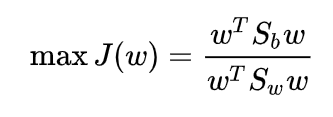
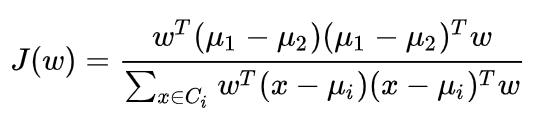
有监督的降维算法，数据是有标签的。

Fisher 判别比：

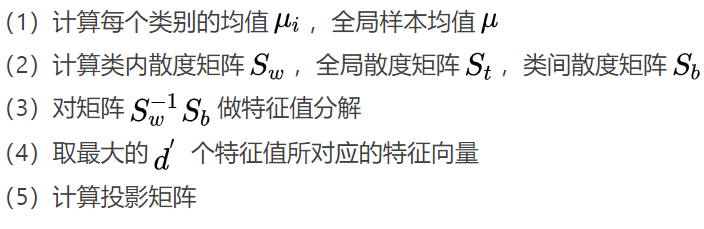
 

FDR=类间协方差矩阵Sb（样本均值之间）/类内协方差矩阵Sw（某类中各样本之间）

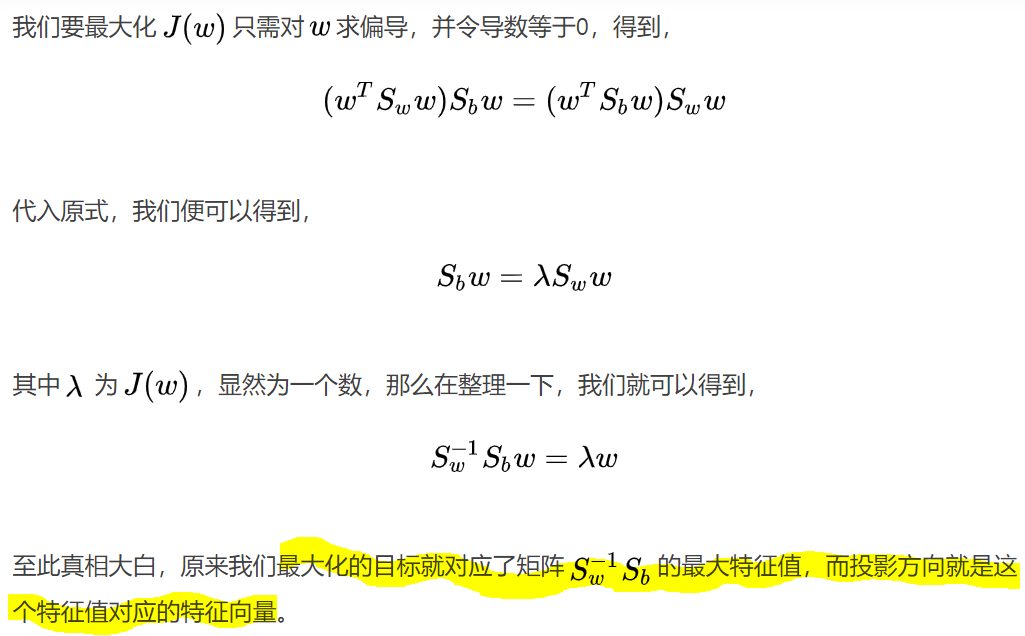
**两类问题的LDA优化目标函数：**

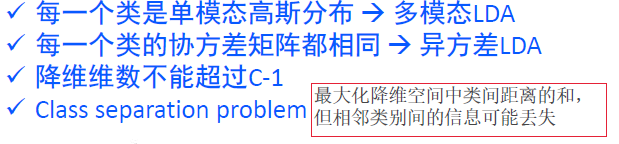


**算法步骤：**

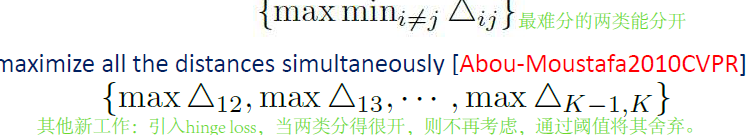


结果的推导过程：





**LDA的缺陷及其改进**：

针对第4个问题的改进：

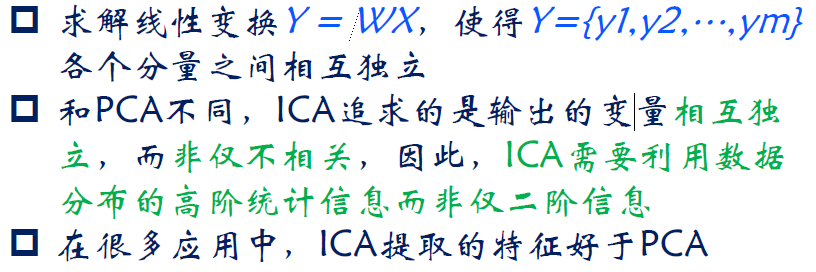
**PCA和LDA的相同点：**

1、都是降维算法；2、假设数据服从高斯分布； 3、都利用了矩阵特征值分解的思想。

**不同点：**

1.PCA无监督（样本无标签），LDA有监督； 2. 前者去掉原始数据冗余的维度，希望投影后数据方差尽可能大（最大可分性），；后者是选择最佳投影方向，使得类内方差小（紧凑），类间方差大（分离）； 3.LDA最多可以降维到C-1； 4. LDA可能过拟合数据

▶ICA（独立成分分析）:



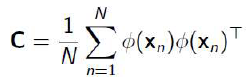
信号高斯分布-PCA；只有**非高斯的信号才能用ICA**

1. **非线性扩展：**

非线性的方法保留了空间邻域信息（局部线性→全局非线性）

▶Kernel PCA（非线性）

Step1：将输入D维的数据利用非线性函数转换到M维空间

Step2：计算M×M的协方差矩阵：

Step3：计算特征向量：

流形学习：

▶（1）邻域线性嵌入（LLE）

高低维的某点用相同重构系数和邻居得以重构。高维空间是朋友，那么降低到低维还是朋友

缺点：很难扩充到新样本，只能批量处理

▶等距特征映射Isometric Feature Mapping（(ISOMAP)）

衡量数据之间的距离不再是欧式，而是用图论的知识，沿着数据的流形找最短距离。

构造D矩阵,使高维空间的测定距离，和低维的欧式距离是相关的。

▶局部保留投影 Locality Preserving Projection（LPP）

定义了离得近的两个点之间的连接。LPP是非线性方法的线性近似（特别是拉普拉斯特征图），它考虑了局部性。（计算复杂度低）

1. **特征选择：**

特征降维：所有的特征变换（连续）；特征选择：只保留部分（离散）

衡量特征好坏的指标：信息熵；距离度量；连续、依赖性、正确性度量。

特征选择的模型：

• Filter model：

–将特征选择与分类器学习分开

–依靠数据的一般特征（信息，距离，依赖性，一致性）

–快速学习任何学习算法

• Wrapper model：

–依靠预定的分类算法

–使用预测准确性作为善度度量

–高精度，计算量大

• Embedded model：

在学习过程中进行特征选择