1. **贝叶斯理论**
   1. **引言**

模式识别的分类问题是根据识别对象特征的观察值将其分到某个类别中去。统计决策理论是处理模式分类问题的基本理论之一, 它对模式分析和分类器的设计有着实际的指导意义。贝叶斯( Bayes) 决策理论方法是统计模式识别中的一个基本方法, 用这个方法进行分类时要求:

( 1) 各类别总体的概率分布是已知的;

( 2) 要决策分类的类别数是一定的。

在连续情况下, 假设对要识别的物理对象有d 种特征观察量x 1 , x 2 , ⋯, xd , 这些特征的所有可能的取值范围构成了d 维特征空间, 我们称x= [ x1 , x 2 , ⋯, x d ] T 为d 维特征向量。这里T 是转置符号。

这些假设说明了要研究的分类问题有c 个类别, 各类别状态用ωi 来表示, i= 1, 2, ⋯, c;对应于各个类别ωi 出现的先验概率 P (ωi)及类条件概率密度函数 p ( x©ω¦ i)是已知的。如果在特征空间已观察到某一向量x , x = [ x 1 , x 2 , ⋯, xd ] T 就是d 维特征空间上的某一个点, 那么应该把x 分到哪一类去才最合理呢? 这就是本章所要研究的主要问题。

* 1. **常用的决策规则**
     1. **基于最小错误率的贝叶斯决策**

在模式分类问题中, 人们往往希望尽量减少分类的错误, 从这样的要求出发, 利用概率论中的贝叶斯公式, 就能得出使错误率为最小的分类规则, 称之为基于最小错误率的贝叶斯决策。

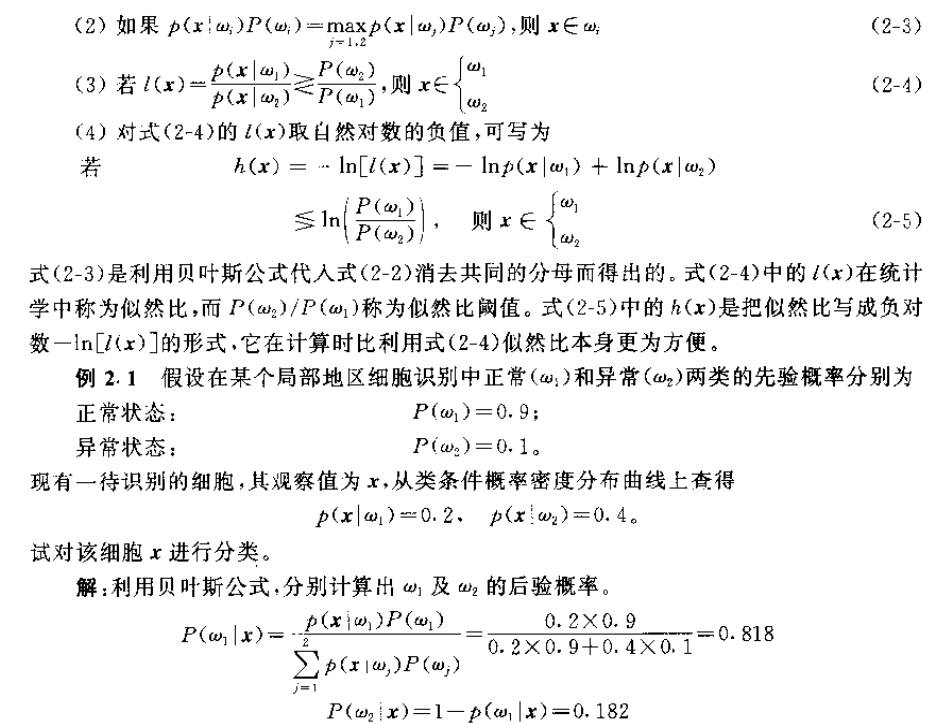
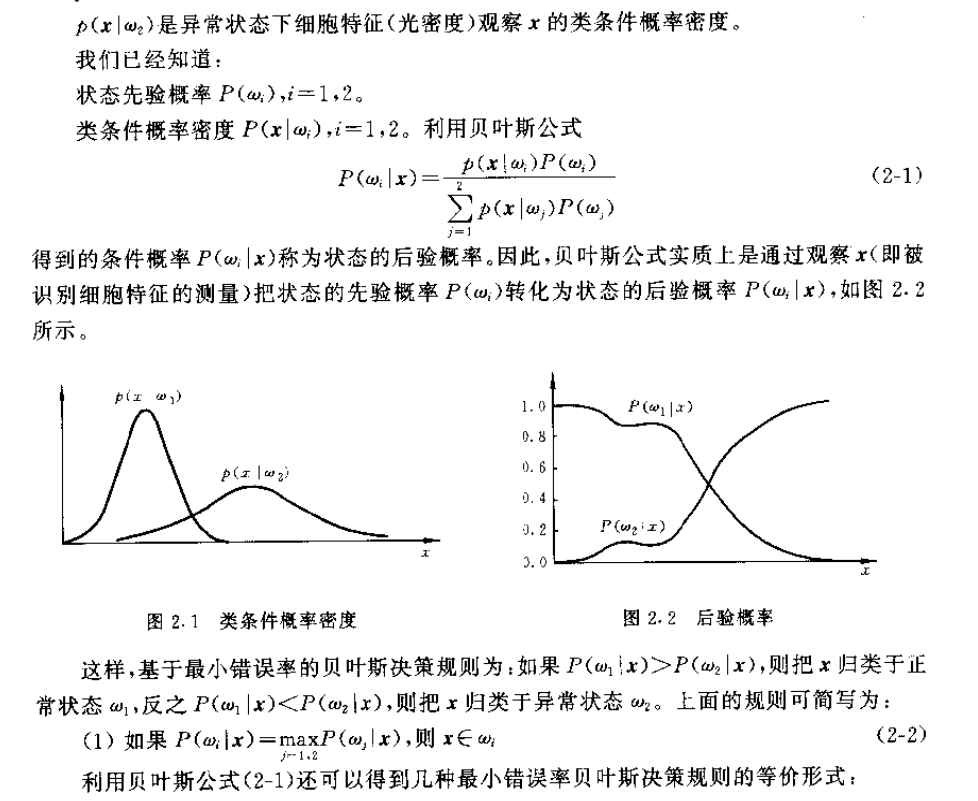
在讨论一般理论之前, 先举一个例子—— 癌细胞的识别—— 来说明解决问题的过程。假设每个要识别的细胞已作过预处理, 抽取出d 个表示细胞基本特性的特征, 成为一个d 维空间的向量x, 识别的目的是要将x 分类为正常细胞或者异常细胞。用决策论的术语来讲就是将x 归类于两种可能的自然状态之一, 如果用ω表示状态, 则

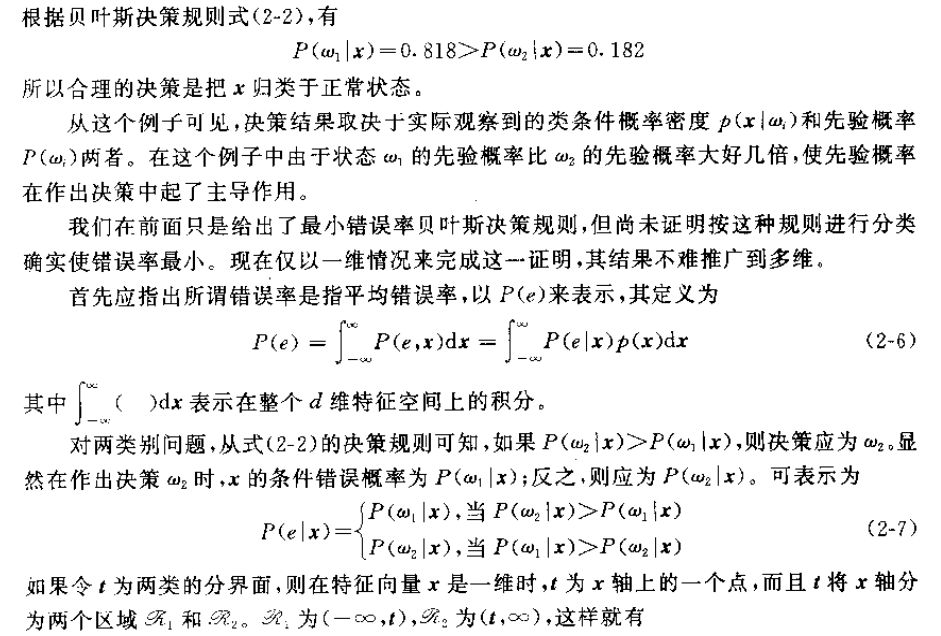
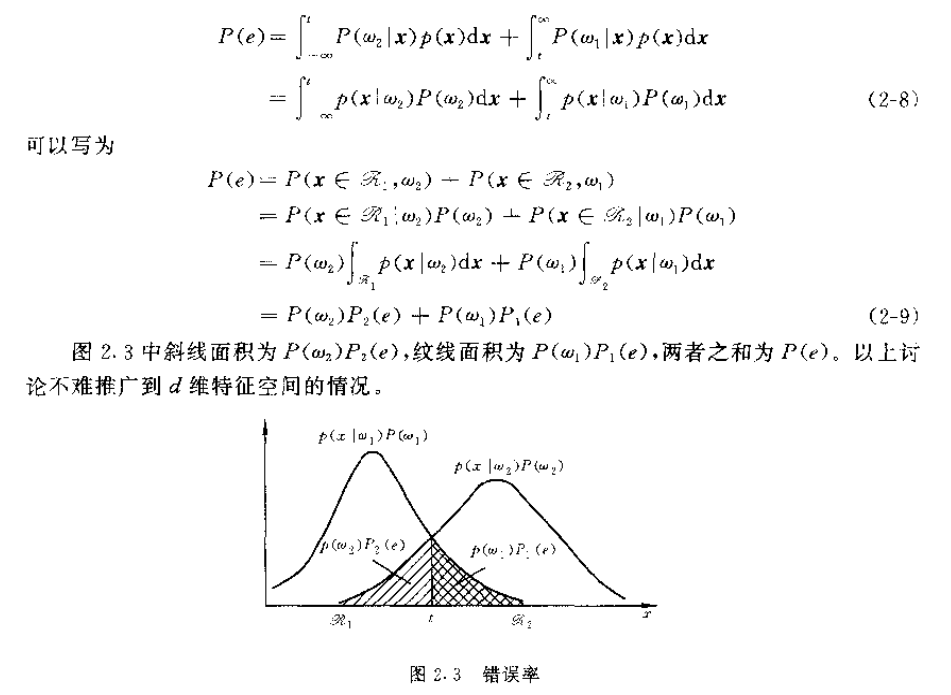
ω= ω1 表示正常

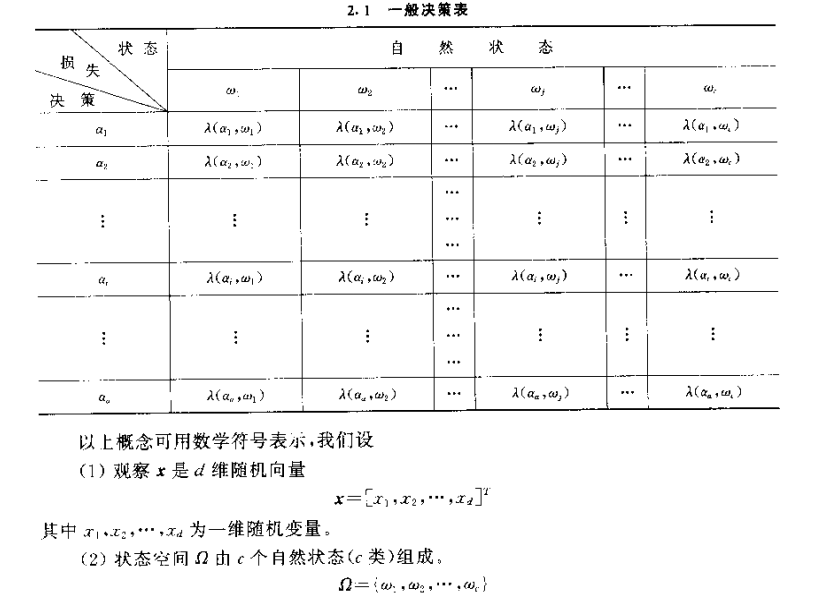
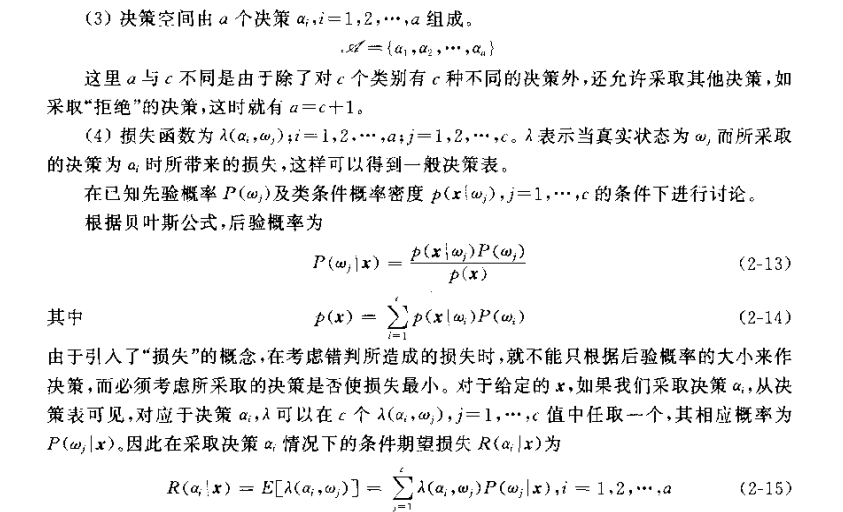
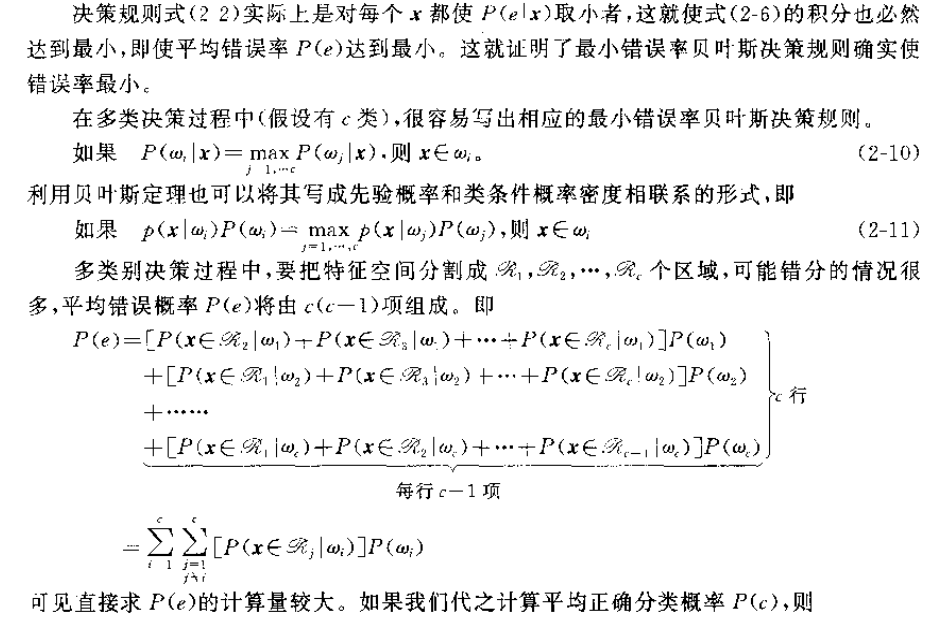
ω= ω2 表示异常

类别的状态是一个随机变量, 而某种状态出现的概率是可以估计的。例如, 根据医院细胞病理检查的大量统计资料可以对某一地区正常细胞和异常细胞出现的比例作出估计, 这就相当于在识别前已知正常状态的概率P ( ω1 ) 和异常状态的概率P ( ω2 ) 。这种由先验知识在识别前就得到的概率P ( ω1 ) 和P ( ω2 ) 称为状态的先验概率。在两类别问题中显然存在P ( ω1 ) + P ( ω2 ) = 1。如果不作细胞特征的仔细观测, 只依靠先验概率P ( ω1 ) , P ( ω2 ) 去作决策, 合理的决策规则就应为: 若P ( ω1 ) > P ( ω2 ) , 则作出ω= ω1 的决策; 反之, 则作出ω= ω2 的决策。在这个例子中由于P ( ω1 ) > P ( ω2 ) , 如果仅按先验概率决策就会把所有细胞都归于正常类, 而根本未达到要把正常细胞与异常细胞区分开来的目的。这是由于先验概率提供的分类信息太少。为此我们还必须利用对细胞作病理分析所观测到的信息, 也就是前面所说的由特征抽取而得到的d 维观测向量。为简单起见, 我们假定只用一个特征( 如胞核总光密度的观察值) 进行分类, 即d = 1。根据本章开始时所做的假设, 在自然状态下观察的类别条件概率分布应为已知, 如图2.1 所示。

p ( x©¦ ω1 ) 是正常状态下细胞特征( 光密度) 观察x 的类条件概率密度。







那么接下来讲这个最大后验概率和最小风险准则作分类判别。这个这一小节有点长，PPT有点多，但是每一页其实数学知识都比较简单，大家应该会看的比较懂一点，所以我们速度稍微快一点，属于较简单的。

从贝叶斯判定上讲，这是上一页的一个PPT，总来上来讲，其实整个模识别如果是从贝叶斯来看就是一个概率的问题，就是猜猜看的问题，尽可能能够猜出问题，使得错误率最小，这是贝叶斯的观点，从概率论上来讲的问题。贝叶斯准则比较简单，如果我们看到的是一个特征，对于一个未知样本来的时候，我们是观测到一个特征，这时候我们要判定的是来的样本属于哪一类？这就跟后验概率关联上了。后验概率转化成先验概率和条件概率，条件概率之后可以看到，每一类的先验概率是多少，然后在每一类的时候出现该特征的情况的概率多少？所以贝叶斯准则其实就是直接用后验概率判定类别的问题，也就是max你的后验概率，把它稍微的展开之后，就是把它变成了一个先验概率和条件概率之间的相乘，如果哪个乘积越大，那么它就属于哪一类，因为分母是一样的，这个属于贝叶斯判别，是比较简单的，就是这个公式，其实本身来讲就是后验概率公式。这个公式很简单，其实最重要的是我们来看一下，就是这个概率怎么去求的问题？。所以接下来我们相对多篇幅，其实是讲概率怎么求的问题，比如说先验概率怎么求，条件概率怎么求的问题。

我们看一下这两个参数的问题。首先这个先验概率很显然，你有多少类？这些肯定等于1，这是一个先决条件，因为是先验概率，所以数据可以得到一个统计。大家可以不看PPT，我们随便说一个数，比如说2020年12月31日是周六和周日的概率是否相等？首先是确定的，但是一下子反应不过来，估计十有八九是认为相等，但实际上好像有一本书，它特别逗，这本书挺好玩的，大家如果有兴趣可以看一本书叫做《How to solve the problem》这本书特别简单，这本书是一个英文版，就是讲一堆很有意思的东西，刚才我说的那个问题可能不见得2020年，可能是2220年，大家看这个就是说某一天是某个星期的概率，它会讲人的直觉实际上可能差别。但是不管怎么样，这些先验概率都可以去求对吧？比如说这个正常细胞和癌变细胞的概率，这些东西都可以在医院数据得到一个统计，如果数据比较大的时候，可能医院每个月统计的数据都会差不多，所以这个事情一般来讲不是一个特别大的问题。比如说，要求得一个汉字的字频，你可以用这个，比如说某一个部分统计样本去做，比如新华社统计资料去做，先验概率相对来讲比较容易去求，只要你的样本足够多，统计意义上是一致的，从统计数学来讲样本多应该是一致的。

第二个是条件概率，条件概率其实就是一个训练的过程，就是说我们有一堆的训练样本，训练样本里我们看到，比如说它的观测，这地方就相当于有一个频度代替了概率，就是用已知类别的训练样本，来估计类条件的先验概率，这个也是可以通过样本也训练得到的。另外一个是比较常见的，也是用的比较多，也是我们这周要讲的事情，就是我们分析样本形成的条件，然后我们用简单的一个已知的概率密度，分布函数来直接的拟合我们知道的条件概率，用的最多的是正态分布，我们这周讲的都是正态分布，包括单峰的正态分布和混合的正态分布的问题。

贝叶斯判定准则有其他的一些形式，这个形式上有很多的变形，比如说两类问题，条件最大似然比的一个判别准则，这是说，我们这个符号变一变，两类问题属于这是先验概率，先验概率放在一起，左边是一个条件概率，比如上面的下标是i，下面是j，i/j，要大于它的先验概率，先验概率是反的，是j/i，如果是大于它就属于第i类，小于就是第j类，为什么这么做？其实比较简单，就是右边这个数是一个常数，所以你来一个样本，只要算这边的比值，我不需要非要常数跟它相乘，然后比，所以左边跟你的位置样本相关，右边是一个常数，这样一变形直接可以计算出来。当然一个特例，如果先验概率都相等的时候，那就直接变成了条件概率，哪个大就属于哪个，这属于必须简单的一个特殊的形式。

这个是属于一种变形，这个变形非常简单只是把右边变成一个常数项。

我们仍然以两类为例，来看这个事情，这是一个之前课件里出现过一个PPT，我们之前应该讲过类似的东西。我们先看一下这张图：

这张图是条件概率密度，这时候属于如果我们两类问题的时候，我们把阈值分界点或者分界面放在这的时候，左边属于第一类，右边属于第二类，这时候所有红色区域，显然这个公式就很简单，这个x属于第二类，但是属于我判别为第一类，那么就是属于这边的东西，这是第二类，但是我判别为第一类。右边是属于第一类，但是我判别为第二类，两个相加是一个错误概率，这是错误概率。这个是我们的分界面，我们来看一下，如果用贝叶斯准则判定的时候，分界面是在哪个地方呢？应该是这个地方，因为这个地方显然不属于贝叶斯判别，对吧？只不过在这点上来讲，这时候它还应该判别为第二类，而不是第一类。所以贝叶斯判定的分界面的时候Xb，就是贝叶斯判别的这个属于分界面。这时候我们可以看到，它判定的时候仍然会有一个错误的概率，但是这个错误概率来讲，是这里面是最小的。这个是属于贝叶斯概率判别的地方，这来下面的公式，都是来解释下面这两项的，包括这个积分。

为什么是贝叶斯概率是最小的呢？因为你每一个地方，其实你都是希望它的错误概率最小，正确的概率最大，所以每一点它都是说正确概率最大的时候，很显然你整体的错误概率最小，我们用这张图，其实讲的是贝叶斯概率它得到的错误概率是最小的。

我们回过头来看上周我们讲的K近邻推导的时候， 我们讲K近邻它的错误率，属于是大于一倍的，小于两倍的。所以同样把它扩展不到多类的贝叶斯判别准则的时候，我们仍然有同样的性质，为什么有同样的性质？同样是说多类的时候，我们在观测的时候类的概率和先验概率、条件概率时候，在每一点上我们都是求最大的，所以，只有最大的时候，我们才判定它属于哪一类。

所以就是在有s个类别的时候，我们相当于把模式空间分割成s个区域，R1R2跟s，它的错的形式很多，我们直接算它正确的概率，正确的概率就是i=1到s，x是属于第i类，同时你把它判定为第i类，这个是它的正确的概率，正确的概率我们把它再展开一下，也就是说一个联合概率变成一个条件概率跟先验概率，每一次我们都取的是这个最大的，所以正确概率是最大的。所以这里面把离散的变成连续，就是一个积分的过程。如果是说你的正确概率最大，很显然正确概率加上错误概率等于1，所以意味着你的错误概率是最小的，1减去你的正确概率。

所以不管你的空间怎么划分，不管样本如何分布，贝叶斯的准则，对所有的x都是选择，所有的观测都是选择使积分值最大的划分方式，所以具有最小的错误概率，也就是说属于比较简单，其实你如果说你有最大的正确概率，那么最小的错误概率也是理所应当的。所以前面的意思是说，你的正确概率是最大的，你的错误最小。这句话读起来稍微拗口一点。

贝叶斯一般形式就是这样，同时我们通过这样的简单的解释，我们能知道贝叶斯能达到一个最小的错误概率，就是讲这么一个知识点。

我们来讲这个时候的再拓展一下，我们会发现这里面有一些小的问题，小的问题是什么？就是说我们没有考虑到所谓的风险，就是说属于错误的风险是不一样的。你两类互相判错是不一样的，就是说不管什么样去做，都可能会错，错的损失我们其实并没有考虑进去。在现实中过程中，实际有很多的情况损失是不均衡的，我们在这个地方，如果加上损失考虑因素，我们其实用的是最小风险判别的准则，来考虑你的分类的问题。

这个地方就比较简单，这个是汉字，汉字是属于判错的话，没什么大的问题，就是还好。另外一个就是数字，数字判错会有比较大的问题，比如说银行，如果是自动用这个机器来识别数字，判错很大问题，银行的数字，一块钱判定为二，两块钱判定为一，显然会损失钱，那么同样是1如果是说你判定为2、7、9，那肯定是完全不一样的，这是1、6、8，对于客户来讲，钱判多了，不会有损失，但银行来讲，它就损失了所以这个是完全不一样的这个损失函数。

就像刚说的损失，分别是0、1、6、8，这个同样是1，你同样判为1，就没有损失就是0，判为2的话损失1，这个地方是6，这个是8。对银行来讲，这个损失完全不一样的，这个是比较重要的一个地方。

在医院里面其实也是这样的情况，医院它是尽量希望误判会比较少，比如正常细胞误判为可能是癌细胞，损失为a，癌细胞误判为正常细胞，损失为b。我想肯定a跟b是肯定是不一样的，说一般来讲是这样的，希望正常细胞误判为癌细胞，就是a一般小于b，在医院里面希望不要漏掉癌细胞，如果误判为正常细胞，就丧失了治疗的机会，这个b我们希望要大一点，但是这个a就会小一些。实际情况是什么样的呢？如果说一个正常人去医院里面，说有癌细胞或者得了艾滋病，这个事情很多人就受不了，精神脆弱一下子就不行了，这时候，我们要采取另外一种机制，做这样的事情，我们一会儿会讲到，你要说是正常细胞判定为癌细胞，正常人判定为你得重病了，这时候，采取的措施是什么措施？复诊。现在各个医院其实很难保证百分百对，尤其是刚开始在稍微差一点的医院，它可能会误判，说你可能有病，这时候的做法，不管是医院还是你人员，你的做法一般都是去好的医院再看一下，县里面不行我到省里面省会医院不行我再到北京的医院看一下，这个事情是自然的一种机制，然后谈到医院机制有点是自发的，但实际上讲，现在的医疗改革，就会往这个方向走，我们叫分析诊疗制度，美国其实已经有这样的这种情况发展，那么这个的话属于将来对与这个地方，a就比较小了，通过分级诊疗，每一次都会降低它的错误概率，这个会比较好一些。

另外一个就是人脸识别抓逃犯的情况，这个情况稍微的有一点不太一样，比如说需要抓逃犯了，一是你不是逃犯，被认为逃犯，那这是一个损失函数，第二个就是你是逃犯，通过化妆淘掉了，这也是一个损失函数，现在社会里可能正常人被误解为逃犯是受不了的，所以不管在哪儿都要打官司，这就属于有损失，这里有公民权利的问题。如果说逃犯被漏掉了，公安部门自己要提高技术能力，这个虽然也有损失，但是要比正常人被误判为逃犯还小一点。

所以举了这些例子，都是说不同的误断有两类问题，i误判为j或者j误判i，它其实在生活中必然是不一样的权重的，所以我们要定义风险函数。如果是说这个是第i类，那么判为第j类的时候，造成的风险函数就是Lij，这个风险权值降下去，那么这时候Rj是什么？就是说来了一个样本，然后你判断它是第j类，这时候是有风险的，所有风险的函数要计算出来，就是等于右边的公式，等于∑的i等于1到S，那么也就是说你来了一个样本你判定为这个第i类的时候，它的风险是多少？你判定第i类，那这时候就是Lij，实际上它是第j类的一个样本风险是多少？

这是Lij，乘上它的后验概率，也就是说这x把s判定为第j类，但是产生的风险，你把所有它是可能是第i类的整个各种的风险做一个平均就是它的整体的风险。

所以最小的风险判别函数就是选取一个j，来了一个样本之后，把它判定为第j类所有的风险最小，就是它的判定准则，风险最小的话是属于跟上面相关，稍微的把后验概率变成一个先验概率或者条件概率的时候，就变成这样的一个式子，唯一区别是加了一个Lij，然后∑的i等于1到S，这里面不一样的一个地方。

我们看是属于你是第i类，第i类的时候你是X，但实际上你把他判定为第j的时候，它的风险加了一个Lij这个因子之后，就变成这样一个公式。

我们同样把刚才S类的情况变成两类的问题来看，这样的公式得比较简单一点。比如说这个两类的问题，比如说r1（X），来了一个样本我判定为ω1类的风险就是L11，首先是第1类的条件概率，这个事情其实也比较简单，我们看一下这XS在先验概率里属于第i类，它有多少概率，这时候属于在第1类的时候出现X，这时候L11，然后看一下第2类，第2类的时候，本来这个x它属于第2类，第2类出现X的概率，这时候实际上属于第2类，但你硬要把它判定为第1类，然后L21，这时候风险是多少？这时候就是第1类的风险。同样的话第2类的风险跟上面是一样的理解。这时候如果说是哪个风险最小我们就判定为哪一类。这是上面小，就判定为第1类，这个是第2类小，就判定为第2类。

把上面的式子稍微做一下变形，变形之后我们看，属于把这个稍微的做了一个减法，把第2类合并同类项，第2类挪过来L211-L22，如果小于L121-L11的话，我们看一下，一般来讲，很显然就是ii自己是这一类，然后判定是这一类，显然这个风险会比较小的。这时候肯定会出现第i类判定为第j类，Lij是大于这个东西，这个假设基本上是成立的。

我们来看把这个公式稍微的变一下，原因很简单希望是右边是一个常数项，左边是来样本的时候一个变量，跟s相关的一个函数。这时候属于这个跟之前是一样的，第1类的条件概率跟第2类的条件概率相除，如果把这个风险的因子拿掉之后，其实就是先验概率，是第2类、第1类先验概率相除，这时候就多了一个风险的因子。这个是第2类的时候判定为第1类，L21-L22，这个地方是L12-L11。规律也比较简单，这个是第二类，前面都是2，后面是1-2。那么这个后面也是一样，这个是1的时候，下标是1，后面是12、11。

这个时候为什么有这个假设？那么很显然首先这个地方他不等于0，相减的话是一个正数，所以整体上这边相当于把风险的因子放在右边，当然是说你这个阈值一算，完了之后就比较简单了，就是说左边的函数就是风险的函数x带进去，如果大于阈值是属于第i类，小于是属于第2类，这样就变得简单一点。

那么这个是关系最小风险判别的公式。

多类的时候也是一样的，多类与的时候，仍然是最小风险判别的一个准则，就是每一类Lj是判定ωj类的风险，这是一个公式，如果说哪个类或者最小的话，它属于该类。所以这个是扩展到多类的情况。

跟最小风险判别和贝叶斯之间有什么关系？其实也很简单，就是说关系就是，把你的风险函数变一下，就是为i≠j的时候，这个函数是0，i=j的时候是0，这个叫克罗内克符号“δ”，我们把这个符号引进去，定义一个风险函数，就是Lij等于什么？等于1-δ的ij，这时候这个函数是说，i=j的时候是等于1的，那么i≠不等于j等于零，对应上的风险什么的，就是说如果说i=j的时候，它的风险就是0，不等于于的时候风险就是1，就是说这个贝叶斯最开始就是这样的形式。

我们探讨了最小风险判别的时候，我们跟贝叶斯之间的关联就是说贝叶斯判别是最小风险判别的一种特例，属于你定义一个风险函数，函数改变这样形式的时候，这时候把它一变形就是贝叶斯的判别。最小风险判别是贝叶斯判别的一个扩展，贝叶斯也同样是最小风险判别的一个特例。所以PPT里面讲贝叶斯它就是这时候的一个特例。

我们来看刚才的最小风险判别的一个准则。同样还有其他的一些判别的准则，这些判别准则都是跟实际问题相关的不同的操作的手法，比如说这个N-P准则，它要做的时候，比如说限制一类的错误概率，不大于某一个常数项，在这个条件下，要另一类的错误概率最小，这个也是经常用的，经常用的比如说刚刚讲的，不管什么样判定的时候，我们希望癌细胞误判为正常的概率尽可能小，它小于某一个值，大概是要至少到十万分之一以下这样的话可能才比较好，概率会比较好一些。我在这样的情况下，我们希望正常误判为癌症的时候，怎么样最小，这个是有一个条件，那么另外一个获得最小。

另外一个逃犯误判为正常，就是说我们希望在漏判的情况下，怎么样使得在正常的误判下也是最小，这些都是一种判别的准则，就是给另外一个作为一个阈值，把另外一错误概率变得最小。其他的恩属于不用先验概率，用一些概率密度函数，这个是属于离散，我们要用概率密度一个延续量的问题。

另外一个判别准则就是刚才提了叫做序贯判别方法，序贯判别方法就是说这个有两种它的目的，第一，降低获取特征的费用。逐步增加特征数目来降低判别风险，比如说这里面我们可以举个例子，比如说现在小孩出生，每一次到医院检查这个一般都是会保险报的，报保险的时候，它不能够说每次检查都很贵，所以来讲，他刚开始是做一个普通的检查，比如说在胎儿检查过程中其中有一个叫唐式筛查，这个大家可能不熟，过几年就知道了，唐式筛查是判别也没有畸形的问题，一般来讲，用血检之后就能够有一个概率，这时候一般来讲跟刚才来讲是一样的问题，比如说他说你的唐式筛查风险在十万分之一你就放心了，一般是没有什么问题了，你不用去复查，但是如果是说风险是百分之一，意思就是一百个里面有一个可能会出问题，就是胎儿畸形的时候，你就得小心了，百分之一，一般人觉得可能没事，但是真的要是出事话，就是百分百，这个时候一般做的事情就是得去高级医院，比如去协和，做的事情这时候我们不能用血检，去做这个事情了，这时候是用穿刺，取不同的体液来判定，是否有畸形的可能性。刚才我说的事情是我经历过的，我们家老大出生的时候，我们去协和经过这么一次，这个就是属于序贯的判别方法。这个是在日常的生活面用到的。主要是说一个是费用的问题，你怎么样去用这个机制去做这个事情，那么我们模识别里面也有类似的方式。

第二个是需要逐次计算和比较新特征费用与所降低的风险，由此决定是否继续加入新特征，这时候一定要注意是加入新的特征，用原有的手段基本上来讲，其实你消除的是一个测量误差，不是用同样的特征去判定，假设你去高级医院做的时候你不能用同样的手段用血检就可以，那样你怀疑的是次一点医院操作的水平问题，必须用新的特征，你才能够得到进一步错误的降低。

另外一个就是跟计算量、存储量、特征相关，也是跟先验概率相关。刚刚讲了胎儿畸形总体上还是比较少的，你要通算一点，概率十万分之一或百万分之一，就可以不管这个事，就这么过去了一般来讲也没事。一般来讲，跟先验概率也相关，最后获得在统计上是一个费希尔的方案，这个是可以采用这个序贯的判别方法。

稍微总结一下，其实统计判别的特点，有这么三的特点，第一个就是属于数学模型清晰，理论分析比较完善。不少结论数学上可以证明，计算比较便捷，很多的方法经常用这种贝叶斯判定给出一个错误线，就是KN的时候它给错误线的时候说它是符合什么样的分布的时候的错误线。也是便于这个机器的实现，切实解决许多实际问题。目前使用最多总体上效果最好基于统计的一个判别的方法。

但是就是说这时候基于统计的方法，一定注意你自己的数据模型和实际问题的内在的实际模型之间吻合的程度，所以还有很多需要的样本，比如密度、先验概率这些东西，这个需要数量比较大，比较精确的参数估计，这个参数估计很重要，但是如果你有这种贝叶斯的假设可能会好一些。

第三点也是特别的重要，这个也是一样的道理，我们在讲模识别的时候，也在强调，就是说特征的选取和筛除的时候，它跟你判定是非常相关的，整个判定的方法的性能，它跟特征是相互依存的，比如说类别的可分程度，密切依存于所选的特征，这个其实我们之前就在课件里面反复强调过这一点，那么我们同样把原来的PPT再看一边。

比如这个是鱼的分类的问题，三文鱼和海鲈鱼，我们可以看到，分布上，近似看一下也有点正态的味道，但是这时候取的特征lightness的时候，显然不管你是用贝叶斯方法来做，它的错误率还是挺大的，这时候我们可以看到，比如说像这一点是贝叶斯判定，既使是这样，我们也不能简单的说，这时候的分布判定是最好的，它的意思是说，在这个特征下贝叶斯判定是最好的。所以要加这个条件，比如说像这个条件下，在这个特征的情况下，比如说这点是贝叶斯的判定，这时候在这个特征意义下它是最好的，显然这个特征换了，同样是贝叶斯方法，获得的错误的概率是不一样的。这也是一样的道理，这个是我们用斜线，同样用贝叶斯的方式获得这样的分界面，判定它是一个正态分布的时候，我们后面有一个很好的理论说法，它大概解析解方程是怎么样的？

对于贝叶斯的判定，我们用一个小的文献做一个总结。这个文献其实是发表在去年12月份的《science》的文章，它其实讲的是什么呢？这里面是有很多的叫Random forest，Random forest是我们下周讲的知识点，叫做随机森林，里面包含了每一棵树叫做角色树，这个是神经网络，这些我们都会讲，这些的方法它们的都需要很多的训练样本，SVM稍好一些这些都需要很多的训练样本。

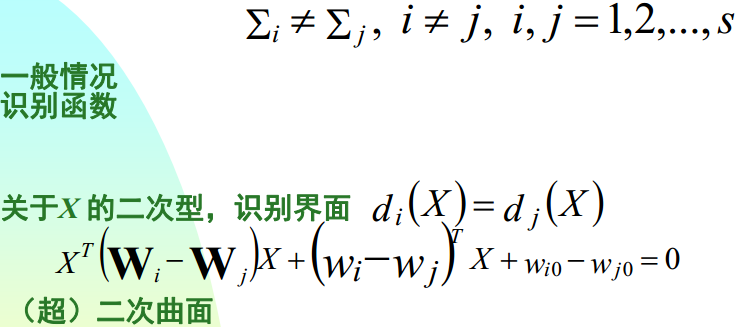
什么叫图灵测试？大家知道什么叫图灵测试？其实图灵测试是图灵讲的一话，我们希望机器达到什么样的功能呢？其实这个功能很多时候也实现了，你在跟一个见面的时候，你不知道对方是人还是机器，如果按我们智能能做到这样的话，那机器就通过了图灵测试。比如说一个聊天软件，你很有兴趣的跟这个软件聊，你以为在跟一个美女聊天，但对面坐了一个机器或者说一个宠物在用键盘敲字，甚至对方是一个两岁的小孩都有可能，比如说像现在我们家的小孩，现在就是两岁玩Pad玩得很牛，发各种表型你都不知道。如果机器人能做到这一点就通过了图灵测试。

我们来看一下这篇文章，大家有兴趣的话可以看一下，是去年12月份发的这篇文章，看的时候要有自己的概念去理解它，我稍微介绍一下，有兴趣的自己去看。

这篇文章叫《Human-Level concept learning through Probabilistic Program Induction》其实来讲就是一个贝叶斯网络的一个方法，那么这个文章本身来讲它的立论是什么？当然能发在《Science》的原因估计有两点，第一点就是这个作者是MIT的教授，第二就是，现在这么几年的时候，大家都在说这个深度学习，那么这个《Science》总喜欢登出一些反潮流的文章，大家说深度学习，我们就一定要讲非深度学习的文章，说这个文章比深度学习好，这篇文章其实它立论就是，说我不用深度学习贝叶斯就比深度学习好，就这样一个方式。

它本身整个的方法并不是特别的难，它是要做的事情，是给了一个不大的数据集，大概只有20类，每个类只有30个样本，一堆的拉丁系的字符集，就是非英文也非汉字的字符集，就是每类里面有大概30个字符，它是希望仅仅用少量的样本，让机器学到什么能力呢？让机器自己写一个字符出来，然后让人去看，说这个是人写的还是机器写的，这时候做了图灵测试，人其实判定不出来是人写的还是机器从里面构造出来的。所以它的文章叫Human-Level concept，就是他觉得用这么少量的样本，就能学到这种书写的风格，使得这个人都被它骗到，这个文章就是这样的情况，从它的技术来讲并不是特别的难，包括整个的方式，贝叶斯架构它是认为一个字符，能分出很多的笔划，笔划都有顺序，然后每个笔划用变量条去做拟合，所以写的时候，会故意把这个变量条拉伸或者是缩放，使得看上去不是一个规则的东西，人看这个比如说是不是机器的字符，你可能会说它比较规整规范，就认为是机器的一个字符，但它会有一些因子去缩放这些东西。就像这些东西，感觉好像故意的像比较挫的人写的，看上去像人写的，实际上是机器骗人的，机器写的。这篇文章方法比较简单，但它用的是贝叶斯方法，大家如果有兴趣可以去看一下。对于文章不同人有不同的简介。

这个是贝叶斯判定，其实刚才就讲完了。

我们接下来探讨的都是参数估计的问题，我们首先讲一下**正态分布**的贝叶斯判别准则。

这个里面，它的条件概率是满足正态分布的时候，贝叶斯判定准则是什么样的？通过这个是贝叶斯的一个特例，同时跟我们之前的比如说，LDA都能关联起来，我们都可以看到那些方法在一定的情况下都是贝叶斯的判别准则的一个特例。

讲这个事情是因为条件概率是非常难以获得的，正态分布是一个很好的模型，这个模型在讲概率论的时候，一般都会讲，什么时候我们要用正态分布？或者什么时候用大数定理？这里都是一样的道理，就是说如果某一独立随机变量，它是由大量的相互独立的随机因素影响所造成的时候，有可能就是用正态分布来做这个拟合，这个是正态分布的一个好处。并且每个随机因素在总体的影响中，都是均匀的小，不是有一个主导的随机因素，如果有主导的随机因素，就不符合正态分布，这是两个假设的时候，很多的随机变量都可以用正态分布来估算。

正态分布主要有两种，第一种是单峰的，单峰的正态分布，接下来一小段时间我们讲单峰，后面讲多峰。单峰里面有单变量，所谓的单变量就是标量，我们从最简单的标量开始看，x如果属于一个单变量的正态分布的时候，它的概率密度是这样的形式，这时候它有很好的性质，性质就是一个均值和方差都特别容易算，这就是为什么用正态分布的一个原因。我们看一下单变量，单峰的正态分布，它大概什么样的情况？它整个的分布回顾一下。这个原则上所有人应该都会。

我们来看一下，这个M是均值，∑是方差，这个概率密度大概长这样子的形式，这时候，就是说在这个一倍方差的时候，它有一个拐点，一般大家都看在一倍方差以内，它的概率已经最够大了，这个外面最够小，所以画正态分布的时候，一般是以M为中心，画一个圆或者椭圆，但并不意味着它是有限的，它是无限的，只不过在这个圆或椭圆里面，概率最够大的，这个里面是有它这样的原因。

我看一下多变量单峰的正态分布概率密度函数，多变量的时候，比如说n维的随机变量这个X是有X1、X2，Xn，也就是而n维的一个向量，这时候在条件概率这个ωi条件概率就等于2π的n/2，∑的1/2，然后在里面的话这是一个指数项，指数项里面这是一个协方差矩阵，左边是跟标准样本之间的差向量，这个是被协方差矩阵所调制的，这是它的一般的形式。这里面就有一个均值，也是一个向量，是一个n维向量Mi1、Mi2和Min。协方差就比较简单了，就是（x- Mi）（x- Mi）两个之间做一个行乘列，是一个n×n的矩阵，这时候它是一个对称的半正定矩阵。对角线就是自己的方差，非对角线是不同类之间的协方差，这里面是11、12到1n，这里面是n1、n2到nn，这是一个表达形式，上面是行列式的表达，下面是矩阵逆，这是一个符号的说明。从这上面来看整个的形式比较简单，是有两大类的位置变量，一个是Mi，一个是它的协方差矩阵∑i，所以来讲相对是比较简单的。

我们看一下属于变量的数目，首先Mi1你统计一下第j类的x量协方差，下标是ωi类就是均值向量。同时协方差也是一样的，x中的第j类特征的方差，这个也很简单。

非对角线的方差，第i的维度和第j的维度它的协差，如果比较好的形式的话，是说不同的维度之间，它是一个统计独立的时候，那么它的协方差的值就等于0。

如果所有的非对角线都为0，联合概率分布就比较简单，就从刚才的特别复杂的形式变成了很多的单变量的正态乘积。这是一个很好的解析方式。

我们来看一下刚才提到整个表达里面我们有多少个要统计的变量？一个是你的均值向量，一个是协方差。这里面有n个均值和协方差它是对称的独立的元素，所以整个的变量是n+n（n+1）/2，我们其实要从训练样本里面统计这么多的数值。

我们来看一下，多变量直观上长什么样？这是一个2维的情况这是一个特征，这整个是一个概率密度，如果从一个切面来看，单维的时候看就是这样，多维的时候长得有点像帽子，如果是说把这个画成一个平面的形式，我们用等概率密度的方式去做，刚才提到这个方向，为什么我们用椭圆来看？我们都是有它一倍的方差来表示，因为在这个里面，这个概率是最大的，所以来讲，我们一般是用椭圆来不断的画它的等概率密度，这时候，可以看到，整个呈现这样的分布，正态的分布产生的样本，是大部分落在中心为Mi，这个Mi是统计的均值的向量，区域内它的形状为∑i球体里面。

这个是刚才提到的方差的概念，在一维里面是一个∑，一倍的，协方差就是∑，如果是2维或高维的时候，就是协方差超球体内。这句话其实来讲就是比较简单，但是这里面我们接下来这么简单的式子，我们也用一个公式去推导一下，比如说等概率密度，这里面我们令它的指数项等于k，就是常数项为常数的时候P（X）不变的。所以我们等于概率密度，它的形状是什么样的？我们相当于是说用一个简单的推导，就是不断的用协方差，属于椭球体内，我们设这个为k，设这个为k的时候，我们其实要求得的是它的短轴和长轴，既然是椭球就是长轴和短轴，它在欧氏空间里面，其实表达的就是一个距离最大，也就是说x和它的点集距离最大，这里面特别简单，这个推导的结论其实我们知道，就是说这个椭球体它是由协方差决定的，协方差里面由什么来决定呢？是由它的特征值，这个结论我们知道，我们稍微的过一下，看到底是不是这样的。

这个地方我们希望最大或者最小，所谓的极值，这个里面的话要求它的最大、最小，这是等概率密度，之间有一个 λ，这时候就是求一下偏导，偏导我们就不用可以刻意的念每一个，偏导完了以后，其实得到一个∑x=λX，那么很显然向量是∑它的特征向量的或者本征向量的，也就是说x的组合方向是协方差矩阵的本征向量方向，就是说左乘对，这个式子我们就不去讲了。

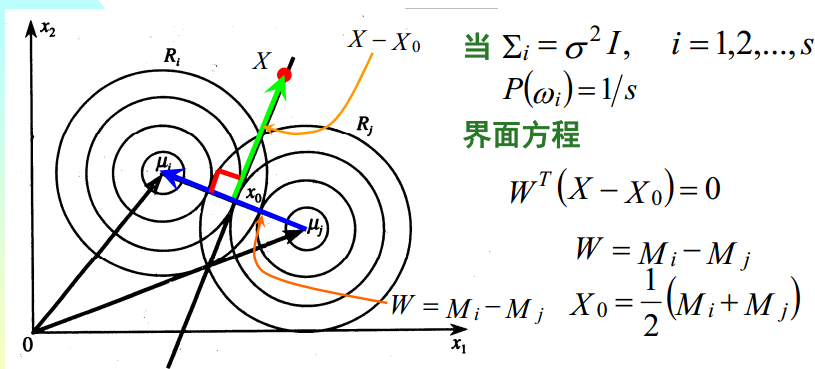
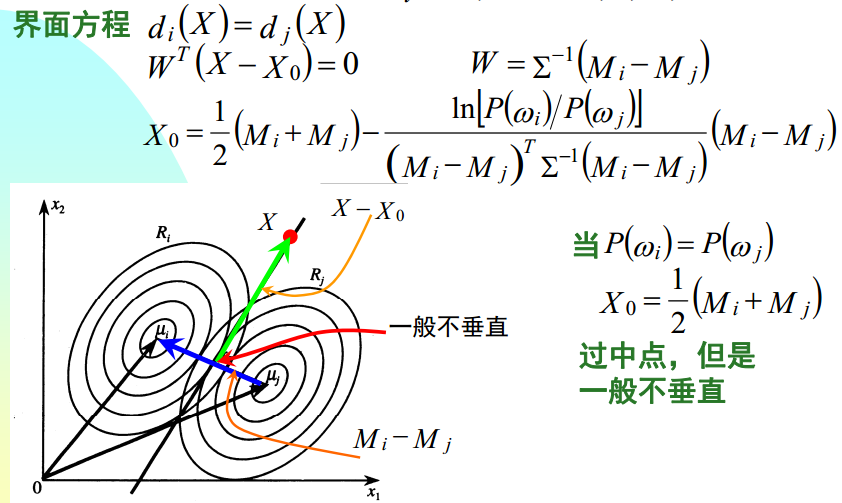
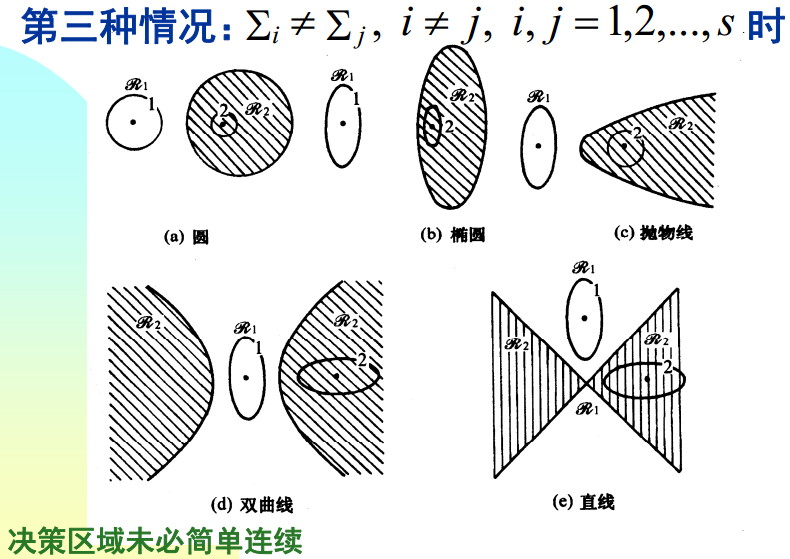
我们直接看这个式子，刚才说的等概率密度，其实来讲在2维情况下，它的椭圆就是一个长轴跟短轴，长轴跟短轴的是和这是它的大小，方向就是说本征向量的方向。

这个就是属于x到它的标准向量的马氏距离，整个就是这样一个方式。

这个就是用了一页的推导，其实要表达的是，我们接下来画图的时候，要画成单峰或者多变量的正态分布的时候，我们就画成这样一个形式，就是说属于椭圆的方式，它的方向就是你的协方差本征向量的方向，它大小的范围就是本征值。其实就是讲这么一个事情。

我们现在休息一下，下节课我们继续讲在单峰的时候，它的概率密度的表现形式和观点。

（62：55-71：40休息）

我们开始第二部分，刚才对单峰正态分布时候的概率密度本身，稍微的介绍一下介绍，包括它的形式什么？接下来主要讲的是正态分布时候的最小错误概率识别函数和识别界面。假设这个样本它符合正态分布的时候，它的最小错误识别概率的函数和识别界面，贝叶斯的识别界面是什么样的？从最小错误概率识别函数上讲，这是贝叶斯的一个方式，包括先验概率和条件概率，先验概率是已知的，条件概率是符合正态分布的，这时候符合刚才讲的正态分布的一个形式，是一个方差和协方差矩阵组成的一个指数的概率分布。这时候我们把这个式子带到上面去，这是一个先验概率，这是一个指数函数，这时候带进去，两面组成一个自然的对数，这样它指数相当于没有了，这时候设一个常数项，减去这个常数项，减去变量，这是一个观测的样本和它的协方差，协方差和它的均值向量，这时候常数项我们把它放在一起，我们首先是跟i无关的拿掉，跟i无关的就是这个东西，这时候它的识别函数在符合单峰正态分布的时候，我们相当于重新把它定义一下，这样就把ln去掉了，这时候是先验概率减去它的协方差ln，减去两倍的这个东西，这个时候是它的识别函数。

这个部分x是一个常数项，来了一个样本的时候，这是一个常数，这是x的2次型曲线，这时候我们要看它的识别界面，也就是d i类和dj类之间等于0，很显然两个X的2次型减去一个x的2次型，通用上来讲就是2次型曲面，现在把二者放在一起带进来，这时候这边是一个i的减去j的2次型，这边是将常数项进行相减等于0。

一般来讲，你的协方差矩阵 ∑i和∑j是不一样的，d i类和d i类是不一样的，这时候，在一般的情况下，识别函数就是这样的一个形式，d i（x）=xTwix+1次项+常数项，这相当于把它展开之后2次的形式。w就是-1/2 ∑i-1。这个是相组合一下∑i-1Mi这个是常数项。

关于x的2次型识别界面，两个是相减或者是相等，这时候它就是相当于说Wi-Wj减去wi-wj，后面是一个常数项，这个是在单峰正态分布的一般形式上的识别界面就会产生这样，很显然，这时候不等于0零的时候，仍然是一个超二次曲线。

这是它的一般形式，接下来我们讨论它的一些特殊形式，从特殊形式来看一下大概识别界面之长什么样的情况？

我们来看一下最简单的情况，它的每一个类都符合各项同性的正态分布，也就是σ2I，整体来讲是一个单位矩阵乘上一个常数项，也就是说，所有的特征是相互独立的，每个特征都有相同的方差，这时候它的矩阵就是这样的形式，对角线的形式。∑i的行列式就是等于σ2n，∑i-1就很简单，每一个对角线上/1σ2。在几何上来讲，各类的样本都分别落在以它的的标准向量为中心的，同半径的圆体内。识别函数把刚才这个d i（x）等于这样的2次型带进去，中间是一个单位正，之后这里面就是一个距离的函数，X-Mi距离加上一个常数项，这是d i（x）的识别函数。

假设说类的概率相等，这其实就是最小欧式距离判别准则。大家可以看这个图，如果每两类之间情况，它的每个特征方差也是一样的，每个都是一个圆，是两类的问题的时候，如果是说它的先验概率是相同的，你的分界面就是这样的，这时候就是最小欧式距离，属于你的协方差是一个单位正乘上常数项，如果它的先验概率是一样的时候，它的两个界面方程就是WT(X-X0)，W=Mi-Mj，Mi-Mj就是原来我们说的投影向量，投影向量是Mi-Mj，X0=1/2（Mi+Mj ），所以这样的界面方程就是二者之间的一个中垂线，这个时候，最小欧式距离在这样的情况下一也是一个贝叶斯的判别，也是最优的情况。

我们来看一下当你的协方差矩阵，仍然是一个单位正乘上一个方向的时候，就是所谓常数项的时候，跟最小欧式距离不一样的时候，是什么时间？比较简单。比如你的先验概率不一样，先验概率不一样是两类的问题，如果贝叶斯判定的时候，比如说分界面在这儿，在这儿甚至包括了红色这一类的它的模板向量，这边都属于第1类，这边属于第2类，为什么会出现这样情况？是因为你的贝叶斯判定的基准X0它是跟先验概率相关的，也就是说某一个先验概率比较大的时候，你就可能往这边偏，这里面也挺简单的，我们举个比较简单直观的例子，如果是啊说某一类先验概率特别大，这时候来了样本你去猜，就直接猜概率大的，哪怕你都错了也无所谓，都错是小概率的这边错了。所以先验概率是很影响你的分界面的。这个是比较简单的情况。

这个地方给了一些图示，比如说先验概率都是相等的时候，比如这里各是0.5的时候就是最小欧式距离的判别。那么同样，先验概率如果是不一样的，就是不一样。这是在三维的情况下长什么样？这是一个包络，也是一个中垂线的概念，这个先验概率是一样的。这个是先验概率一样的情况，一个球体的情况，如果说先验概率不一样，刚刚提到了，比如说这里面第一类的先验概率70%，第二类30%，很显然，这时候的贝叶斯判定就不是简单的两个概率密度之间相等的地方，那么就是在这个地方，这个是反而贝叶斯判定的一个分界面。如果更夸张一点，这个是0.9这个是0.1的时候，这个分界面就会更加的偏一点，所以这个事情是很简单的，只是大家的直观上如果看到一个概率不要简单的就是认为是二者相等，跟先验概率有密切相关的事情。

这里面是一个三维的表示比如0.8，0.2，那么我可以看到，三维情况下，分界面在这个地方，虽然图示上画这一类，这个是它的标准向量，那么这个圆是它的协方差，这个是大小，但是分界面跟这个是有交叉的，这个是极端的情况，基本上这是大概率出现的，这是小概率出现的，这个分界面可能是落在这儿，从概率上，密度上来看基本上灰色的是第1类，红色属于第2类，这块直观上两个相等的时候是不一样的。

反过头来看，我们上半部分讲的时候，两个相等的时候大家看错误概率，大家看那个时候要注意，那时候纵坐标不是概率密度，是先验概率×概率密度，所以那时候是相等的，稍微有点区别。其实这个图书上都有，只是大家在看这个图的时候有各种的看法，用不同的形画出来的。

刚才我们把第一部分，最简单的部分讲了一下，最简单的部分是说协方差都是一个单位正，意味着每一个维度上的特征都是一样的，也就是各向同性的，最小欧式距离的特点就是说，它的先验概率是一样的时候，就是最小欧式距离，不一样的时候它的投影向量是一样的，只是贝叶斯判定的方式不一样。我们来看一下第二种情况，第二种情况是第一种情况的扩展，也就是每一个协方差都是相同的，但是不同的地方它具有一定的形式，特殊的一般形式，都是∑。每一类的协方差矩阵都是相同的，但是它具有任意协方差的形式，就是说这个矩阵是一样的，在这时候我们来看一下，就是贝叶斯的判定长什么样的情况？

在几何上来讲，如果是满足这样的分布，每一类样本，都分别落入以代表向量Mi为中心的五，同方向的椭球内，跟刚才的球体内不一样，是椭球体内，也就是说属于每一维特征上它的协方差是不一样的，这时候它的识别函数d i（x）=1/2，这时候是马氏距离，马氏距离相当于被∑调式了一下，∑-1，加上它的先验概率，如果先验概率同样的时候，这个常数项就可以没有，这时候d i（x）就是马氏距离，计算x到各类中心的距离，这个平方，判定X属于距离最小的这时候，时候来样本，你分别跟每一类的标准距离做一下它的马氏距离，算完之后最小它属于哪一类，这个是它几何的意义啊，如果是不相等的话把这个放进去。

识别函数是这样的情况，如果忽略就是长这样的情况。这个是∑-1，这个因为是相面相等两个就会相减就没有了，所以没有2次相，我们看到这个界面没有2次相。

我们来看一下它的方程，反过头来，识别这两类的时候2次相就没了，整个识别函数就是一个1次项加上常数项，1次项就等于∑-1×Li，常数项是一个这样的形式。那么我们来看一下它分界面的情况，也就是界面方程的一个情况，界面方程的情况刚才提到了，它的向量就是你的（Mi-Mj）∑-1，常数项是长成这样的一个形式，看一下就行了。

我们重点看一下投影的矩阵。这个投影矩阵里面，我们看一下这个矩阵的形式，是（Mi-Mj），左边是∑-1，这个形式跟我们上节课和这节课刚开始时候的某一个公式非常像，那个公式这里面SW-1，很明显这二者间肯定是有关系的，是什么关系呢？我们来看一下这个特例就明白了，我们先看这个方式，如果这两类之间它具有相同的分布，这两个椭球体朝向各方面都一样，这时候分界面不是两个之间的中垂线，如果中垂线大概是这样的情况，这时候它是这个方向，然后旋转了一下，就是∑-1调式了一下，就是沿着你的倾向性有一条直线。

所有来讲一般不垂直对这个进行一个旋转，如果是说二者相同，这时候的X0=1/2（Mi+Mj），这个我们在前面做分类的时候提到，就是在费希尔准则或者LDA里面，我们要判定的时候，比如哪一边算是第1类，哪边是第2类？如果两个类概率相等的时候，这个X0=1/2（Mi+Mj）是最佳的贝叶斯判定的阈值。判定的这条直线，还有这条直线，也是过中点，但一般不垂直。

这时候，我们就把贝叶斯的准则和前面的费希尔准则二者之间进行了一个关联。

这是一个图示的情况，我们来看一下图示的表示，这个是所说的椭圆，为什么画这成这样？是因为这是一个椭圆的中心，椭圆的长轴和短轴都是跟它协方差相关，协方差它的本征向量相关。画成这样它的分界面比如就长成这样，这时候如果不在两个椭球之间，跟它的先验概率是一样的，这边比如说是先验概率都是50%、50%，这边就不太一样了，这边先验概率要大一点，这边要小一点。这个是用另外一种图示表示。这时候是分界面大概是这样的，这个是50%先验概率，50%先验概率，两个椭球体，这个是等概率的椭球体。它的分界面是这样的形式，这边分界面貌似切了一个椭球体，这个椭球体为什么切呢？因为它属于协方差的本征值相关的椭球体，先验概率不一样的时候，它可能更往先验概率小的一面偏，这个是第2类情况，所以第2类情况是属于贝叶斯判定跟费希尔准则、LDA是相关的，大家看这些图就要知道这么一个概念。

我们看一下第三种情况，第三种情况也是一个一般的情况，我们从特殊到一般，特殊到一般每一个类的协方差都不等，很显然它的识别的界面就是一个2次型，也就是一般的形式，它的每一类的识别函数具有一般的形式，跟刚才提到的，我们把它变成2次型的时候，有这样的形式，X2+X+常数项，两个边的识别界面就是差向量，两个W矩阵相减，和w矩阵向量相减，加上一个常数项，具有一般的形式。它的识别界面是任意的2次型都是有可能的，我们看一下它的一般的形式。

这里面把它变成了一个二维的形式，就是说看它的形式会长什么样？我们逐一稍微看一，第一个属于它的这个识别界面，识别区域有可能是一个圆，比如说这是第一类的分布，这是第二类的分布，如果这时候贝叶斯判定，它的识别界面有可能就是这样的一个圆，这个圆里面的都属于第二类，圆外面的都是第一类，因为这边的方差大，就是你的识别界面有可能是一个圆。

同样的话，它有可能是椭圆，椭圆我们看的时候，要注意是这对图，下面这个标好像有点错位了，看的时候不能把这三个放在一起，这两个放在一起，这时候是第一类，这个是第二类，同样这个方差大一点，这个方差小一点，因为这个从它的每一个维度上，这时候分界面有可能是椭圆的，这时候分解面甚至有可能是抛物线，从这个形式上来讲，这一类具有这样的一个的分布。这时候，它的分界面有可能是一个双曲线，这时候是一个什么概念？这是第一类，第一类是标准样本，它是协方差分布，这个是长轴，这个是短轴，这个地方是它的标准的特征，代表向量，这是短轴，这是长轴，这里面灰色的都是第一类，这个是第二类，这个很有意思，有的时候会退化成两条交叉线，这两条交叉线分布是一样的，这个是往上挪了一下，就变成一个交叉线，双曲线变成交叉线，也意味着两条直线是双曲线的一个特例。

从这上面来讲觉得有点神奇，分界面或者说判断区域，比较有意思，各种情况都有可能，我们来看一下这个三维的情况，首先这里面有几类，决策区域未必是简单连续的，这个说的是什么？就是说比如像这样的情况，这两个区域其实是不相连的，不是简单连续的，属于在第三种情况的时候，它决策区域未必是简单连续的，这是一个结论。

我们看一下三维的情况， 三维的情况就会反映到刚才提到的各种可能性。我们看一下一维的情况，看一下这个图，同样是一维的时候，这是一维的高斯分布，这也是一维的高斯分布，但是同样是一维的高斯分布，它的形态是不一样的，这个胖，这个瘦一点，从贝叶斯判定来讲，假设它的先验概率是一样的时候，我们可以看到，在这一小段属于比如说第二类，两边是第一类，对不对？这个时候第一类它的决策区域显然不是简单连续的，是被第二类区分开了，中间是第二类，两边是一类。这是一维的正态分布的情况，看一下高维的时候，它的三维情况。

这个时候，跟刚才很类似，这是一个扁的，属于第一类，这个扁的是第二类，这个中间这段是概率比较高的，两边是概率低的，很显然从贝叶斯看，这个判定很简单，中间这一条红的是这个类，灰的是属于另一类，因为外面灰的都比较平一点，它的概率就会高一点，这个是刚才的情况，虽然是两类，但是两类它的协方差分布是不一样的，在这样的一个对角的区域，红的概率比较高，在这个方向灰的概率高，所以分解面是长成这样。

这是刚才抛物线它在概率密度上的表达形式，这个是灰色这一类，这个是红色这一类，在概率上很容易看到，这一边灰色的势力范围，这边是红色势力范围，把它投影到二维的情况，曲线上来讲，它的分解面就是一个抛物线的状况。这里面是一个椭圆的状况，这是一类，另外还有红的这一类，灰的势力范围就是这个区域，外面都是红色的势力范围。

我们再看一下另外一个，就是说它分布也是不一样的，一个是向这样方向的分布，一个是向这样的一个方向分布，这时候从这上面来看，很明显这一片，对应红色的要大一些，旁边是灰色的大一些。这个是刚才比较特殊的圆的一个情况，圆的情况也是两类，一个是灰的，一个是红的，很显然里面是红色区域，外面是灰色区域，它的分界面就是这样的一个圆，这样看也是非常有意思的一个情况。

这是多类的情况，这就更复杂了，但是你把它画出来的时候，很能理解它们为什么下面会变成这样的，就是说整个的空间会变成不同的区域，比如说这里面，把整个空间给区分成S个区域，这时候它的区域各种情况都有，很显然属于不同2次曲面，超曲面等各种情况。

这个是四个正态分布的二维情况，对应到这里面显然就各种情况都有。

上面我们讲到的这些，是在单峰的时候我们做贝叶斯判别情况，单峰的时候还是比较简单的，单峰的计算也很简单的，如果是计算单峰情况，我们之前提到，我们相当于由n+n/2，n×（n+1）参数估计，这些参数估计通过算它的均值、它的协方差，都可以用一般的统计上的数字很容易计算出来，单峰我们都会用，作业里面也是跟大家的单峰一个情况，那么我们来给大家稍微的扩展一下，遇到多峰的时候，要怎么做？

个多峰在很多情况也经常会发生的，所谓的多峰就是，它仍然是一个高斯模型，它是一个参数化的模型，是由高斯密度来估计目标的分布。单峰情况我们之前提到了，我们花了应该有一小节多的时间，给大家看各种各样的图，包括有两种特殊情况，跟最小欧式距离，跟LGA都关联上。它一般形式看上去好像也比较复杂，即使这样，单峰的情况仍然是一个简单的、理想的状况，大家还是比较熟的用这个情况，我们看一下常见的我们实际中，经常可能你的一个随机变量不是用单峰描述的，而是用高斯混合模型来做的。

高斯混合模型它的表达形式是什么？i=1到K，这个k就是你的核有多少？每一个ni是一个简单的单高斯，这个是若干的高斯的线性组合，就叫混合高斯，就是说混合高斯比单高斯的情况其实要更加的通融一些，就是单高斯是混合高斯的一种特例，我们给大家扩展一下，单高斯到混合高斯的情况。

公式表达就是刚才一样的情况，是由k个加权的高斯函数线性组合的表示，我们看直观上会有怎样的形式？比如说像这样红色的点，看上去显然不是一个单峰高斯分布，但是右边来看，我们可以看到这个似乎是有三个高斯叠加起来，所以这个是单独高斯函数，这个是单独高斯函数，这个分布我们相当于是三个单独高斯函数的叠加。这是混合高斯的情况。

接下来讲，这时候要做的事情，我们假设有这样的样本登录的时候，这个混合高斯应怎么去估计？这里面估计比刚才要多的是πi怎么估计？这个πi是怎么样的情况？它落在那个单峰里面，相当于多了这么一个估计的情况。

估计也很简单，我给大家一个思路，具体算的时候，用一个迭代方式去做的，我们现在要定义另外一个变量Zi，是这什么意思？Zi1、Zik，K表示你有K个单独高斯函数，Zik表示这个样本Xi是否属于第K个高斯函数？这时候zik它是有两个去取值，就是说属于还是不属于，不属于是0，属于是1，1的话表示属于第K个高斯函数，0的话是不属于第K的高斯函数。这时候，有这样的表达形式，∑的k=1到K，zik=1，这里面虽然是连加的形式，但是实际上zik只有一个1，其他都是0，它只能属于一个的核函数。所以来讲，比如说这个P（Zi）它的密度，它的概率等于相乘，就是在π的k=1到K的时候相乘，里面πik×Zik，这里面等于0的时候就等于1，如果是等于1的时候就是一个单峰高斯函数，所以这个时候，就是P（Zi），我们知道第二个样本它就属于某一个单峰高斯函数，它是属于那个高斯分布，所以表达形式比较简单不复杂，但这里面它就是表达的是它属于某一个单峰高斯函数，这里面虽然是一个连乘，但是因为是0的时候就是1，我们来看一下鸡生蛋，蛋生鸡的问题。

我们知道什么的时候我们能算什么？我们先假设说，我们明确知道每一个样本，i是样本，它属于某一个高斯函数，某一个单峰的，zik假设是已知的时候，我们求一下整个的高斯函数的分布，就是说这个k高斯函数长的什么样的情况？比如说这里面我们对第k高斯核，我们无外乎也要算它的均值和方差，这地方没有k，那就等于什么？嗯1/nk，属于高斯函数里面有多少个样本？zik属于它就等于1不属于它等于0，所以里面连加的符号是I等于1到N，这个N是所有的数，这里面每个样本乘上一个zik，就假设zik是已知的时候，那∑方差也比较简单，之前把zik拿掉，就是大家知道的情况，这时候我们加上zik，这时候的标号就是1到N，这是所有的样本，zik就相当于不属于第K的和的话，它等于0，它不影响的，所以表达形式一样的，唯一是前面的除数是不一样的，是落在第k个高斯函数里面，它的样本有多少？是nk。如果ik是已知的时候，我们能够很容易求到你每一个单峰高斯函数是什么样的？每个单峰高斯函数无外乎是均值方差。

我们可以看到，把这个公式再做一下变形，变形是把这里面由每个这个东西，变成做一个期望，比如说像这个地方，里面xi里面不是求zik，zik的一个期望，这时候右边的公式跟左边的公式唯一的区别就是把这个zik变成了一个期望的公式放在里面了，为什么要这样放呢？是因为我们给了一个观测的时候，我们不知道zik是多少的时候，所以用它的均值替代里边的zik，因为在统计过程中我们并不知道zik这时候我们要把它均值放进去做它的一个估计，这个时候，同样，均值和方差能求出来。 那么我们来看一下，这组公式就是刚才的这个，相当于给了一个观测之后，我们给出它的均值来解。鸡生蛋，蛋生鸡，这里面就是一个循环的过程，实际过程中我们求证它的期望的时候，我们要假定我们是知道的，我们知道每个高斯函数的分布，我们要求期望，有了zik之后们就可以再计算它的高斯分布，有了期望我们就相当于又得到了新的高斯分布，这个是鸡生蛋，蛋生鸡的问题。到底在这个两个互相耦合的时候，应该怎么做比较简单？

既然两个互相关联，我们就打破一个求另外一个，打破一个是什么意思？假设我们知道了它属于哪两个高斯分布，我们先假设θ是知道的，这时候求每一个样本属于这个分布的，它的期望是多少？这个是我们假设完了之后，我们就可以得到E{zik}，得这个之后我们再重新算一遍，完了之后再这样迭代过来，可以看到已知的分布求每一个样本属于这个分布的期望。有了这个东西，我们回过头来看，你重新更新它的分布的情况，这样就是不断的迭代，最终我们就可以得到一个稳定的值。这个是一个鸡生蛋，蛋生鸡的问题，也是一个EM算法，最大期望的一个算法。因为EM算法不是我们的重点，我们给大家介绍在这种鸡生蛋，蛋生鸡的情况应该怎么做？你先给另一组，然后去进行迭代，不断的迭代之后，我们得到参数出来。

给大家一个例子，可能对这个事情就更加的容易理解了。我们来看一下一组样本，绿色是一组样本，我们直观上都给出来了是这么一坨，这么一坨，但实际上我们并不知道它到底属于怎样的高斯分布？所以我们就是给了一个初始估计，这个估计比如你认为这个高斯均值在这儿，方差是这样，这时候方差还是一个单位矩阵的方差，是一个很简单的一个估计，均值很显然并不是特别好估计，我们来看一下这样一个情况，这时候有了这么两坨，，这时候应该怎么弄呢？我们看这一坨样本里边，到底假设我们知道就是这么两类，这些样本哪些是这一类？哪些是这一类？很简单，从方向来讲，这边靠最近，应该属于这一类，进行了一次迭代，每一个样本都打上了标签，我们看一下，这是一个一堆的样本，我们对它进行一个高斯的抽检，高斯的抽检显然是这样的分布和这样的分布，我们就是又重新更新了一次你原来的高斯的估计，有了这个估计之后，我们看同样你的均值在这儿，协方差在这儿，这时候我们再来进行一次迭代的估计，有了这个估计之后，我们来看一下，这个是从这儿到这儿，这是第一次迭代，这是第二次迭代，假设有这两类，这时候把样本再重新划分一下，这时候，蓝色得就更多一些，红色的在这边，这边也有蓝色的一部分，再进行一次迭代的时候，这应该是在第5次的时候，不是同一次了，这时候你的分布是这样的。对每一个样本再重新看一下，在这个时候概率是多少？最终你的高斯分布就变成这样了，这是一个和，这是一个和，分差和方向都很明确。你先假定一个分布，你算每一个点属于这个分布的概率，有了这个之后，重新再看新的分布，这样不断的周而复始的去迭代，最后也能迭代出你的高斯分布出来，两个可能的高斯分布。这个是给大家的一个例子，当然如果从实现的角度上来讲，可能你自己需要去把程序再编一编，再迭代一下，包括它的迭代次数会怎样，自己要去实践，比PPT讲的要复杂一点点。

这个就是高斯模型里面它学习的过程，就是参数估计的过程，最终能不能迭代出两个和出来。

我们讲完混合高斯分布之后给大家一些例子，就是怎么去用？ 第一个就是比较简单的情况，叫背景建模，这已经是很多地方都会用，比如最简单的目标检测，假设教室要检测有多少个学生，比较简单，在大家还没有来的时候拍张照片，大家都坐好了再拍一张照片，两张照片一相减就是大家都出来了，这个是很直观，因为背景是教室不动，多出来的就是情景，那么这是一个情况。

实际上来讲问题会比这个复杂一点，比如说我们要检测人出来，里面有很多的因素，比如说里面有灯光的因素，我们的灯光是25瓦内会闪，就是说你拍的时候是一种灯光，和你人进来的时候频率是不一样的，两个相减的时候，可能大家都出来了，可能还有噪声，背景建模没有大家想象的这么简单，但是思路是一样的。背景建模怎么弄？我们把它理解成，个背景是由几个高斯函数来弄的，我们先举一个简单的例子，这个PPT是一个背景，我们来检测光点，PPT上每一点，虽然投影仪它有闪烁，但是我们让它有两个高斯分布，这时候它每一点上，你都可以把它的量分布统计出来，这时候如果把激光笔放在上面，那么我们的激光笔的值肯定不是高斯分布中间位置，这个是背景概率比较小，背景概率比较小的时候，你就可以认为这个可能是前景。比如要在PPT上检测红点，就可以用我们这样的一个方式去做。这是高斯分布的情况。

怎样去迭代？判定高斯建模的时候，我们是一帧一帧去判定的，每一帧每一帧，每一点的去迭代的，拟了这样的高斯分布，这时候，逐帧迭代的情况。

再往上一页，关于高斯建模的时候，有这样一篇文章，是1999年的文章，当时讲了另外一个例子，这个例子可能更加便于大家理解背景建模要用高斯来做，是什么？比如说我们这个有一棵树，树被风吹，这时候叶子是来回倒，但我们希望这棵树是背景，要不然始终检测是树叶，他认为用高斯分布，表明了这个地方，这段时间被叶子遮挡了，另外时间没有被遮挡，所以有两个高斯函数去模拟，这个是在户外的背景建模的时候，他认为比单高斯好，刚才我提到了比如说室内的部分，大家直观的觉得，虽然灯有闪烁，但用单高斯就够了；在户外的时候，就复杂一些，用多高斯。他们实验里面用的室外的场景，用的是多高斯，k允许3-10以内，更多了意义也不大。

同样给一个例子关于背景建模的例子，左边的这个图，大家看到应该很眼熟，这个是第3教学楼，这个应该是看过的，这个就是背景建模，人都已经出来了，但是我们看到，人里面有空洞，这个空洞就是，一是他穿得衣服跟背景比较像，这个像素被判定为背景，这也是很正常的时候，但如果要判定前景的时候，你要做一些形态学的办法，把人区分出来，这个是背景建模的一个情况。

右边是一个前影情况，这好像在欧洲一个地铁站的情况。这是一个室内情况，很容易人都出来了，大家看到有影子也在里边，所以这时候其实也是有一个很不好的地方，有影子的时候，对这个人要进行一些形态的估计，比如说身高、速度，都会有影响，你不知道他哪个是脚，所以有些研究工作就是专门研究去影子化，这里面也有一些工作，但是都不解决本质问题，因为如果从背景建模或者前景检出的时候，影子的确很难区分。

这个下面是同样的例子，大家看到是一个前景的分割，然后要做一个骨架的识别，做骨架识别一定要把前景分割出来，否认受不了，所以前景分割是其中的一个必要的环节，当然它的数据是深度数据，深度数据比较简单，把背景建出来，把前景的区域弄出来。这个也是我们之前用了混合高斯建模的方式做这样的事情。

背景建模或者前景检出是很多分析中很重要的环节，从今年开始，业界里面突然兴起用这种深度摄像头做监控，大家前几年看多智能监控有很多的文章，最开始的时候都是用光学摄像头去做，然后大家有很多的办法，比如说目标检测、行为分析，还有这种运动的跟踪，这在实际的场合里面其实用的都很少，是因为很难以区分出目标和场景，哪怕现在有很多的技术去做，比如说这里面的光照，阴天，突然出现的车子，从这个高速上过桥，从亮的地方到暗的地方，跟踪这个车是很难的，跟踪都非常费劲。所以在实际的监控中，目前所谓的智能监控在工业中用的还是比较少的，就是说今年的开始大家都用深度摄像头，最重要的是把前景分割出来，这地方应用的场合都在高端的监控里面，包括两个地方，第一地方是银行监控，第二个是监狱，现在银行和监狱都升级了，用这种深度摄像头做这种事情。

还有一个例子，叫目标模型的描述。这个工作本身是用来做跟踪的，但是这里面，也同样用到了混合高斯模型，这个模型在里面怎么样根据它的模型，使你的跟踪更有效，大家不用刻意的明白很多，我们主要看它的框架，看每个框架里混合高斯模型怎么做？

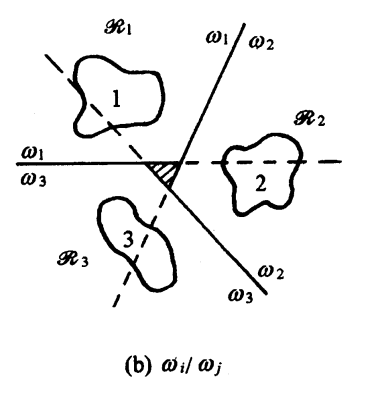
这地方整个用的是局部特征点来代表物体，比如说这个书，里面有很多的局部区域，之前我们在模板里面也讲了，就是说局部区域，它用这种圆来拟这个区域，它是用这个局部的区域来代表目标，希望你在挪动书或者旋转书的时候，仍然能找到这个目标这里面就有一个假设，就说每一个点，你在每一帧在的过程中它变化都挺小，都相似，所以来讲它应该符合一种高斯分布，这时候用这种高斯分布去模型拟合它，完了就可以始终跟踪它。

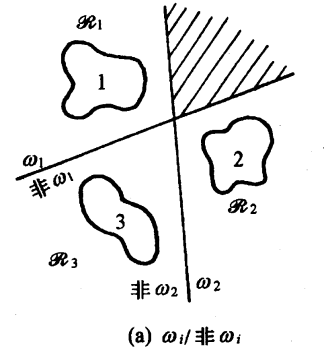
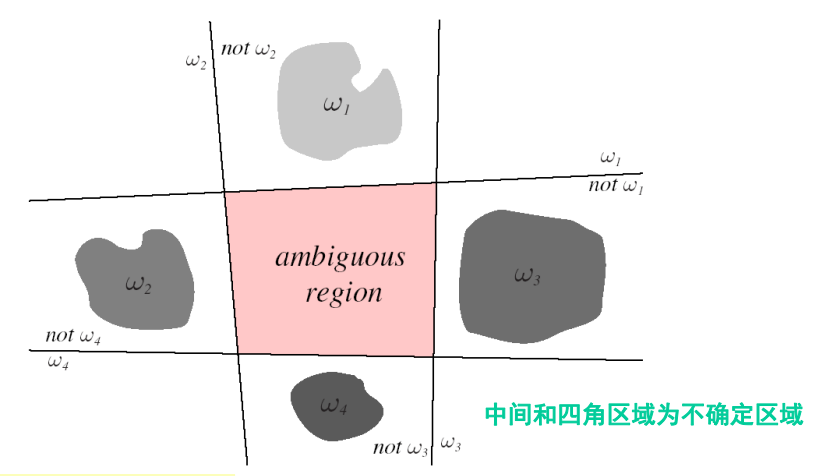
我们来看一下流程，流程就是用到高斯分布讲这样的事情，比如这是一个初始帧，首先跟踪有一个目标，目标是一个特征的提取，局部提取特性，之后目标的模型的建立，建立完了之后，可以看到从进来的头像，这里面同样做一系列的事情，这个是一个模板匹配的过程，接下来是有了这个东西之后，得到一个跟踪，跟踪完了之后你要对你的每个区域做一个更新，更新就是混合高斯模型的模型，在现有的区域更新你每一个特征点的参数，每一个参数它是一个高斯模型，这样混合高斯模型就是建模过程中所用到的一个情况。这个是跟踪的结果，大家看一下这个例子。 在对每一个局部区域，是用合混合高斯进行建模，比如说像这样每一个字符都是局部区域，这时候旋转过程中，其实每个局部区域都可以对应上的。另外一个大家看一下三个目标下面的版块，三个目标大家看到有什么样的共同特点？接近于平面。我为什么说目标都是接近平面的目标？是因为我们看上面的方案的时候，写的里面有用一个ransac建模，那样对一个平面来讲它相对于有一个限制，所以它外面的光环都是四边形。

**多分类问题**

* 1. **基本概念**

这一节我们主要讲多类问题的判别情况。

具体来讲我们看多类s类，其实有两种思路去考虑，第一种思路就是把它变成一个S类的两类，比如我们看这个图，这个是三类问题，可以把它看成三个两类问题，比如讲123的时候，这个1如何，2、3分开，我只考1，这是第一类，我们看跟非1类怎么样，之前我们学的分界面可能是这样的，如果第二类作为单独一类，1、3是一伙的话，分界面可能是这样的。如果3是一类的时候，可能会变成这样，它相当于说是有3个。比如我们仅仅把1作为一类，2作为一类，就变成灰色的区域，或者说阴影的区域，类别判定就出现了混淆。那么以1为主，属于第一类的区域，2类是属于第二类区域，第一组也属于第一类区域，这两个都放在一起了。那么这两个到底属于哪个区域，把它变成S的两个问题，就变成有一点点缺点的区域，所以这个问题就不能简单的把它拓展开，S类问题，我把它变成S两类问题解决，实际上就会在整个的空间里面就出现一些不确定，既是1，又是2，那这时候怎么做的一个问题。所以这个事情是不能简单的把两类问题直接扩散到多类问题上去做。

接下来也是一样的，就把S类问题转化成S个两类问题的时候，同样就会变成正规区域，就变成很不确定，简单的画一下就看到变成了不确定的因素，就不好。简单比喻，在谁与谁交界的地方，有些地区比较偏，正好加在几个省之间，一般管理上就比较乱，一些偏远地区，好像有两个管理在里面，所以这个事情就变得比较复杂了，那么我们这个区域就是类似的情况，这是一个思路。

第二个思路，也挺直观的，我们把两类问题，多类问题怎么向两类问题做延伸。这种思路是我们把S类问题变成CS2，我有S类，那么那两个之间是一个两类问题，所以我们只要去求这么多累的线性函数进行判别，每一个线性函数对其中两类进行分别，我们就可以把两类问题直接运用过来，这个思路会不会好点。同样我们简单去看一下这个思路的问题，我们先看这个图，这是三类问题，我们先看简单一些，比如说三类问题的时候，要用已知的方法去做，比如1、2之间，我可以用这样一个界面，虚幻就过来了。2、3之间的分界面，用以前的数据也学到了，当然1、3之间也有这么一个东西。我们来看一下阴影部分就变成这样了，阴影部分如果说这么以农，在判定这个的时候属于第一类，判定这个属于第二类，判定这个属于第三类，就变成了不知道属于哪个区域了。所以也存在不确定的区域，如果这时候转化成多个两类问题判别的时候，也是会存在一个不确定的因素，所以这是斜线的阴影，是未知的。

所以直接用两类问题的求解办法去求解多类问题的时候，的确会有一些困难，所以使得这是一个问题。这一周我们讲多类问题，怎么去做的问题，我们看多类问题的线性判别函数的话应该怎么去做这样的事情。

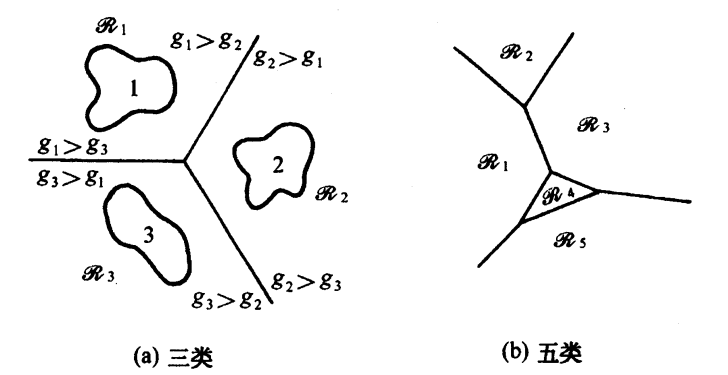
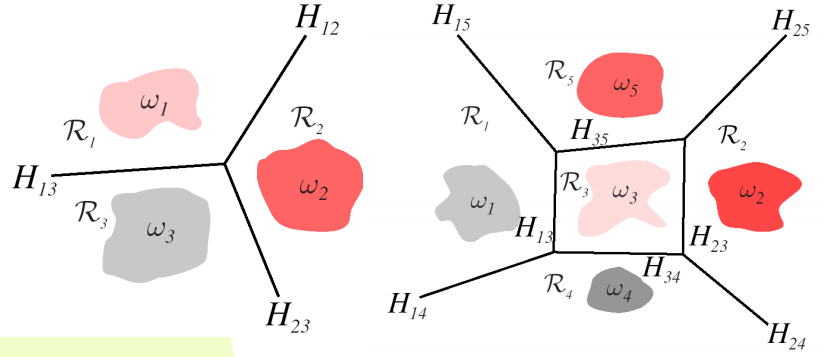
* 1. **线性机器**

在多类问题中，若可用一个线性机器将全部样本正确分类，则称该组样本为线性可分。

这叫做线性机器，定义S个识别函数，di（X）=WIT X＋wio ，i ＝1，2，……S，这其实是我们非常熟悉的，像我们转化到线性识别函数的时候都已经讲了。如果说对于所有的J不等于I，存在drx大于djx，就x属于di，这个意思就是说j不等于i，当你有一个x的时候，你有一个识别函数，那么这个识别函数在di的时候，机械度都比其他高的时候，那么就属于它，这样我们的区域就画出来了。那么在这样的情况下，在整个特征空间分割成s个决策区域，S个决策区域属于R1、R2到Rs，那么s在这个区域里最大，就说ds是最大的，s落在那个区域，这个线性的函数应该是最大的。

那我们来看一下，如果说di和dj的区域相邻，那么这两类的界面就是超平面的一部分，Hij的一部分，这两类间界面的超平面。那么这个超平面的方程跟上面一样，（Wi-Wj）X+（wio－wjo）＝0，这根我们上面的方程也是一样的，同样属于之前的线性两类问题时候的话，差是很重要的，到分界面的距离，其实也是等于你线性方程之间的识别函数差，那么这在之前最小欧式距离提到W，这边就是ds，应该表明X到两类之间的距离，这属于跟前面是一样的，线性机器低。

同样提到权向量的差比权向量本身更重要，二者之间的差比单独的wi或wj是更重要的。那我们看在多类问题的时候，什么叫线性可分，如果在多类问题中，可以用一个线性的机器加全部样本的正确分类，则该组样本是线性可分的，那么这时候就可以看到，每一个类别的话，识别函数都是线性的，那这个线性同时就变成了两俩之间也是分界面，也是线性的，如果说存在这样的情况的时候，就是线性可分的，这是多类问题中的线性可分类的情况。

那么这一块跟我们线思维函数的定义，基本是一样的，只不过把明显的两类问题搁在多类。在这种情况下，看分界面的情况，这时候分界面也智慧有Cs2分界面，最多的两俩之间相间，会有Cs2=s（s－1）/2区域的对，只有相邻区域的话才会有这个分界面。但实际上分界面跟超平面常常少于这个东西，这是一个三类问题，在三类同时判别的时候，我们希望找到这三个分界面，这三个分界面都是线性的函数。那么在这个区域里面ds是最大的，这里面分的对数有Cs2有C3的组合，这上面是分界面的对，这个是没有问题的。

我们看下一个，在1234五类问题的时候，这也是线性可分的，这时候两俩之间的分界面都是这样的情况。完了之后我们看一下，如果说坏的情况有C52的分界面，那么C52分界面就是说十个Ky，C52，这里只有八个，他跟他不相同，所以就不存在五类分界面的问题。实际上分界面一般常常会少于这么一个情况，这是属于基本的概念。

这里五个问题，分界面八个，我们实现了多类问题进行分类，多类问题目标就希望把这些东西找到，这是我们要做的一个事情，也是我们接下来解决的一个问题。为什么希望找到这样的情况，是因为我们的目标是一个线性的机器，这样的话产生的决策区域是凸形的、单连通的，简单、便于分析，并且常适用于单峰的分布，在任何区域都是凸形的、单连通的，便于分析，这是目标，希望找到这样的分布，多类的时候线性机器我们为什么要这么做的原因。

要做这样的事情，我们要用的方式是决策树的一个概念，做这样的事情，这是我们主要讲的一个部分，我们先看一下决策树的基本概念。用它来表示决策和相应的决策结果对应关系，首先它是一个树，树中每一个非叶节点表示一个决策，有了决策就有了结果。所以决策和结果是里面的关键词了。我们会在下一节详细介绍决策树。

1. **决策树**
   1. **基本概念**

我们先看一下决策树的基本概念。用它来表示决策和相应的决策结果对应关系，首先它是一个树，树中每一个非叶节点表示一个决策，有了决策就有了结果。所以决策和结果是里面的关键词了。

那么树里面的每一个非叶节点表示的决策，该决策的值导致不同的决策结果，这个不同的结果在特殊空间里面进行划分，会影响后面的决策选择，导致不同的结果产生不同的叶节点情况。所以在常用过程中，我们把多类分类问题，把它转化成分层次的两类问题去考虑，也不是说分层次就一定是两类问题，因为分层次三类问题和分层次四类问题可以，但一般来讲分层次的两类问题比较简单，一分为二去做这样的一个事情。

那么这时候，决策的基本思路，看上去都不会很难，整体的想法不要依次把问题解决掉，不要想着一口吃成胖子，不要一次就搞定，但也不要一天就变成瘦子，这是一个持之以恒的过程，不要想着一次就搞定所有的问题，不要试图用一种算法在N维特征空间里面把所有的问题都搞定，每一次使用一个特征，或者一个特征集，把当前的模式情况一分为二，变成两个候选子集，这样的做法把它相对变成两类问题，有多类问题的时候把它分一分类，还是把它当成两类问题去做，分完之后，其实总体上要达到那个目标，在你的子集里面，比原来的问题要复杂一点，稍微的直观一些，你要面对的是十类问题的时候，做一次切分，可能属于两个五类问题，那是会好一些，每个五类问题再做就好很多，这是一个目的。每一次我希望把你解决问题的对象分一分，分开来做，这样解决较为简单的一件事情。

也就是说依此类推，直到把所有的类分开，希望每一个候选子集只有一个候选的模式类，这样的话就比较好了，如果单有决策树来讲，我们希望每个叶子节点，它就属于某一类的，这样的话就比较简单，最后你只要走到这一类的时候，就知道它属于哪一类。

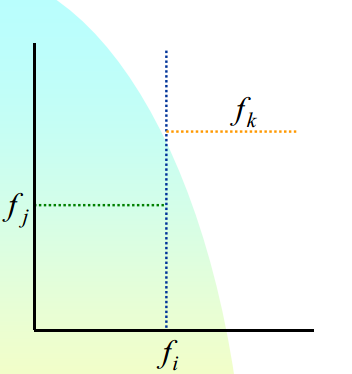
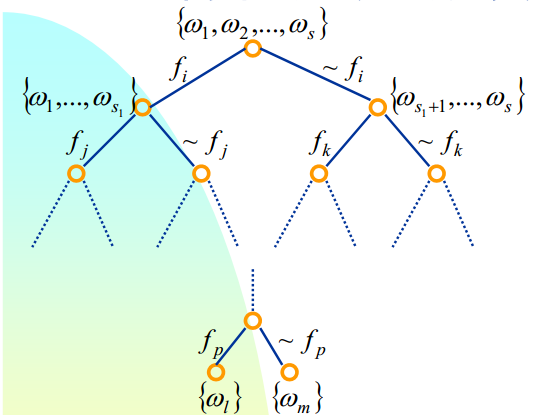
实际上我们后面提到的情况没那么简单，再怎么决策树的时候，叶子节点都不可能只有某一类，还是若干类的混合在一起，这样的话应该怎么弄，后面还会讲怎么解决。

我们看决策树的方法，它的历史可以稍微看一下，我们在讲的时候，希望给大家知道历史，知道它的传承包括发展的情况。他的方法来讲，1966年开始了决策树的算法的提出，这时候提出了CLS的学习系统。那么1979年的时候，一位学者给出ID3的算法，I是迭代，D是二分，其实来讲是二分的方法，也就是每次分成两类，那么在1986年的时候，进行总体的简化，其实作为一个比较典型的算法，后来有ID4，包括ID5算法，最后还有个ID3直接变成C4.5的算法。另外叫CART算法，CART算法本身来讲它是决策树，可以分为五个理论，我们会在第三小节详细介绍CART。

* 1. **二叉树分类与分段线性分类器**

我们看一下简单二叉树分类，是怎么样的一个过程。假设我们要解决的是一个s个分类问题，如果通过s分类会分成这样的情况，至少说目标要做分类的情况。那么我们来看一下，比方说这个问题，我们可以看看它的一般思路，也就是说我们看一下跟这是否具有特征，分成两个子集，比如看看这里面候选特征，有没有模糊特征，有没有Fi，那么把整个的类别分两个子集，这个时候我们就把问题一分为二，看到这个的时候，可能有同学就会问，我怎么判断他有这个特征，没这个特征，其实一个比较简单的方式，大家都有这个特征，但是他的尺度不一样，在某一唯独上，比如S1类里面，这个尺度比较大，另外比较小，概念上来讲，我们看有没有这样的一个特征，这属于一种解释，但实际上在做的过程中是很简单的一种方式。再根据看是否有这个特征集Fj，把前面一小戳的子集再分成两个子集，那么相当于在这个地方来讲，我们就不考虑这么多，同理我们也可以看，不断递归的感觉，我把这个问题变小再去划分一下，这时候看到它的类别就变小了，那我们根据这个情况再一分为二，因为你的类别数是有限的，所以每一次都在变小，很显然每一次一分为二的时候，比如在这个节点上，只有两类问题的话，很简单，找一个特征就把它区分开。也就是说每一个子集只有两类再画一次就结束了，相当于就不再往下递归拆分了。

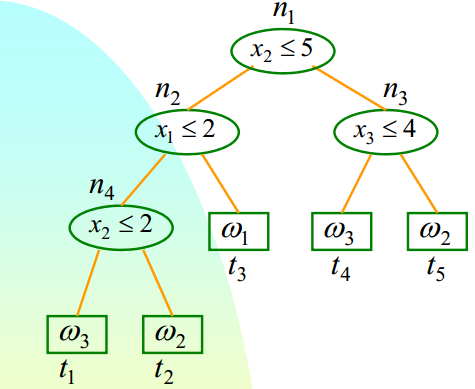
第二比如上面的时候考虑到某一个特征，下面能不能考虑同样的特征呢？完全是可以的，所以不同层次的特征可以相同的，不是说前面用了后面就不能用，这个的话完全可以用，并且这边用了，这边也可以用，这是没有关系的。所以特征可以相同，也可以不同，如果能找到同一层特征都相同的话，这里面的话，就是说我们要把S真的要划分出来的话，至少要log2S，这是很简单的，一分为二 一分为二，我们至少要这么做才能达到每个叶子节点上最后只有某一类，这是最少的一个可能性，大家要注意最少的特征，接下来可能比这要多，这不是说属于完全平衡数才可以。

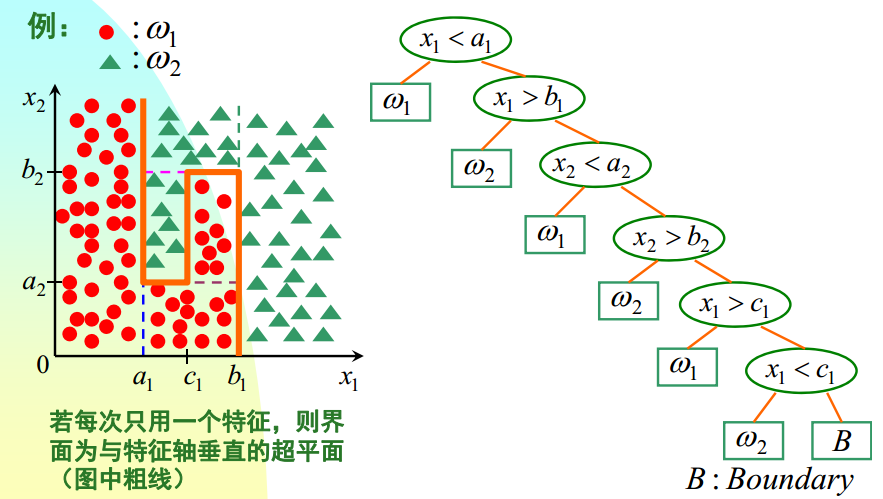
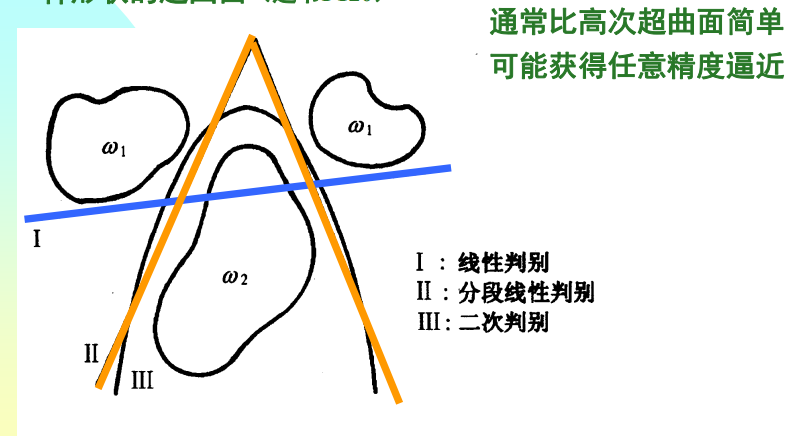
这是最简单的情况，那么从概念上再看一下，二叉树分类器对特征空间的划分到底是什么样的情况，其实很简单。就是说我们可以把它看成是在特征轴垂直的超平面划分特征空间，每一次一分为二的，这是一个示意图，那么我们刚才看到了，提到S问题的时候，我们刚开始说把整个特征空间给划分，我们可以看到Fi的特征，比如Fi小于它还是大于它，我们就把整个空间给划分出来了。比如在左边，我再举一个Fj的特征，这个也是大于它小于，同理Fk大于它小于它，你可以看到它是在多维空间里面有一个超平面进行切分，最后我们把整个的空间切分很多小的，如果是理想状况的话，最后每个小区域就是属于这一类，这是比较好的情况，也是我们希望达到的情况。

其实关键就在于能不能找到相应的特征，把所有的类别能够分开，我们能不能通过学习找到这样的一个特征，这个特征把它区分开，这是一个问题，我们怎么样能够找到这个特征。

同时还有一个问题，比如我们举了很多的特征，但是我们怎么样去挑选好的特征，能够把它给区分开，这也是我们要做的事情。那我们来看一下例子，这个例子跟刚开的示意图非常的接近，只是说我们把它用实际的例子讲出来，实际的数填进去看一下这样的情况，那我们看一下123，三类的分类问题的情况，每一次选一个特征把它一分为二，确定一下搜索的类别，我们来看一下这样的情况，比如有三类的问题，我们看取第二类特征，取第二类特征的话，这是X2小于等于5，那么在第二特征上，比如说这个值小于5的时候，在左边，在大于5的时候，在右边，这么一弄就可以看一下左边的节点可能有一些类别，右边是一些类别，那我们看一下第一类的特征，X1小于等于2，最右边是X3小于等于4，如果在这样的时候，我们先看左边，X1小于等于2的时候，又在第一个维度上，相当于把这个空间也切分一下。如果X1大于2的时候属于第一类，这是一个叶子节点。满足这样的情况只有第一类的，那么第一类区间什么的，就说X2小于等于5，同时X1大于5，这样的空间都属于第一类。

那么这边X3的问题，我们具体细分，这是有的特征跟叶节点用的特征是一样的，同样可以用，第二类同城X2小于等于2，这就比较好了，这是第三类第四类，大于2的时候是第二类，小于等于2是第三类，那么这一部分我们就搞定了，123都在这儿。

那左边的情况同样都包含了123，那么这样的做法，希望能够把这个问题简化，右边只有第二类、第三类，那么通过第三类的特征去看，第三类小一点第四，这时候大于四就是第二了，那么可以看到左右分类，是有可能重叠的，这边也有第二第三，这边也有第二第三。形成一个问题，你只要这样去做的话，最终每一个叶子节点上是有某一个领域就可以了，我们可以看到能够得到X的一个类别，我们很迅速的知道他属于某一类，这样的话就是一个决策树的情况，

决策树简单的做法也跟大家介绍一下，他可能运用在什么地方，完了之后再介绍具体的算法，再给一些实际的例子。刚才提到用决策树的方式，或者说树状分析，每个节点采用的都是线性的判别准则，很方面，进行多类的分类。实际上用这样的树结构，其实还可以做，哪怕是两类问题，我们也可以做，多用两类问题的话，也是一个分段线性的分类，那我们来看一下这样的问题，这是示意图。

如果说红色的原点和绿色的三角形，搁在以前不好弄，不是说没有，可以同一线性，或者什么样的方式去做这样的事情。实际上来讲，这种分布还挺有规律，就是说我们不要指望用一个简单的事情就搞定了，我们可以一步一步的来，那我们来看这样的一个情况。对这两类问题，原来我们用树状分类器来做这样的事情。比如这里面X1小于a1，那么这时候左边属于第一类，我们就不管它了。我们看右边，专门考虑右边的情况怎么样。大于a的时候，两个混杂在一起。那我们再找一个值，比如x1大于b1，同样是第一个维度，但相对来讲，再选一个余值，大于b1的时候，这边就变成叶子节点，变成第二类。同样这一部分区域显然是x1大于a，和x1小于b，就是这么一个分类。相当于就变成这么一个圈，带状的部分，那我们可以继续把这个问题进行简化了，简化的时候就说x2小于a2，我们看x2小于a2的时候已经分好了，也是属于第一类，不管怎么样，把这个东西变得更简单了。

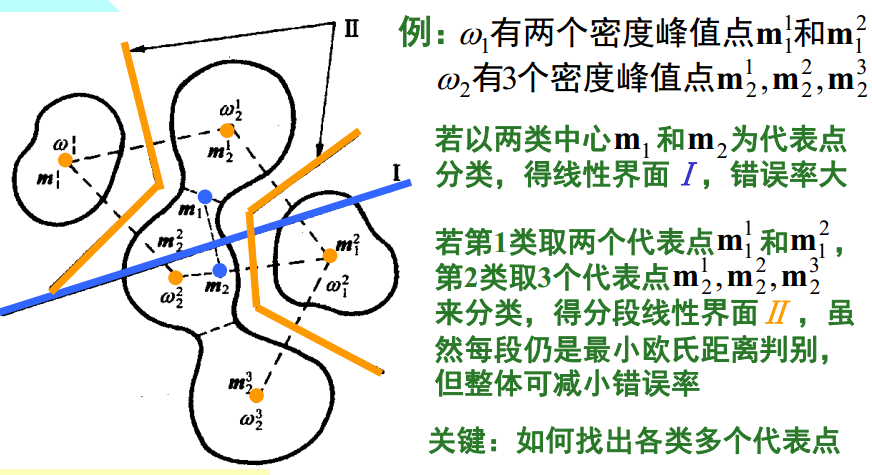
那这时候x2大于b2，再把这一部分抛去，在这一区域分好的时候，单看这个就变成线性了，咱们这一个东西x1小于c1，这时候这一块大于这一块。这时候我们可以看到分界面，通过树状分类器我们得到这么一个分界面，首先也是非线性的，但是每一个都是分段线性，也就是分段线性来做这样一个分类问题的话，我们可以把处理原来不太好做的事情。所以我们可以看到，决策树和树状分析我们仍然可以做到很好的结果。

那么也就是说分段线性函数的定义，决策面是由若干段超平面组成，用于逼近各种形状的超曲面，那么这是我们希望做的一个事情，通常比高次的曲面简单，并且可以获得任意精度的逼近，那么任意的精度也特别简单，比如说算圆的面积或者说算圆周率的时候，最开始都这样一环，用多边形来分析这个圆，然后这个pai（音）等于多少，这一块其实是很好理解的。那么就是说用分段线性来去做分界面的话就非常简单。

可以举这样一个例子，这也是三类分类的问题，这个看上去比较完美的分界面，我们求这个就比较麻烦了，怎么去做这样的事情比较麻烦。那么这时候我们这样去做就很简单，我们可以看到这个问题，也是我们接下来要讨论的问题，虽然是两类的问题，一类和两类，但是比较不凑巧的事情是说，第一类不是一个简单连续的，是分成两坨了，如果原来的算法是标准的人工法去替代的话，那么有问题，那么这是一个理想的状况，理想的状况，求解是不难求，比如现在他是一个局面，比如三次或者更高次都有可能，求起来比较麻烦，那么这时候用这么一个方式去分开的话是比较好，相对来讲是两段线性替代一个二次型曲线。

那么我们可以看一下类似这样的问题，我们怎么用树状的分段线性去做这样的事情。这都非常简单，基于到类中心距离的分段线性的识别，这都是概念性的，很简单，我们过一遍，然后举具体的例子就可以了。

最简单就是欧式距离的分类的判别，有一个标准模版进行中间线的切分，进行一个分类，这是我们开始在做这样的一个事情，这个适用于贝叶斯概率的判别情况是正态分布、协方差矩阵相同、各个特征独立且方差相同、各类先验概念相同，这是属于最小欧式距离做的事情。

那我们看下一个，其实在样板分布的时候，其实并不是说单峰分布，可能是多峰分布，也不见得是正态分布，那我们看示意图，这比刚才稍微夸张一点点，同样是属于两类的问题，那么第一类是包含这么一个区域，正态分布的时候。但是第二类的话看的就更加奇形怪状，感觉是故意的。那么这时候要做两类的判别，就变得复杂了一些，如果从概念上讲，我们想做的事情把它变成分段，比如第二类的话复杂，我们最好是把它切成三个代表，原来就是一个代表，就是说小状形以它为代表，这边不同意，这边也不同意，那么这时候分成三个代表，大家的利益得到一个平衡。同样第一类也是说你的标准模版有它，有它，两个代表，那么这时候去做，我们如果说把它变成了二类、三类问题的时候，就显得比较简单一点。

首先这和他之间的分类是这个，他和他之间的分类是这个，那么同理这边和这边的区分，可以把这个折线给求出来，那么这也是分段线性的概念，这是我们理想的状况，要做这样的事情。关键把这些代表给找出来，这些代表是谁，不能够随便说，那么这是其中的一个关键点，那么这个代表怎么找出来，其实也有一半儿，在讲到聚类的时候应该去想，就是说像这样哪怕是有标签，我们也做聚类，这可能是一个办法。

那我们把刚才的说法作为程序上或者步骤上进行一个归纳，首先通过聚类的方式，把多峰的分布分成若干个子集，比如把di1，把它变成Li一类，Li子集，看不见的区域也切分一下，就分为Li的子区域，那么对于每一个子类取一个代表点，这样的话，可能他们有一个标准的模版，那么对于每一个子类的话，定义它的识别函数，这时候就变成识别函数，同时定义一个判别的规则。哪怕是两类问题，你的分布相对来讲不是简单，也许比较复杂的时候，你可以把它分成多个子区域，那么多子区域完了之后，就把它变成一个由树状分类去做的事情，相当于求每个子类的情况，同样我们可以用树状分类的方式去解决。

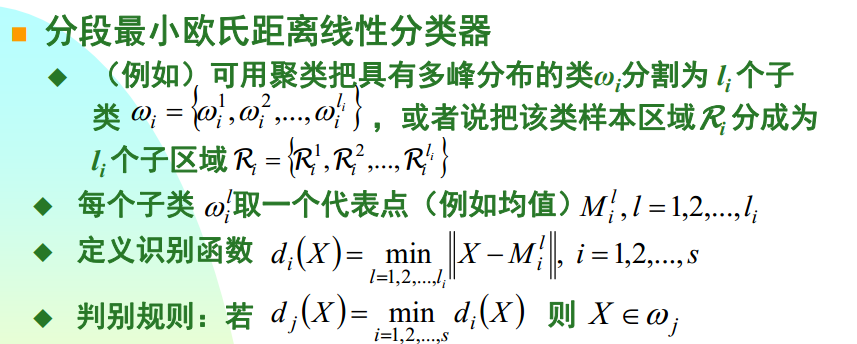
当然这里面应用出来是在字符识别，不同的字体，从它的特征上来讲，有比较大的不同，那可以把不同的字体作为不同类去判别，分完之后，不同的字体硬要把它放在一个类别的时候，那个类别就比较大，不好操作。这里面这样本质还有一些希望，如果你把它变成一类的时候，怎么样切分原始的情况本身，类面具也就会比较小，这是整个的基本思路。

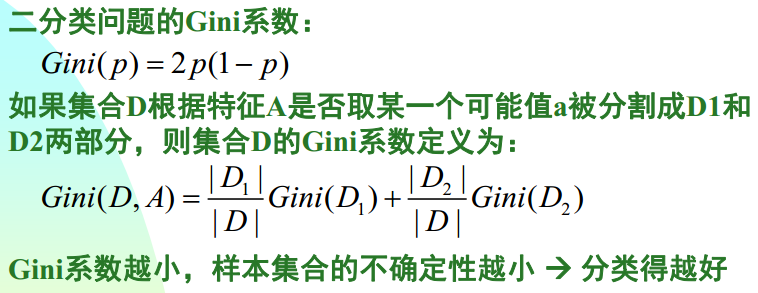
* 1. **ID3和C4.5**

李航统计学习方法P63

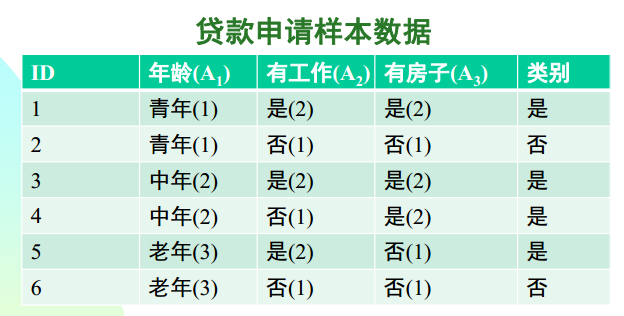
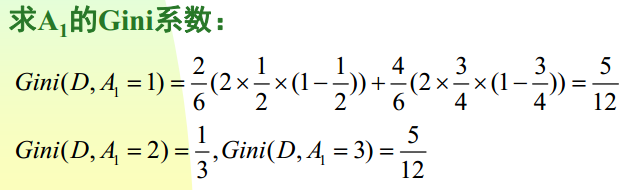
* 1. **CART**

分类回归树算法：CART(Classification And Regression Tree)算法采用一种二分递归分割的技术，将当前的样本集分为两个子样本集，使得生成的的每个非叶子节点都有两个分支。因此，CART算法生成的决策树是结构简洁的二叉树。CART的算法，刚才提到CART本质来讲，就是（01：32：14），这是一个缩写。那在每个节点上，都采用这个划分法来划分子节点，那么也就是每一次都寻找一个最优的特征，最优的特征来分辨整个数据，使得最小化的叫基尼系数，叫分类树，如果这个是准则的话，你的决策，最小化，你的Gini系数的分类树。或者说最小化你的平方数差（音），就是回归树，最好平方数差属于连续的，所以这时候的话就是回归树。所以他既可以用于分类问题，也可以用于回归问题。

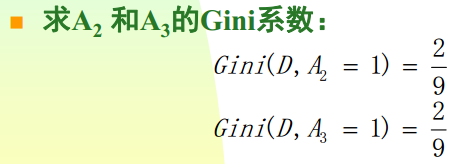
当然从整个概念上来讲，为了减少过离合的话，可能会需要一些剪枝而控制决策树的规模，这些术语，在本科的时候学树结构应该都会有，因为之前提到的决策数据开始就是个树，关于树的概念蛮多都用得到，树的思维结构，我相信大家都应该知道，剪枝是什么样的情况，剪枝有集中方式，先剪枝和后剪枝的概念。从这个来讲其实都是一些后树处理的方式，这个地方我们都不去详细讲，我们先讲本身算法的核心内部。

最核心的内部提到什么叫Gini系数，这个概念虽然简单，但我们还是要提一下算法，就是说基尼系数本身来讲就是一个公式，比如说一个饭店的基尼系数就是这么一个公式，属于这一类的概率就是P，不属于就是1-P，基尼系数二分布里面一个概率的函数，就是2P（1-P）。本身来讲这是一个计算的方式，所以很简单，叫基尼系数。

如果我们看集合D根据某个特征A是否取某一个可能的值，被分割成两个子集，它是二次函数的分类，所以它根据某一个特征，把每个子集分成两个，比如说第一部分和第二部分，那么分完之后，整个结合了D的基尼系数，这是一个数据集，A的话是说取某个特征的决策，算它的基尼系数，那么基尼系数本身是怎么算的，D1/D，也就是你把它划分完之后，第一第二分别有一个基尼系数，完了之后，再乘上它各自的权重，那么就变成A的角色情况怎么样，这是一个概念。那么你的目标选择，就是说基尼系数越小，样本的不确定性越小，分类得越好。其实你的目标是说不确定性是越小越好，完全不确定分类就结束了，这个其实是说你整个目标是希望你的系数越小越好。如果是转化到信息论角度来讲，基尼系数跟信息类是很接近的，是一个logP，是一样的道理，概率越小，上面越大，上面越大就是信息量越大，就给你带来激情，一个偶然的发生，比如说中奖，信息量很大，一下子成富翁了。比如说买一个毛巾没什么意思，中不中奖都无所谓。我们要做的事情，就是希望比较确定，基尼系数越小越好。

我们来看一下基于基尼系数的分类树怎么做的，买房子或贷款吧，银行最现实的，什么样的人能贷款，什么样的人不能贷款。银行肯定要设计一个表，就是说你申请贷款的时候，我怎么样能够快速判定给谁放贷，这也是一个算法。那么我们用了最简单的情况来说，第一个年龄，其次工作、房子、类别，这个例子去银行的很少，六个，我们区分成两类，年轻人、中年人跟老年人，有工作的没工作，有房子没房子，有三个特征做输入，最后来判定。那么在这种情况下做训练的时候，首先来了一个样本之后我们怎样去划分，我们做训练的时候，我们得先看一下（01：38：32），怎么弄。我们看A1的特征，那么A1能取的值有三个，A1可以等于1，可以等于2，等于3，那我们先看一下A1等于1，A1等于2和等于3，那这都是同一计算。那么A1等于1的时候是怎么算的？相当于把整个的样本一分为二，青年还是非青年，那么就说青年是两个人，这是2和6，那么在这个子集里面，两种可能性，yes或no，能贷款或不能贷款的都是二分之一，那么这时候两倍的p乘以p，就算完了。右边样本4/6这相当于比重，就说在这四个里面，能贷款的是3，不能贷款的是1，就是四分之三和四分之一，那么这时候2×3/4×（1-3/4），等于5/12，这个就算算完了。那么A1等于2的时候是三分之一，A1等于3的时候是5/12。很显然由A1来做决策的时候，很可能A1等于2的时候，它的基尼系数最小，是4/12。

那么下面我们同样去算，A2等于1，因为这里面A2要么取1，要么取2，那么A2等于1的时候，它划分之后是九分之二，A3等于4的时候也是九分之二，这时候把A2作为最优特征去划分，也可以，也可以把A3作为划分，都没有问题。那这是从数值上看，从现实的角度来看隐含的是什么呢，先看一下这个哥们儿有没有工作，有工作的话再说，或者说看这哥们儿有没有房子，有房子也可以容易划分，其实就是这个意思。

比如我们取A2等于1的方式来看，A2等于1的时候，是的话就可以贷款了，就是说有工作就贷款，比较符合实际情况。那么没工作的时候，变成区间2、4、6，把整个区分开了，也就有工作的可以放贷，没工作的还要考虑。2、4、6一样的变少了，只有三个，这时候跟前面一样的算法，那么这时候可以看到的是A3，A3等于1的时候就等于房子，有房子的时候可以看2、4、6，有房子的话都贷款了，没房子的都不贷，这也比较简单。其实总结来讲，好像跟年龄都没啥关系，第一看你有没有工作；第二看你有没有房子，如果满足这些条件就可以贷款。那么通过这个样本我们可以得到这样的一个特征，看出了建模的过程。

我们把决策树单个来进行一下讨论，决策树简化特征空间的划分，判别简单明确，用这样的方式比较好，把空间一分为二，可以逼近一个曲线，比较好做。同样便于用硬件或软件的实现，微软。对于多类的问题，也能提高分类的速度。如果对人大样本库人脸识别就可以做很多类的事情，粗分类和细分类，这里面还是有一些问题。比如说树结构的设计，特征的选择，分类规则都不是特别的容易，比如说特征的选择是非常重要的，树的结构，包括后续的剪枝，也是比较复杂的过程，整个过程还是有难度的。

1. **随机森林**
   1. **基本概念**

首先我们要先介绍Bagging方法，Bagging方法是ensemble methods中获得用于训练base estimator的数据的重要一环。 正如其名，Bagging方法就是将所有training data放进一个黑色的bag中，黑色意味着我们看不到里面的数据的详细情况，只知道里面有我们的数据集。然后从这个bag中随机抽一部分数据出来用于训练一个base estimator。抽到的数据用完之后我们有两种选择，放回或不放回。

既然样本本身可以bagging，那么feature是不是也可以bagging呢？当然可以！bagging完数据本身之后我们可以再bagging features，即从所有特征维度里面随机选取部分特征用于训练。在后面我们会看到，这两个‘随机’就是随机森林的精髓所在。从随机性来看，bagging技术可以有效的减小方差，即减小过拟合程度。

通俗来讲，bagging + decision trees，我们得到了随机森林。将决策树作为base estimator，然后采用bagging技术训练一大堆小决策树，最后将这些小决策树组合起来，这样就得到了一片森林(随机森林)。随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。

* 1. **算法流程**

随机森林的建立，基本就是两个步骤：随机采样与完全分裂。

（1）随机采样

首先是两个随机采样的过程，random forest对输入的数据要进行行、列的采样。

对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为N个，那么采样的样本也为N个，这选择好了的N个样本用来训练一个决策树，作为决策树根节点处的样本，同时使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现over-fitting。

对于列采样，从M个feature中，选择m个(m << M)，即：当每个样本有M个属性时，在决策树的每个节点需要分裂时，随机从这M个属性中选取出m个属性，满足条件m << M。

（2）完全分裂

对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。分裂的办法是：采用上面说的列采样的过程从这m个属性中采用某种策略（比如说信息增益）来选择1个属性作为该节点的分裂属性。

决策树形成过程中每个节点都要按完全分裂的方式来分裂，一直到不能够再分裂为止（如果下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到了叶子节点，无须继续分裂了）。

我们用LearnUnprunedTree(X,Y)表示生成一棵未剪枝的决策树的过程，以下简写LUT (X,Y)：

LearnUnprunedTree(X,Y)

输入：

X是RxM的矩阵，Xij表示第i个样本的第j个特征。

Y是Rx1的向量，Yi表示第i个样本的类别标签。

输出：

一棵未剪枝的树

如果X的所有样本值都相同，或Y的所有类别标签相同，或者R<2，则产生一个叶结点，该结点的类别即是X中最多数的类别。

否则

从M个特征中随机挑选m个

这m个特征中，信息增益最大的记为p。（信息增益的计算方法见下文）

如果特征p的取值是非连续的（如性别：“男”，“女”）

则对p的任一取值v

用Xv表示特征p取值为v的样本，Yv为其对应类别

Childv =LUT(Xv,Yv)

返回一个树结点，在特征p处分裂，孩子的数量与特征p的不同取值数量相同。第v’个孩子即是Childv = LUT(Xv,Yv)

如果特征p的取值是连续的（如温度，长度等），设t为最佳分裂阈值

XLO 表示 特征p的值<t的样本集合，YLO为其对应类别

ChildLO = LUT(XLO, YLO)

XHI 表示 特征p的值>=t的样本集合，YHI为其对应类别

ChildLO = LUT(XHI , YHI)

返回一个树结点，在特征p处分裂，有2个孩子，分别是ChildLO = LUT(XLO, YLO) 和ChildLO = LUT(XHI , YHI)。

首先，以上是未剪枝决策树的生成过程，一般很多的决策树算法都会包含剪枝过程来避免over-fitting。但是由于随机森林的两个随机采样的过程保证了随机性，所以就算不剪枝也不容易出现over-fitting，这也是随机森林的优势之一。

其次，按上述办法生成的每一棵决策树的分类能力很有限（从M个feature中选择m让每一棵决策树进行学习），但是组合在一起形成森林之后分类能力就大大加强了，这点很像adaboost里面的弱分类器组合成强分类器的思想，并且最后都是通过带权重的方式组合起来。

最后，随机森林有2个参数需要人为控制，一个是森林中树的数量，一般建议取很大。另一个是m的大小，推荐m的值为M的均方根。

所以我们可以总结随机森林的优点如下：

（1）比较适合做多分类问题，训练和预测速度快，在数据集上表现良好；

（2）对训练数据的容错能力强，是一种有效地估计缺失数据的一种方法，当数据集中有大比例的数据缺失时仍然可以保持精度不变和能够有效地处理大的数据集；

（3）能够处理很高维度的数据，并且不用做特征选择，即：可以处理没有删减的成千上万的变量；

（4）能够在分类的过程中可以生成一个泛化误差的内部无偏估计；

（5）能够在训练过程中检测到特征之间的相互影响以及特征的重要性程度；

（6）不会出现过度拟合；

（7）实现简单并且容易实现并行化。