模式识别

题　目	最大似然估计与贝叶斯估计
学　院	信息科学与工程学院
专　业	控制科学与控制工程
姓　名	任静
学　号	Y30190741
指导教师	赵海涛

2019 年 12 月

目录

最大似然估计与贝叶斯估计

摘要

在求解典型的模式识别问题中，估计先验概率通常没有太大的困难，最大的困难在于估计类条件概率密度。其中涉及两个主要的问题，首先是在很多情况下，已经有的训练样本数通常比较少；然后是当用于表示特征的向量 x 的维数较大时，就会产生严重的计算复杂度问题。为了降低问题的求解难度，通常我们会事先知道参数的个数，并且先验知识允许我们把条件概率密度进行参数化。

参数估计问题是统计学中的经典问题，目前已经有一些具体的解决方法。通过对模式识别的学习，本文中我主要讨论两种最常用并且很有效的方法，即最大似然估计和贝叶斯估计。虽然这两个方法得到的结果是一致的，但是这两个方法的本质却有很大的差别。最大似然估计和其他的一些类似方法相同，都是把待估计的参数看作是确定性的量，只是其取值未知。最佳估计就是使得产生已观测到的样本即训练样本的概率为最大的那个值。与此不同的是，贝叶斯估计则把待估计的参数看成是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，这样就利用样本的信息修正了对参数的初始估计值。在贝叶斯估计中，最为典型的效果就是:每得到新的观测样本，都使得后验概率密度函数变得更加尖锐，使其在待估计参数的真实值附近形成最大的尖峰，这一过程被称为“贝叶斯学习”过程。无论使用何种参数估计方法，在参数估计完成后，都要使用后验概率作为分类准则。

本文对最大似然估计和贝叶斯估计这两种估计方法进行了梳理与总结。

1 引言

最大似然估计（*maximum likelihood estimation, MLE*）一种重要而普遍的求估计量的方法。最大似然法明确地使用概率模型，其目标是寻找能够以较高概率产生观察数据的系统发生树。最大似然法是一类完全基于统计的系统发生树重建方法的代表。这个方法最早是遗传学家以及统计学家罗纳德·费雪爵士在 1912 年至 1922 年间开始使用的。似然是对 likelihood 的一种较为贴近文言文的翻译，“似然”用现代的中文来说即“可能性”。因此，若称之为“最大可能性估计”则更加通俗易懂。

给定一个概率分布 D，假定其概率密度函数（连续分布）或概率聚集函数（离散分布）为 $f(D)$，以及一个分布参数 θ，我们可以从这个分布中抽出一个具有 n 个值的采样，通过利用 $f(D)$，我们就能计算出其概率：但是我们可能不知道 θ 的值，尽管我们知道这些采样数据来自于分布 D。那么我们如何才能估计出 θ 呢？一个自然的想法是从这个分布中抽出一个具有 n 个值的采样 $x_1, x_2, ..., x_n$，然后用这些采样数据来估计 θ，一旦我们获得就能从中找到一个关于 θ 的估计。最大似然估计会寻找关于 θ 的最可能的值（即在所有可能的 θ 取值中，寻找一个值使这个采样的“可能性”最大化）。

这种方法正好同一些其他的估计方法不同，如 θ 的非偏估计，非偏估计未必会输出一个最可能的值，而是会输出一个既不高估也不低估的 θ 值。要在数学上实现最大似然估计法，我们首先要定义可能性：并且在 θ 的所有取值上，使这个函数最大化。这个使可能性最大的值即被称为 θ 的最大似然估计。注意这里的可能性是指不变时关于 θ 的一个函数。最大似然估计函数不一定是惟一的，甚至不一定存在。

贝叶斯分析方法（*Bayesian Analysis*）是贝叶斯学习的基础，它提供了一种计算假设概率的方法，这种方法是基于假设的先验概率，给定假设下观察到不同数据的概率以及观察到的数据本身而得出的。其方法为：将关于未知参数的先验信息与样本信息综合，再根据贝叶斯公式得出后验信息，最后根据后验信息去推断未知参数的方法。

在贝叶斯统计理论中，统计推断中的相关量均作为随机量对待，而不考虑其

是否产生随机值。概率被理解为基于给定信息下对相关量不完全了解的程度，对于具有相同可能性的随机事件认为具有相同的概率。在进行测量不确定度的贝叶斯评定时，与测量结果推断或不确是度评定相关的每一个物理量均被分配一个随机变量，分布宽度常用标准差表示，反映了对未知真值了解的程度。

按照贝叶斯理论，与测量或相关评定工作有关的每一个物理量均被分配一个随机变量，尽管每一个估计量和它所表示的相关被测量是不相同的，但它是用来估计被测量的待定真值的。为了简单起见，估计量的值和该被测量均用相同的符号表示，如用表示样本，同时也用它表示样本值，这可从上下文区别，不会发生混淆，因为样本是随机变量，而样本值是一些常量，这与经典统计理论是不同的。

2 最大似然估计

最大似然估计通常在训练样本增多时具有非常好的收敛性，同时，最大似然估计通常比其他方法要简单，很适合应用在实际过程中。

2.1 基本原理

假设我们有c个样本集，即$D_1, D_2, ..., D_c$，其中任意一个样本集D_i中的样本都是独立的根据类条件概率密度函数$p(x|w_j)$来抽取的，因此每一个样本集中的样本都是独立同分的随机变量。假设每一个类条件概率密度函数$p(x|w_j)$的形式都是已知的，其未知的部分就是具体的参数向量θ_j的值。因此，一旦知道了参数向量θ_j的值，那么整个类条件概率密度函数也就确定了。

为了简化对问题的处理，总是假设属于类别D_i的训练样本对于参数向量θ_j的估计不提供任何信息。也就是说，假设每个参数向量θ_j对它所属的类别起的作用都是互相独立，互不影响的。因此每个参数向量只对自己的类别中的样本起作用，这就允许我们对每个类别可以分别处理，同时也使得记号得以简化，因此在这种情况下，用于表示不同类别的下标可以省略。在这样的假设条件下，我们将有c个独立的问题，其中的每一个问题都可以表述成下列形式：已知样本集D，其中每一个样本都是独立的的根据已知形式的概率密度函数θ抽取得到的，要求使用这些样本估计概率密度函数中的参数向量θ的值。

假设D中有n个样本：$x_1,x_2,...,x_n$，且样本独立抽取，则下式成立：

$$p(D|\theta)=\prod_{k=1}^{n}p(x_k|\theta) \qquad (1)$$

根据定义，参数向量θ的最大似然估计就是使上式达到最大值的那个参数向量θ，可由图 1 进行直观的理解。

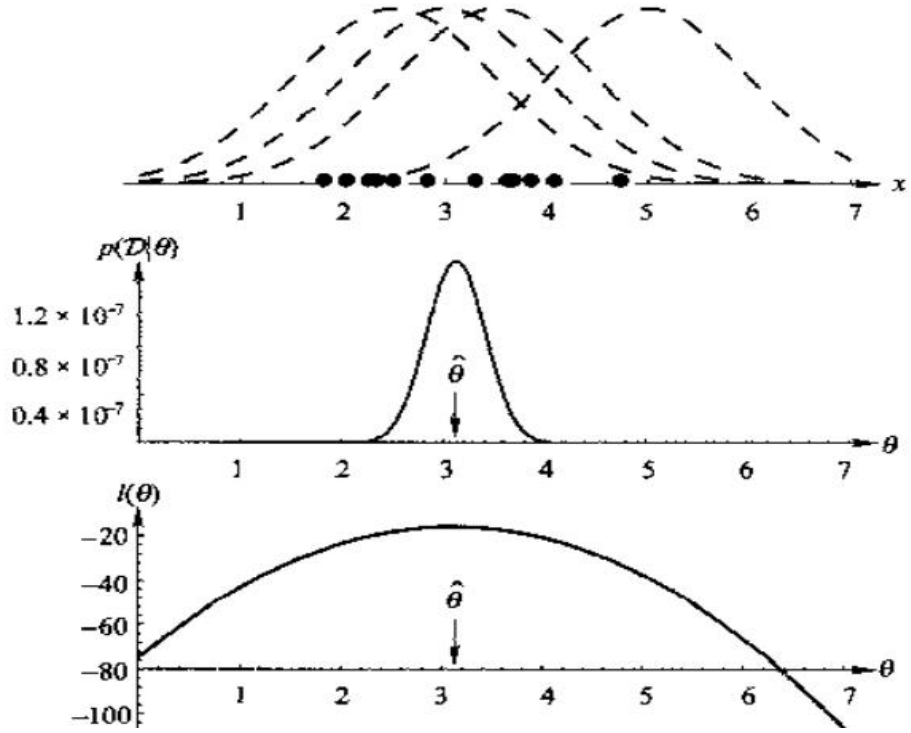

图 1 图像函数

位于最上方的图显示了一维情况下的一些训练样本，而这些样本都服从一个方差已知，而均值未知的一维高斯分布。位于中间的图显示了似然函数$p(D|\theta)$关于均值的图像函数，使得似然函数取得最大值的点标记为$\hat{\theta}$，这个使得似然函数取得最大值的点，也是使得在最下方的图中所示使得对数似然函数$l(\theta)$取到最大值的点。注意，对于似然函数$p(D|\theta)$和条件概率密度函数$p(x|\theta)$，虽然他们看起来相像，但是似然函数$p(D|\theta)$是一个关于θ的函数，而条件概率密度函数$p(x|\theta)$却是一个以θ为参数而关于变量x的函数。最下方的图中曲线下的面积没有什么实际的意义。

如果$p(D|\theta)$是一个可微函数，那么这个的值就可以用标准的微分运算来求

得。如果实际的待求参数的个数为p，则θ参数向量θ可以写成如下的向量形式：

$$\nabla_\theta = \begin{bmatrix} \dfrac{\partial}{\partial\theta_1} \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial}{\partial\theta_p} \end{bmatrix} \tag{2}$$

定义对数似然函数为：

$$l(\theta) = \ln p(D \mid \theta) \tag{3}$$

我们可以把求使对数似然函数最大的θ的过程写成规范的形式如下：

$$\hat{\theta} = \arg\max_\theta l(\theta) \tag{4}$$

其中的对数似然函数显然是依赖于样本集的，结合公式（1）和（4）得下列结论：

$$l(\theta) = \sum_{k=1}^{n} \ln p(x_k \mid \theta) \tag{5}$$

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k \mid \theta) \tag{6}$$

这样，我们就得到了一组求解最大似然估计值的必要条件，这组条件是由 p 个方程所组成的方程组：

$$\nabla_\theta l = 0 \tag{7}$$

这里所求的解θ可能是真正的全局最大值点，也可能是局部极值点，或者仅仅是$l(\theta)$的一个拐点。我们还必须检查所得到的解是否是位于定义域空间的边界上。如果所有的极值解都已经求得了，我们就能确定其中必有一个是全局的最大值点。接着，我们必须对所有的可能解进行检查以确定其中的真正的全局最优点。如果训练样本个数越多，其中的样本越具有代表性，那估计值也就越接近真实值。

2.2 高斯情况：μ未知

为了加深对最大似然估计方法的理解，我们这里将深人讨论当训练样本服从多元正态分布时的情况。设这个多元正态分布的均值为μ，而协方差矩阵为Σ。首先，为了简单起见，我们将先分析当协方差矩阵Σ已知，而均值μ未知的情况。

在这样的假设下，我们考虑一个训练样本点x_k，有下面的式子成立：

$$\ln p(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu) \tag{8}$$

$$\nabla_\mu \ln p(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu) \tag{9}$$

这里我们用μ标识θ是为了强调参数向量θ中的未知量为μ。结合式(6),(7),(9)我们可以得到，对μ的最大似然估计值必须满足下式：

$$\sum_{k=1}^{n} \Sigma^{-1}(x_k - \mu) = 0 \tag{10}$$

两边乘以协方差矩阵Σ,并且进行一些简单整理后,我们得到下述公式：

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{11}$$

这个公式说明：对均值的最大似然估计就是对全体样本取平均，也就是均值的最大似然估计等于样本均值。因此有时也把这个结果记为$\hat{\mu}_n$，以强调依赖于训练样本的个数n这一事实。其几何意义是,如果把样本集看作是一个由点组成的云团，则这个样本均值就是这个云团的质心。样本均值还具有其他的一些优秀的统计性质。通常在实际应用中,即使不知道这是最大似然估计方法得出的结果,我们往往也直接使用样本均值作为实际均值的估计。

2.3 高斯情况：μ和Σ均未知

实际应用中,多元正态分布的更典型情况是,均值μ和协方差矩阵Σ均未知。这样，参数向量θ就是由这两个成分组成。我们首先考虑单变量的情况，其中参数向量θ的组成成分是：$\theta_1 = \mu$，$\theta_1 = \sigma^2$。这样对于单个训练样本的似然函数为：

$$\ln p(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \tag{12}$$

对上式关于变量θ求导得：

$$\nabla_{\theta l} = \nabla_\theta \ln p(x_k \mid \theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix} \tag{13}$$

利用式(7)，我们得到对于全体样本的对数似然函数的极值条件如下：

$$\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}\left(x_k-\hat{\theta}_2\right)=0 \tag{14}$$

$$-\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}+\sum_{k=1}^{n}\frac{\left(x_k-\hat{\theta}_1\right)^2}{\hat{\theta}_2^2}=0 \tag{15}$$

其中的$\hat{\theta}_1$，$\hat{\theta}_2$别是对于θ_1，θ_2的最大似然估计。

把$\hat{\theta}_1$，$\hat{\theta}_2$用$\hat{\mu}$，$\hat{\sigma}^2$代替，并进行简单的整理，我们得到下面的对于均值和方差的最大似然估计结果为：

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k \tag{16}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\hat{\mu}\right)^t \tag{17}$$

当高斯函数为多元时，最大似然估计的过程也是非常类似的，当然，也将更加复杂。对于多元高斯分布的均值μ和协方差矩阵Σ的最大似然估计结果为：

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k \tag{18}$$

$$\hat{\Sigma}=\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\hat{\mu}\right)\left(x_k-\hat{\mu}\right)^t \tag{19}$$

由此可以看出协方差的最大似然估计则是n个矩阵的算术平均。因为实际的协方差矩阵是关于矩阵的数学期望，所以可以看到协方差的最大似然估计结果也是非常直观和令人满意的。

2.4 估计的偏差

在上面的分析中，对方差的最大似然估计是有偏的估计。也就是说，对所有可能的大小为n的样本集进行方差估计，其数学期望并不等于实际的方差，原因如下：

$$\varepsilon\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]=\frac{n-1}{n}\sigma^2\neq\sigma^2 \tag{20}$$

假设一个分布的方差σ^2非零，如果考虑仅有一个样本的极端情况。在这种情况下，估计值的数学期望为0，所以不等于σ^2。类似的，对协方差矩阵的最大

似然估计也是有偏的。对协方差矩阵的无偏估计则如下式所示：

$$C=\frac{1}{n\text{-}1}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})^t \tag{21}$$

如果一个估计器对于所有的分布都是无偏的，那么它就被称为绝对无偏的。如果某一个估计器在样本数n很大时，能够趋于无偏估计，则这个估计器被称为渐进无偏的。在许多模式识别的实际问题中，如果训练样本集足够大，那么渐进无偏估计算子得出的结果是可以被接受的。

显然，$\hat{\sum}$是渐进无偏的估计，但当样本数n很大时，这两个结果几乎是相同的。但是，同时存在这两个相似却又不完全相同的估计方法，这总是令人迷惑的。我们总是希望某一估计能够使得最后的分类结果为最优，而这一要求却是比较抽象的。无疑，使用最大似然估计的结果是合理的，在实际中也是相当有效的。

如果我们对于产生已知样本分布的数学模型及其参数向量θ的建模都是可靠的，那么最大似然估计就能够有很好的结果。但如果我们的数学模型本身就有错误，我们是不能够保证基于那个不正确的模型的估计方法仍然能得到最优分类器。所以正确的模型带来的误差的影响是非常巨大的，也就是说，需要对数学模型有较可靠的知识。如果初始假设的数学模型与实际的情况有较大偏差的话，那显然无法保证设计出来的分类器会是最优分类器。

3 贝叶斯估计

虽然使用贝叶斯估计方法得到的结果与最大似然估计的结果很相似，但这两个方法在本质上是很不同的：在最大似然估计方法中，我们把需要估计的参数向量θ看作是一个确定而未知的参数。而在贝叶斯学习方法中，我们把参数向量θ本身看成一个随机变量，已有的训练样本使我们能够把对于θ的初始密度的估计转化为后验概率密度。

3.1 类条件密度

贝叶斯分类方法的核心是后验概率$P(w_i|x)$的计算。贝叶斯公式告诉我们如何根据类条件密度$P(x|w_i)$和各类别的先验概率$P(w_i)$来计算这个后验概率。但是，在这两个概率也未知的情况下，我们能做的就是希望利用现有的全部信息来

计算后验概率$P(w_i|x)$。其中的“现有的全部信息”如下：一部分为我们的先验知识，比如未知概率密度函数的形式，未知参数的取值范围等；另一部分信息则来自于训练样本本身。用D表示现有训练样本的集合，那么我们把后验概率$P(w_i|x)$进一步写成$P(w_i|x，D)$的形式，用来强调训练样本在估计过程中的重要性。根据这些概率，我们就能够设计出贝叶斯分类器，那么贝叶斯公式变为：

$$P(w_i|x，D)=\frac{P(x|w_i，D)P(w_i|D)}{\sum_{j=1}^{c}P(x|w_j，D)P(w_j|D)} \tag{22}$$

这一公式指出，我们能够根据训练样本提供的信息来确定类条件概率密度$P(x|w_j，D)$和先验概率$P(w_j，D)$。

尽管公式（22）具有更大的一般性，但实际上我们通常可以认为先验概率可以事先得到，或者仅通过简单的计算就能够求得先验概率。因此，我们通常把$P(w_i，D)$简写成$P(w_i)$。 而且，由于我们处理的是有监督的学习，因此完全可以把每一个样本都归到它所属的类中去，即把全体训练样本依据类别分到c个次样本集：$D_1,D_2,...,D_c$中去。如同在讨论最大似然问题时一样，如果$i\neq j$，那么样本集D中的训练样本就对$P(x|w_j，D)$没有任何影响。这样得到化简后的公式如下：

$$P(w_i|x，D)=\frac{P(x|w_i，D)P(w_i)}{\sum_{j=1}^{c}P(x|w_j，D)P(w_j)} \tag{23}$$

其次，由于能够对每一个类别进行分别处理，因此公式中为了说明各个类别的记号都可以省略，简化了公式的形式。所以，就其实质来说,我们要处理的是c个独立的问题，每一个问题都是如下的形式：已知一组训练样本D，这些样本都是从固定但未知的概率密度函数中独立抽取的，要求根据这些样本估计$P(x，D)$，这就是贝叶斯学习的核心问题。

3.2 参数的分布

虽然具体的概率密度函数$P(x)$未知，但我们假设其参数形式是已知的。所以

惟一未知的就是参数向量θ的值。为了明确的表示$P(x)$的形式已知而参数的值未知这一事实，我们强调条件概率密度函数$P(x|\theta)$是完全确定性的。在观察到具体的训练样本之前,我们已有的关于参数向量θ的全部知识就可以用已知的先验概率密度函数$P(\theta)$来体现。对训练样本的观察，使得我们能够把这个先验概率密度转化成后验概率密度函数$P(\theta|D)$，并且,我们希望这个后验概率密度$P(\theta|D)$在θ的真实值附近有非常显著的尖峰。

因此，基本目标是计算后验概率密度函数$P(x|D)$，并且使得它尽可能精确地逼近$P(x)$。我们把联合概率密度$P(x,\theta|D)$对θ进行积分，也就是

$$P(x|D)=\int p(x,\theta|D)d\theta \qquad (24)$$

其中积分是对整个定义域进行的。现在我们能够把$P(x,\theta|D)$写成乘积$P(x,\theta|D)\,P(\theta|D)$的形式。由于对测试样本$x$和训练样本集$D$的选取是独立进行的,因此$P(x,\theta|D)$就等于$P(\theta|D)$。也就是说，只要我们能够得到参数向量$\theta$的值，$x$的分布形式就完全已知了。这样，上式可以写为：

$$P(x|D)=\int p(x|\theta)p(\theta|D)d\theta \qquad (25)$$

上式就是贝叶斯估计中最核心的公式，它把类条件概率密度$P(x|D)$和未知参量的后验概率密度$P(\theta|D)$联系起来。如果后验密度$P(\theta|D)$在某一个值θ附近形成最显著的尖峰，那么就有$P(x|D)\approx P(x|\theta)$，也就是说，用估计值$\theta$近似代替真实值所得的结果。当然，这个结果的前提条件是要求$P(x|\theta)$必须光滑，并且积分拖尾的影响足够小。这些条件通常很典型，但也并非一成不变，有时会有例外的情况。总的来说，如果我们对参数向量θ的真实值并不十分有把握的话，那么该方程指导我们应该把$P(x|\theta)$对所有可能的θ求平均,这样得到的结果将最令人满意。

4 贝叶斯参数估计:高斯情况

在这一节中，我们对高斯正态密度函数的情况，用贝叶斯估计方法来计算θ的后验概率密度函数$P(\theta|D)$和设计分类器所需的概率密度函数$P(x|D)$。其中我们假设$P(x|\mu)\sim N(\mu|\Sigma)$。

4.1 单变量情况：$P(\mu|D)$

我们先考虑只有均值μ未知的情况，这里先处理一维的情况，也就是

$$P(x|\mu)\sim N(\mu|\sigma^2) \tag{26}$$

其中惟一的未知数就是均值μ。而且，我们认为所有的关于均值μ的先验知识都包含在先验概率密度函数$P(\mu)$中，我们假设均值μ服从

$$P(\mu)\sim N(\mu_0|\sigma_0^2) \tag{27}$$

在选择好了均值μ的先验概率密度函数以后，设想从均值μ的分布$P(\mu)$中选取一个具体的μ值，一旦这个μ值被选定，它就成为μ的真实值，由于我们已经认为$P(x|\theta)$是完全已知的，也就是完全确定了变量x的概率密度函数。然后，再从变量x的概率密度函数中，独立的抽取n个样本：$x_1,x_2,...,x_n$，记$D=\{x_1,x_2,...,x_n\}$。应用贝叶斯公式，得到

$$P(\mu|D)=\frac{P(D|\mu)P(\mu)}{\int P(D|\mu)P(\mu)d\mu}=\alpha\prod_{k=1}^{n}P(x_k|\mu)P(\mu) \tag{28}$$

其中α是依赖于样本集D的归一化系数，这个系数不依赖于μ。这一公式说明了训练样本能如何的影响对μ值的估计。它把先验概率密度$P(\mu)$和后验概率密度$P(\mu|D)$联系了起来。因为$P(x_k|\mu)\sim N(\mu|\sigma^2)$，和$P(\mu)\sim N(\mu_0|\sigma_0)$，得如下式子：

$$P(\mu|D)=\alpha\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]$$
$$=\alpha^{'}\exp\left[-\frac{1}{2}\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2+\frac{\mu-\mu_0}{\sigma_0}\right] \quad (29)$$
$$=\alpha^{''}\exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0{}^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k+\frac{\mu_0}{\sigma_0{}^2}\right)\mu\right]\right]$$

上式中的不依赖于μ的那些因子都被归入系数α，$\alpha^{'}$，$\alpha^{''}$中了。这样，我们发现$P(\mu|D)$是一个指数函数,其中的指数部分为μ的二次型。也就是说，$P(\mu|D)$实质上还是一个正态分布。因为这一事实对任意大小的样本集均成立，因此$P(\mu|D)$在样本个数n增加时仍保持正态分布。我们把$P(\mu|D)$称为复制密度函数，把$P(\mu)$称为共轭先验。如果写成下面的形式：$P(\mu|D)\sim N(\mu_n|\sigma_{\mathrm{n}}^2)$，也就是如下的式子：

$$P(\mu|D)=\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] \quad (30)$$

那么对公式（29）和公式（30）应用对应项相等的原则，就可以求得μ_n和σ_{n}^2：

$$\frac{1}{\sigma_n^2}=\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2} \quad (31)$$

$$\frac{\mu_n}{\sigma_n^2}=\frac{n}{\sigma^2}\hat{\mu}_n+\frac{\mu_0}{\sigma_0^2} \quad (32)$$

其中,n是样本均值

$$\hat{\mu}_n=\frac{1}{n}\sum_{k=1}^{n}x_k \quad (33)$$

进一步求解μ_n和σ_{n}^2得：

$$\mu_n=\left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n+\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \quad (34)$$

$$\sigma_{\mathrm{n}}^2=\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \quad (35)$$

上述方程显示了先验知识和样本观测结果的结合，并且形成后验密度$P(\mu|D)$的。总的说来，μ_n代表了在观察到n个样本后，我们对μ的真实值的最好的估计，而$\sigma_n=0$。反映了对这个估计的不确定程度。根据公式(35)可以了解到σ_n^2是n的单调递减函数，并且在n趋于无穷大时，σ_n^2趋于$\frac{\sigma^2}{n}$,也就是说，每增加一个观察样本,我们对μ的估计的不确定程度就能减少。当n增加时，$P(\mu|D)$的波形变得越来越尖，并且在n趋于无穷大时,逼近于狄拉克函数。这一现象通常就被称为贝叶斯学习过程，该过程如图 2。

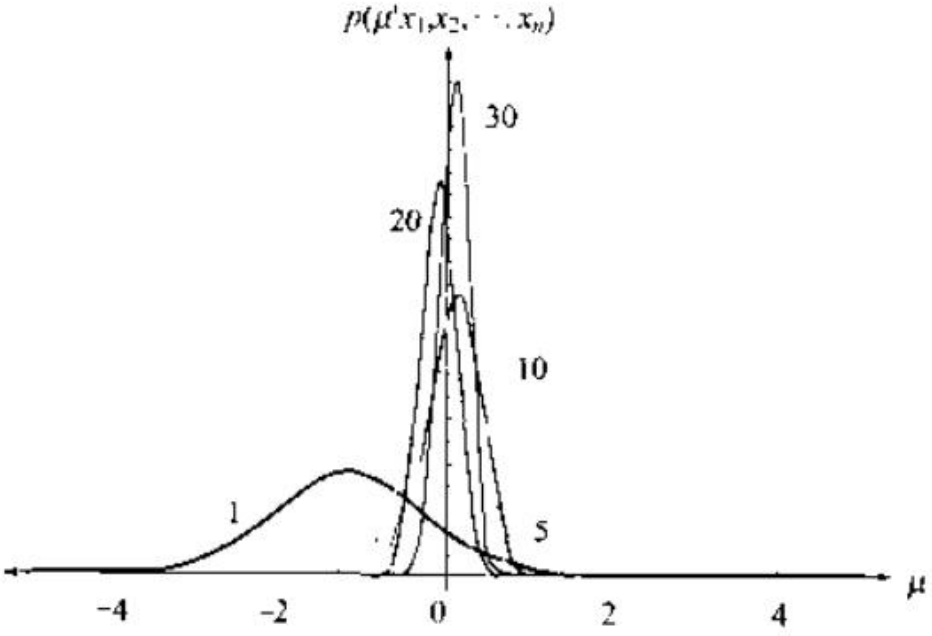

图 2 贝叶斯学习过程

根据公式（34），我们知道，在通常情况下，μ_n都是$\hat{\mu}_n$和μ_0的线性组合，两者的系数均为非负，并且和为 1，也就是说μ_n位于$\hat{\mu}_n$和μ_0的连线上。如果$\sigma_n\neq 0$，则当n趋于无穷大，μ_n趋近于样本均值。如果$\sigma_n=0$，这是一种退化的情况，也就是说，我们对先验估计μ_0是如此确信，以至于任何观察样本都无法改变我们的态度。在另一种极端情况中如果$\sigma_0>>\sigma$，也就是说我们对先验估计μ_0是如此的不确信，以至于我们直接把样本均值$\hat{\mu}_n$当作了μ。总的来说，先验知识和经验数据各自的贡献之间的平衡取决于σ^2和σ_0^2的比值，这个比值被称为

决断因子。如果该值不是无穷大，那么当获得了足够的样本后，μ_0 和 σ_0^2 的具体数值的精确假定就变得无关紧要了，同时 μ_n 将收敛于样本均值 $\hat{\mu}_n$。

4.2 单变量情况：$P(x|D)$

在得到了均值的后验密度 $P(\mu|D)$ 之后，就可以计算类条件概率密度 $P(x|D)$ 了。根据式（25），（26），（30），我们得到如下的式子：

$$\begin{aligned} P(\mu|D) &= \int p(x|\mu)p(\mu|D)d\mu \\ &= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu \\ &= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n) \end{aligned} \tag{36}$$

其中

$$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2 x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu$$

也就是说，作为 x 的函数，类条件概率密度函数 $P(x|D)$ 正比于

$$\exp\left[-(1/2)(x-\mu_n)^2/(\sigma^2+\sigma_n^2)\right]$$

因此 $P(x|D)$ 是一个正态分布，均值为 μ_n，方差为 $\sigma^2+\sigma_n^2$，即

$$P(x|D) \sim N(\mu_n,\ \sigma^2+\sigma_n^2) \tag{37}$$

也就是说，为了得到类条件概率密度函数 $P(x|D)$，其参数形式为已知的 $P(x|\mu)\sim N(\mu|\sigma^2)$，我们只需用 μ_n 替换 μ，用 $\sigma^2+\sigma_n^2$ 替换 σ^2 就可以了。在效果上，μ_n 被当作 μ 的真实值看待，而这时的方差比起来说 σ^2 相对增加了，原因是均值 μ 的不确定性增加了对 x 的不确定性。这就是最终的结果：$P(x|D)$ 就是类条件概率密度函数 $P(x|\mathrm{w}_j,D_j)$，结合先验概率 $P(\mathrm{w}_j)$，我们就完全掌握了设计贝叶斯分类器所需的概率知识。在这点上，贝叶斯估计方法与最大似然方法不同，因为最大似然方法只是估计 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 的值，而不是估计 $P(x|D)$ 的分布。

4.3 多变量情况

对于多变量的情况，在协方差矩阵Σ已知，而均值μ未知的情况下，并不能把单变量的结果作简单的推广。我们在这里将大略的描述分析的过程，如同一维的情况，我们假设：

$$P(x|\mu)\sim N(\mu|\Sigma)\text{且}P(\mu)\sim N(\mu_0|\Sigma_0) \tag{38}$$

其中的Σ，Σ_0，μ_0均假设为已知。在观测到样本集D中的n个互相独立的样本$x_1,x_2,...,x_n$后，我们使用贝叶斯公式得到：

$$\begin{aligned}P(\mu|D)&=\alpha\prod_{k=1}^{n}p(x_{\mathrm{k}}|\mu)p(\mu)\\&=\alpha^{'}\exp\left[-\frac{1}{2}\left(\mu^{'}\left(n\Sigma^{-1}+\Sigma_0^{-1}\right)\mu-2\mu^{'}\left(\Sigma^{-1}\sum_{k=1}^{n}x_k+\Sigma_0^{-1}\mu_0\right)\right)\right]\end{aligned} \tag{39}$$

进行配方和变量代换，上式可以简化表示为

$$P(\mu|D)=\alpha^{''}\exp\left[-\frac{1}{2}(\mu-\mu_n)^t\Sigma_n^{-1}(\mu-\mu_n)\right] \tag{40}$$

这样，$P(\mu|D)\sim N(\mu_n|\Sigma_{\mathrm{n}})$，并且再一次的，我们又得到了复制概率密度。对式（39）和式（40）应用对应项相等的原则，得到分别类似于式（34），式（35）的等式如下：

$$\Sigma_n^{-1}=n\Sigma^{-1}+\Sigma_0^{-1} \tag{41}$$

$$\Sigma_n^{-1}\mu_n=n\Sigma^{-1}\hat{\mu}_n+\Sigma_0^{-1}\mu_0 \tag{42}$$

其中$\hat{\mu}_n$是样本均值

$$\hat{\mu}_n=\frac{1}{n}\sum_{k=1}^{n}x_k \tag{43}$$

在对上述的几个方程求解均值μ_n和协方差矩阵Σ_n时，需要用到的恒等式如下：

$$\left(A^{-1}+B^{-1}\right)^{-1}=A(A+B)^{-1}B=B(A+B)^{-1}A \tag{44}$$

其中矩阵为非奇异矩阵，经过推导进一步解得如下的式子：

$$\mu_n=\Sigma_0\left(\Sigma_0+\frac{1}{n}\Sigma\right)^{-1}\hat{\mu}_n+\frac{1}{n}\Sigma\left(\Sigma_0+\frac{1}{n}\Sigma\right)^{-1}\mu_0 \tag{45}$$

$$\Sigma_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\frac{1}{n}\Sigma \tag{46}$$

如果我们利用积分

$$p(x|D) = \int p(x|\mu)p(\mu|D)d\mu \tag{47}$$

那么可以进一步证明：$P(x|D) \sim N(\mu_n, \Sigma + \Sigma_n)$。然而，这一结果可以用另一种简单的方法来得出：因为x可以看成两个互相独立的随机变量的和，其中一个变量为服从$P(\mu|D) \sim N(\mu_n, \Sigma_n)$的变量$\mu$，另一个变量为独立随机变量$y$，服从分布$P(y) \sim N(0, \Sigma)$。因为两个独立的正态分布的向量随机变量的和仍然为一个正态分布的向量，其均值为各自均值的和,其协方差矩阵为各自协方差矩阵的和，这样我们就得到了如下的关系式：

$$P(x|D) \sim N(\mu_n, \Sigma + \Sigma_n) \tag{48}$$

到目前为止，我们就完成了在参数服从高斯分布的这种情况下，从单变量到多变量的推广。

5 贝叶斯参数估计：一般理论

我们已经看到了在多元高斯分布的情况下,如何应用贝叶斯估计方法去获得后验概率$P(x|D)$。在一般情况下，只要未知概率分布能够被表示成参数形式，则这一方法就能得到同样的使用。一些基本的假设如下：

（1）条件概率密度函数$P(x|\theta)$是完全已知的，参数向量θ的具体数值未知。

（2）参数向量θ的先验概率密度函数$p(\theta)$包含了对于θ的全部先验知识。

（3）其余的关于参数向量θ的信息就包含在观察到的独立样本$x_1, x_2, ..., x_n$中，这些样本都服从未知的概率密度函数$P(x)$。

最基本的问题就是计算后验概率密度函数$P(\theta|D)$，因为一旦求得后验概率密度函数$P(\theta|D)$，我们就可以利用式(25)来计算$P(x|D)$：

$$P(x|D) = \int P(x|\theta)P(\theta|D)d\theta \tag{49}$$

根据贝叶斯公式，我们有

$$P(\theta \mid D)=\frac{P(D \mid \theta)P(\theta)}{\int P(D \mid \theta)P(\theta)d\theta} \tag{50}$$

再根据样本之间的独立性假设得：

$$P(D \mid \theta)=\prod_{k=1}^{n} p(x_k \mid \theta) \tag{51}$$

这就完成了对问题的正式解答，同时，式（50）和式（51）阐明了与最大似然估计之间的关系。假设$P(D \mid \theta)$在$\theta=\hat{\theta}$处有一个非常尖的峰值，如果先验概率$P(\theta)$在$\theta=\hat{\theta}$处非零，并且在周围的某一邻域内变化不大，那么$P(\theta \mid D)$也在同一地方有一个峰值。这样，式（49）表明了$P(x \mid D)$将趋近于$P(x \mid \theta)$，而这一结果也正是根据最大似然方法得到的结论。如果$P(D \mid \theta)$的峰值非常尖，那么先验知识中对θ的真实值的不确定性几乎可以忽略。在这种情况下（也包括其他的更为一般的情况），是贝叶斯估计方法而不是最大似然估计方法，告诉我们如何根据所有的现有信息来计算条件概率密度函数$P(x \mid D)$。

到目前为止，我们已经得到解,但是还有许多有趣的问题值得进行研究。其中的一个问题就是执行这些计算的复杂度如何。另一个问题是$P(x \mid D)$能否可靠的收敛到真正的$P(x)$以及收敛速度问题。下面将简要的讨论收敛性问题，为了明确地表示集合中已有的样本个数n，我们采用这样的记号：$D=\{x_1,x_2,...,x_n\}$。然后，根据公式（51），如果$n>1$，那么我们有

$$P(D^{n} \mid \theta)=P(x_{n} \mid \theta)P(D^{n-1} \mid \theta) \tag{52}$$

将上式代人公式（50），并且结合贝叶斯公式则得到下面的结果：

$$P(\theta \mid D^{n})=\frac{P(x_{n} \mid \theta)P(\theta \mid D^{n-1})}{\int P(x_{n} \mid \theta)P(\theta \mid D^{n-1})d\theta} \tag{53}$$

注意，当尚未有观测样本时，令$P(\theta \mid D^{n})=P(\theta)$。反复运用上述公式，能够产生一系列的概率密度函数：$P(\theta)$，$P(\theta \mid x_1)$，$P(\theta \mid x_1, x_2)$等等，这一过程被称为参数估计的递归的贝叶斯方法。这是我们遇到的第一个增量学习或在线学习

算法，其特点是学习过程随着观察数据的不断获得而不断进行。如果这一概率密度函数的序列最终能够收敛到一个中心在参数的真实值附近的狄拉克函数，那么就实现了贝叶斯学习过程。当然，我们还将遇到许多非增量的学习方法，其中所有的训练样本必须在学习过程开始前就全部获得。

在原则上，为了计算$P(\theta|D^{n})$，等式（53）要求保留D^{n-1}中的所有训练点。然而，对于某些分布，可能几个与$P(\theta|D^{n-1})$相关的参数的估计就足以包含所需的全部信息了。这样的参数被称为对应与某个特定分布的充分统计量。有些书的作者认为递归学习这一概念特指只使用某种充分统计量，而不是训练样本本身的的情况，而在本书中，我们把这种特殊情况称为真正的递归贝叶斯学习。

对于通常能遇到的典型的条件概率密度函数$P(x|\theta)$，后验概率密度函数序列一般都能收敛到狄拉克函数。因此，这就意味着，只要训练样本的数量足够多，就能够确定惟一的一个最适合这些训练样本的θ的值。也就是说，参数θ能被条件概率密度函数$P(x|\theta)$惟一确定。在这种情况下，概率密度函数$P(x|\theta)$被称为可辨识的。这一性质的严格证明需要确切知道概率密度函数$P(x|\theta)$和$P(\theta)$的形式，但证明过程本身并不是十分困难的。

然而的确存在这样的情况，即不同的θ值，而产生的$P(x|\theta)$都相同。在这种情况下，θ不能由$P(x|\theta)$惟一确定，并且$P(x|D^{n})$将在所有可能的θ值的附近都形成尖峰。然而幸运的是，这种可能的不确定性并不会带来严重后果，因为在公式(25)中，参与运算的概率密度函数$P(x|\theta)$对所有可能的θ值都相同。也就是说，无论条件概率密度函数$P(x|\theta)$是可辨识与否，$P(x|D^{n})$总是会收敛到$P(x)$，然而，这种不确定性总是客观存在的。

5.1 最大似然方法和贝叶斯方法何时有区别

对于先验概率能保证问题有解的情况下，最大似然估计和贝叶斯估计在训练样本趋近于无穷时效果是一样的。然而在实际的模式识别问题中，训练样本总是有限的，因此，我们很自然地就会想到：在什么时候，最大似然估计和贝叶斯估计这两种方法将表现出不同，并且在这种情况下，我们应该选取哪一种方法。

决定我们的选择的标准有如下几个：其中的一个标准就是所使用的方法的计算复杂度。在这个标准下，最大似然估计法是较好的选择，因为运用最大似然法，将只涉及一些微分运算或梯度搜索技术以求得θ，而如果采用贝叶斯估计方法，则可能要求计算非常复杂的多重积分。这又引出了另一个标准：可理解性。在许多情况下，最大似然法要比贝叶斯估计方法更容易理解和掌握，因为它得到的结果是基于设计者所提供的训练样本的一个最佳解答，而贝叶斯估计方法得到的结果则是许多可行解答的加权平均值，反映出对各种可行解答的不确定程度，这就使得贝叶斯估计方法比最大似然估计方法更难于直观理解。也就是说，贝叶斯估计方法的结果反映出对所使用的模型的剩余的不确定性。

另一个选择的标准是我们对初始的先验知识的信任程度，比如对概率密度函数$P(x|\theta)$的形式。最大似然估计得到的结果$P(x|\theta)$的形式是与初始假设的形式一致的，而这一点对于贝叶斯估计就未必成立。就像在一些例题中一样，$P(x|D)$的初始假设为一个均匀分布，而贝叶斯估计得到的结果$P(x|\hat{\theta})$的形式却与初始假设的形式不同。总的说来，通过使用全部$P(\theta|D)$中的信息，贝叶斯估计方法比最大似然方法能够利用更多有用的信息。如果这些信息是可靠的话，那么有理由认为贝叶斯估计方法比最大似然估计方法能够得到更准确的结果。而且在没有特别的先验知识的情况下，贝叶斯估计方法与最大似然估计方法是很相似的。同时如果有非常多的训练样本，使得$P(x|\theta)$形成一个非常显著的尖峰，而先验概率$P(\theta)$又是均匀分布，那么前而所说的 MAP 估计在本质上也是与最大似然估计相同的。

然而，如果$P(\theta|D)$的波形比较宽，或者在$\hat{\theta}$附近是不对称的（这一不对称性并不是因为选取训练样本的过程而造成的，而是问题本身所决定的），那么，最大似然估计和贝叶斯估计产生的结果就不相同了。通常非常明显的不对称性显然表示了分布本身的某种特点。贝叶斯方法能够利用这些特点，而最大似然法却忽略了这些特点。而且贝叶斯估计方法对偏差和方差之间的折中研究的更加透彻，而这一折中是与训练样本的个数密切相关的。

当使用最大似然估计或贝叶斯估计的结果设计分类器时，采用的的方法为：

对每一类别都计算后验概率密度函数，并且根据最大后验概率对测试样本进行分类（如果还知道风险矩阵,那么我们也能够考虑进分类风险所带来的影响）。使得系统产生的最终分类误差的来源有如下几个。

（1）贝叶斯误差（或不可分性误差）：这一分类误差是由子不同的类条件概率密度函数$P(x|\omega_i)$之间的互相重叠引起的，这种分类误差是问题本身所固有的，因此永远无法消除。

（2）模型误差：由于选择了不正确的模型所导致的分类误差。只有当设计分类器时，计的模型形式中包括了正确的模型的时候，这一误差才可能消除。然而设计者总是根据对问题的先验知识和理解来选择模型，并不是在后续的估计过程中选择模型。因此，这一误差在最大似然和贝叶斯估计中的影响都是类似的。

（3）估计误差：这是由于采用了有限样本进行估计所带来的误差。这一误差的影响可以用增加训练样本个数的方法来减小。

这三种误差各自对整个问题的影响程度是因问题而异的。如果能够使用无限多的样本，那么估计误差就能够消除，因此这时全部的分类错误对于最大似然估计方法和贝叶斯估计方法来说都是一样的。

综上所述，在理论上，贝叶斯估计方法有很强的理论和算法基础。但在实际应用中，最大似然估计更加简便，而且，设计出的分类器的性能几平与贝叶斯方法得到的结果相差无几。

5.2 无信息先验和不变性

总的说来，关于$P(\theta)$的先验知识来自设计者对具体问题的理解和掌握，这其实是超出了分类器设计的范畴。然而，在某些情况下，还是有一些原则，这些原则能够使我们对先验概率分布$P(\theta)$的假设不会过于糟糕。这就引出了有关无信息先验（知识）的概念。

当我们在处理每一类别的先验概率时，如果没有其他的特别信息，那么我们都简单地假设每一类的概率相同。类似地，在贝叶斯估计方法中，我们对每一个参数也有一个无信息的先验估计。假设我们要使用贝叶斯方法从一组训练样本中估计其位置参数μ和尺度参数σ（例如，对于高斯分布，这两个参数就是均值和标准差，对于三角形分布，这两个参数就是中心位置和宽度等等）。对于这两个

参数，我们做的先验假设如下所述。

首先考虑位置参数μ。显然，我们要求这个先验分布不依赖于原点的具体位置，也就是说，我们对位置参数μ要求具有平移不变性。有这样的平移不变性的惟一分布就是在整个一维空间内的均匀分布。当然，这个分布其实是不合适的，因为这样就会有$\int p(\mu)d\mu=\infty$。

其次，考虑尺度参数的分布$P(\sigma)$的先验假设。显然，空域度量的单位米、英尺、英寸——应该与先验概率的形式无关，也就是说，我们要求尺度参数σ具有尺度不变性。考虑一个新的变量$\tilde{\sigma}=\ln\sigma$。如果σ被一个正的系数α所改变大小，即$\sigma\to\alpha\sigma$，这就使得新的变量产生了一个平移：

$$\tilde{\sigma}\to\ln\alpha+\ln\sigma=\ln\alpha+\ln\tilde{\sigma} \tag{54}$$

这样，如同对位置参数μ的要求一样，我们要求σ对所有可能的值都有均匀分布，也就是要求尺度参数σ必须具有如下分布：

$$p(\sigma)=\frac{1}{\sigma} \tag{55}$$

当然，这样的分布也是现实中无法实现的。

总的说来，如果已经知道必须满足的不变性，例如，平移不变性或对离散分布要求样本选取的顺序的无关性，那么就会对先验概率的可能具有的形式带进约束。如果我们能找到满足这种约束的分布，那么最后的结果就称为对这些不变性要求是无信息的。

我们容易认为使用无信息的先验分布形式能够达到客规性，即样本本身能发挥出最大的作用，但这种想法还是欠考虑的。比如，在估计一个高斯分布的标准差σ时，我们希望保证先验分布是无信息的，但是这样的保证这并不能使得σ^2也是无信息的。对于 MAP 估计器，要实现不变性将更为困难。因此，在贝叶斯估计方法中，不变性的考虑是非常有用的。

6 本章小结

如果我们已经知道某个类条件概率密度函数的参数形式，那么就可以把寻找这个分布本身的问题简化为学习（估计）分布的参数的问题（对每一个类w_i用

参数向量θ表示），估计结果就可以直接用于设计分类器。

最大似然估计方法寻找的是能最好的解释训练样本的那个参数值——也就是说，使得观测到现有的训练样本的概率最大化（在实际应用中,为了简化计算起见,通常使用的是对数似然函数）。而在贝叶斯参数估计中，这些参数被认为是某种具有先验概率密度的随机变量,而训练样本的作用就是把先验密度转化为后验密度。递归贝叶斯法是通过逐次修正的办法来更新贝叶斯参数估计的结果。

虽然贝叶斯估计方法在理论上更有说服力,但在实际应用中通常更多地使用最大似然估计,因为最大似然估计方法更容易实现,并且在大训练样本的条件下,得到的分类器的效果也较好。参数向量θ的充分统计量s是一个关于全部样本的函数，包含了训练样本中有助于确定θ的所有有用信息。一旦知道了已知形式的概率模型（比如，指数族函数）的充分统计量,我们就只需要从训练样本中估计这些充分统计量，就可以进行分类器设计了。

参考文献

该论文是我阅读《模式分类经典著作》第二版写的关于最大似然估计方法和贝叶斯估计方法的知识梳理与总结，所以主要的参考文献即为《模式分类经典著作》第二版。