第 3 章　概率密度函数的估计

3.1 引言

在第 2 章中，我们讨论了设计贝叶斯分类器的方法，即在先验概率 $P(\omega_i)$ 和类条件概率密度 $p(x|\omega_i)$ 已知的情况下，按一定的决策规则确定判别函数和决策面。但在实际工作中，类条件概率密度常常是未知的。以例 2.1 来说，我们不可能直接知道先验概率和类条件概率密度 $p(x|\omega_i)$。但是我们可能从经验中知道玉米和杂草的大致比例，因而可能推断出先验概率 $P(\omega_i)$。此外我们还可能得到一些玉米和杂草的样本。这就需要我们从这些样本中去估计出玉米和杂草的类概率密度 $p(x|\omega_1)$ 及 $p(x|\omega_2)$。这就是本章要讨论的有关概率密度函数的估计问题。

在实际中，我们能收集到一些样本，而未知的则可能是：

1．类条件概率密度，即各类的概率密度分布　$p(x|\omega_i)$；

2．先验概率 $P(\omega_i)$。

我们的最终任务是利用样本集设计分类器。一个很自然的想法是把分类器设计过程分为两步：第一步，利用样本集估计 $p(x|\omega_i)$ 和 $P(\omega_i)$，分别记为 $\hat{p}(x|\omega_i)$ 和 $\hat{P}(\omega_i)$。解决这样的问题可以利用统计推断中的估计理论。第二步，再将估计量 $\hat{p}(x|\omega_i)$ 和 $\hat{P}(\omega_i)$ 代入第 2 章的贝叶斯决策规则中，完成分类器设计。我们将这样的分类器设计过程称为基于样本的两步贝叶斯决策。

利用两步贝叶斯决策方法得到的分类器性能与第 2 章理论上的贝叶斯分类器有所不同。我们希望当样本数目 $N \to \infty$ 时，基于样本的分类器能收敛于理论上的结果。为此，只要说明 $N \to \infty$ 时，$\hat{p}(x|\omega_i)$ 和 $\hat{P}(\omega_i)$ 收敛于 $p(x|\omega_i)$ 和 $P(\omega_i)$ 就可以了。这在统计学中可通过对估计量性质的讨论来解决。

一旦得到了 $p(x|\omega_i)$ 和 $P(\omega_i)$，我们就可以利用第 2 章的方法实现一个分类器。因此，我们本章的主要任务是利用样本集估计 $p(x|\omega_i)$ 和 $P(\omega_i)$。一般来说，有两类方法估计概率密度函数。一类是参数方法。在参数方法中，假设函数形式是已知的，未知的是函数的参数。通过估计参数来完成概率密度函数的估计。这里我们只考虑两种常用的方法。一种是最大似然估计方法，另一种是贝叶斯估计方法。虽然这两种估计的结果通常是近似相等的，但从概念上和观点上来说它们是完全不同的。最大似然估计把参数看作是确定而未知的，最好的估计值是在获得实际观察样本的概率为最大的条件下得到的。这时的参数估计基本上依赖于使用的样本。而贝叶斯估计则把未知的参数当作具有某种分布的随机变量，考虑了未知参数的先验分布，从而得到对参数的更好的估计。

另一类方法是非参数估计。在非参数估计中假设概率密度函数的形式是未知的，要求我们直接推断概率密度函数本身。我们知道，在统计学中常见的一些典型分布形式不总是能够和实际中的数据分布吻合。这就迫使我们必须考虑非参数估计方法。本章仅讨论两种推断类条件概率密度的方法——Parzen 窗法及 k_N 近邻

法。而直接利用样本设计分类器的非参数方法则放到后面几章去讨论。

3.1.1 参数估计的基本概念

参数估计是统计推断的基本问题之一。下面介绍参数估计中的几个基本概念。

(1) 统计量：样本中包含着分布的信息，我们希望通过样本集把有关样本的分布信息抽取出来，就是说针对不同要求构造出样本的某种函数，这种函数在统计学中称为统计量。如：样本的均值，方差等量。

(2) 参数空间：在参数估计中，我们总是假设概率密度函数的形式是已知的。未知的是分布中的一些参数。通常把未知参数列为一个向量，记为θ。在统计学中，将未知参数向量θ的全部可能取值组成的集合称为参数空间，记为Θ。

(3) 点估计、估计量和估计值：点估计问题就是要构造一个统计量$d(x_1,\cdots,x_N)$作为参数θ的估计，这被称为θ的估计量。如果$x_1^{(i)},\cdots x_N^{(i)}$是属于类别$w_i$的$N$个样本观察值，代入统计量$d$就得到对于第$i$类的的具体数值，这个数值被称为$\theta$的估计值。

(4) 区间估计：除点估计外，还有另一类估计，它要求用区间(d_1,d_2)作为θ可能取值范围的一种估计。这个区间被称为置信区间，这类估计问题被称为区间估计。

估计总体分布的具体参数是一个点估计问题。我们下面介绍两种主要的点估计方法：最大似然估计和贝叶斯估计，它们都能得到相应的估计值。当然评价一个估计的“好坏”，不能按一次抽样结果得到的估计值与参数真值θ的偏差大小来确定，而必须从平均的和方差的角度出发进行分析。统计学中有一些关于估计量性质的分析。我们将利用统计学中的方法对参数估计的结果进行分析。

3.2 最大似然估计

3.2.1 最大似然估计方法

在最大似然估计方法中有以下假设：

(1) 待估计的是(非随机)未知的量。

(2) 假定有c个类，则总的样本由c个样本集$\chi_1,\chi_2,\cdots,\chi_c$构成，其中$\chi_j$中的样本都是从概率密度为$p(x|\omega_j)$的分布中独立抽取出来的。

(3) 类条件概率密度$p(x|\omega_j)$具有某种确定的函数形式。例如，正态分布、指数分布、γ分布、β分布等等，但其参数向量θ_j未知。例如一维正态分布$N(\mu_j,\sigma_j^2)$，未知的参数为$\theta_j=(\mu_j,\sigma_j^2)$。为了表示$p(x|\omega_j)$同$\theta_j$有关，就把$p(x|\omega_j)$记成$p(x|\omega_j,\theta_j)$。

(4)假定χ_i中的样本不包含关于$\theta_j(j\neq i)$的信息，也就是说不同类别的参数在函数上是独立的，这样就可以对每一类分别处理。也就是说χ_i中的样本只对θ_i提供有关信息，而没有关于$\theta_j(j\neq i)$的任何信息。

有了这些假设，就可以按照下面的方法分别处理c个类别的概率密度函数的估计问题。

已知一个包含有N个样本的样本集，即

$$\chi = \{x_1, x_2, \cdots, x_N\} \qquad (3\text{-}1)$$

我们假设这些样本是从一个概率分布函数 $p(x)$ 中独立抽取出来的。在有些情况下，我们可以知道或者可以假设这个概率分布函数 $p(x)$ 的函数形式，例如，正态分布、指数分布、γ 分布、β 分布等，但其参数向量 θ 未知。我们的任务就是根据样本 χ 估计出参数 θ。

需要说明的是，对于观测到的有限的数据集 χ，通常情况下总有很多的（有时是无穷多的）参数 θ 值对应的概率分布可以产生它。实际上，对于任意参数 θ 对应得分布 $p(x)$，只要该分布函数在每个数据点 $x_1, x_2, \cdots, x_N$ 处非零，它都可以成为候选分布。因此，要估计参数 θ 是一个病态问题。需要进一步说明的是，密度估计问题本质上是一个病态问题，因为对于观测到的有限的数据集 χ，一般来说有无穷多的概率分布可以产生它。

最大似然方法是估计参数 θ 的一种方法。因为产生数据集 χ 的参数 θ 可能很多，因此，最大似然方法从中挑选 θ 的原则是使得函数 $p(\chi \mid \theta)$ 最大。由于我们假设数据是独立同分布的，因此有

$$l(\theta) = p(\chi \mid \theta) = p(x_1, x_2, \cdots, x_N \mid \theta) = \prod_{k=1}^{N} p(x_k \mid \theta) \qquad (3\text{-}2)$$

式(3-2)是 θ 的函数。我们把 $l(\theta)$ 或者 $p(\chi \mid \theta)$ 叫做似然函数。

似然函数 $l(\theta)$ 给出了从分布函数中抽出 $x_1, x_2, \cdots, x_N$ 这样 N 个样本的概率。当参数 θ 取不同值时，$l(\theta)$ 也可能不同。当得到 N 个样本 $x_1, x_2, \cdots, x_N$ 时，我们想知道这组样本“最可能”来自哪个密度函数(θ 取什么值)。这个“最可能”的函数就是使得 $l(\theta)$ 值为最大的密度函数。因此，我们的任务就是通过最大化似然函数 $l(\theta)$ 来求取 θ。

例如，我们想知道一个班级微积分考试成绩的分布。我们经常假定这个分布是一个正态分布。如果随机问了一个学生的成绩为 80 分，那么我们认为这个班平均成绩是 80 的可能性要比其它分数，如 20，更大。实际上，根据后面的分析可以知道，参数平均成绩 80 是所有平均成绩中使得这个学生的成绩被抽中可能性最大的参数值。

我们用 $\hat{\theta}$ 表示使似然函数的值最大的 θ 值。它是样本 $x_1, x_2, \cdots, x_N$ 的函数，记为 $\hat{\theta} = d(x_1, x_2 . \cdots x_N)$，$\hat{\theta} = d(x_1, x_2 . \cdots x_N)$ 叫做 θ 的最大似然估计量。

可以采用一些优化方法求最大似然估计量。当参数 θ 只有一个，并且似然函数连续、可微时，最大似然估计量是下面微分方程的解，即

$$\frac{dl(\theta)}{d\theta}=0 \qquad (3\text{-}3)$$

在很多情况下，使用似然函数的对数往往比使用似然函数本身计算起来更容易些。因为对数函数是单调增函数，因此使对数似然函数最大的$\hat{\theta}$值也必然使似然函数最大。所以可以定义对数似然函数$l(\theta)$为

$$H(\theta)=\ln l(\theta) \qquad (3\text{-}4)$$

这时最大似然估计量就是

$$\frac{dH(\theta)}{d\theta}=0 \qquad (3\text{-}5)$$

的解。

图 3.1 给出了关于参数θ的最大似然估计的图示。

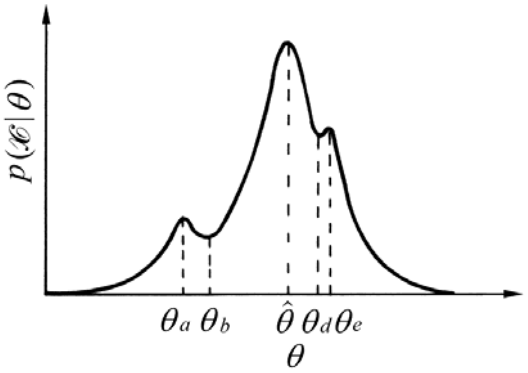


图 3.1 参数θ的最大似然估计的示意图

如果未知参数不是只有一个，而是有S个，则θ可表示为具有S个分量的向量

$$\theta=\left[\theta_1,\theta_2,\cdots,\theta_S\right]^T \qquad (3\text{-}6)$$

用∇_θ表示梯度算子

$$\nabla_\theta=\begin{bmatrix}\dfrac{\partial}{\partial\theta_1}\\ \vdots \\ \dfrac{\partial}{\partial\theta_S}\end{bmatrix} \qquad (3\text{-}7)$$

$H(\theta)$为对数似然函数

$$H(\theta)=\ln\left[l(\theta)\right]=\ln p(\chi|\theta)=\ln p(x_1,x_2,\cdots,x_N|\theta) \qquad (3\text{-}8)$$

当N个样本独立同分布时，式(3-8)可写为

$$H(\theta)=\ln\prod_{k=1}^{N}p(x_k\mid\theta)=\sum_{k=1}^{N}\ln p(x_k\mid\theta) \tag{3-9}$$

因而有

$$\nabla_\theta H(\theta)=\sum_{k=1}^{N}\nabla_\theta \ln p(x_k\mid\theta) \tag{3-10}$$

令

$$\nabla_\theta H(\theta)=0 \tag{3-11}$$

式(3-11)对应的S个方程是θ为最大似然估计量的必要条件。如果式(3-11)的解$\hat{\theta}$能使似然函数值最大，则$\hat{\theta}$就是θ的最大似然估计。有时式(3-11)可能有多个解。例如，在图 3.1 中，有 5 个解。虽然$\theta_a,\theta_b,\theta_c,\theta_d$都是解，但它们并不使似然函数最大，只有$\hat{\theta}$才使似然函数最大。所以，当计算出式(3-11)的解后，需要分析其解是全局最优，局部极值，或者不是一个局部极值点。

用式(3-11)求极大值是通常的做法，但有时不一定行得通。例如，随机变量x服从均匀分布，即

$$p(x\mid\theta)=\begin{cases}\dfrac{1}{\theta_2-\theta_1} & \theta_1<x<\theta_2\\ 0 & \text{其他}\end{cases} \tag{3-12}$$

但参数θ_1,θ_2未知。设从该分布中独立地抽取出N个样本$x_1,x_2,\cdots,x_N$。则其似然函数为

$$l(\theta)=p(\chi\mid\theta)=\begin{cases}p(x_1,x_2,\cdots,x_N\mid\theta_1,\theta_2)=\dfrac{1}{(\theta_2-\theta_1)^N}\\ 0\end{cases} \tag{3-13}$$

对数似然函数为

$$H(\theta)=-N\ln(\theta_2-\theta_1) \tag{3-14}$$

用式(3-11)求

$$\begin{aligned}\frac{\partial H}{\partial\theta_1}&=N\bullet\frac{1}{\theta_2-\theta_1}\\ \frac{\partial H}{\partial\theta_2}&=-N\bullet\frac{1}{\theta_2-\theta_1}\end{aligned} \tag{3-15}$$

从式(3-15)方程组中解出的参数θ_1和θ_2至少有一个为无穷大，这是无意义的结果。所以我们必须用其他方法来找最大值。从式(3-12)看出，当$\theta_2-\theta_1$越小，则似然函数越大。而在给定一个有N个观察值$x_1,x_2,\cdots,x_N$的样本集中，如果我们用x'表示观察值中最小的一个，用x''表示观察值中最大的一个，显然θ_1不能大于x'，θ_2不能小于x''，因此$\theta_2-\theta_1$的最小可能值是$x''-x'$，这时θ的最大似然估计显然是

$$\hat{\theta}_1 = x', \hat{\theta}_2 = x''$$

在实际求解最大似然估计量时，如果似然函数比较复杂，我们还需要采用其他一些更为复杂的求解方法。在本章后面描述的 EM 算法就是这样的一个例子。

3.2.2 正态分布时参数的最大似然估计

本小节以正态分布为例说明上述参数估计方法的使用。

在多维正态分布情况下，当均值向量 μ 和协方差矩阵 Σ 未知时，它们就构成所要估计的参数向量 θ 的分量。为简单起见，我们仅考虑单变量正态分布时利用最大似然估计方法来估计其均值 μ 和方差 σ^2。此时，$\theta_1 = \mu$，$\theta_2 = \sigma^2$，$\theta = [\theta_1, \theta_2]^T$，概率密度分布形式为

$$p(x|\theta) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (3\text{-}16)$$

其中 μ，σ^2 为未知参数。我们的任务是从样本集 $\chi = \{x_1, x_2, \cdots, x_N\}$ 中求出 μ 和 σ^2 的最大似然估计值 $\hat{\mu}$ 和 $\hat{\sigma}^2$。

由式(3-10)和式(3-1)知，最大似然估计量 $\hat{\theta}$ 为方程

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \nabla_\theta \ln p(x_k|\theta) = 0 \qquad (3\text{-}17)$$

的解。对于一个一维的正态分布，有

$$\ln p(x_k|\theta) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \qquad (3\text{-}18)$$

因此

$$\nabla_\theta \ln p(x_k|\theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix} \qquad (3\text{-}19)$$

将式(3-19)代入式(3-17)，得出最大似然估计 $\hat{\theta}_1$、$\hat{\theta}_2$ 满足下列条件

$$\begin{cases} \sum_{k=1}^{N} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \\ -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \end{cases} \qquad (3\text{-}20)$$

以 $\hat{\mu}=\hat{\theta}_1$， $\hat{\sigma^2}=\hat{\theta}_2$ 代入式(3-20)，解上述方程组，得

$$\hat{\mu}=\frac{1}{N}\sum_{k=1}^{N}x_k \qquad (3\text{-}21)$$

$$\hat{\sigma^2}=\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu})^2 \qquad (3\text{-}22)$$

对于多元正态分布，分析方法是类似的，只是运算复杂些。其最大似然估计的结果在形式上也与单变量情况类似，为

$$\hat{\mu}=\frac{1}{N}\sum_{k=1}^{N}x_k \qquad (3\text{-}23)$$

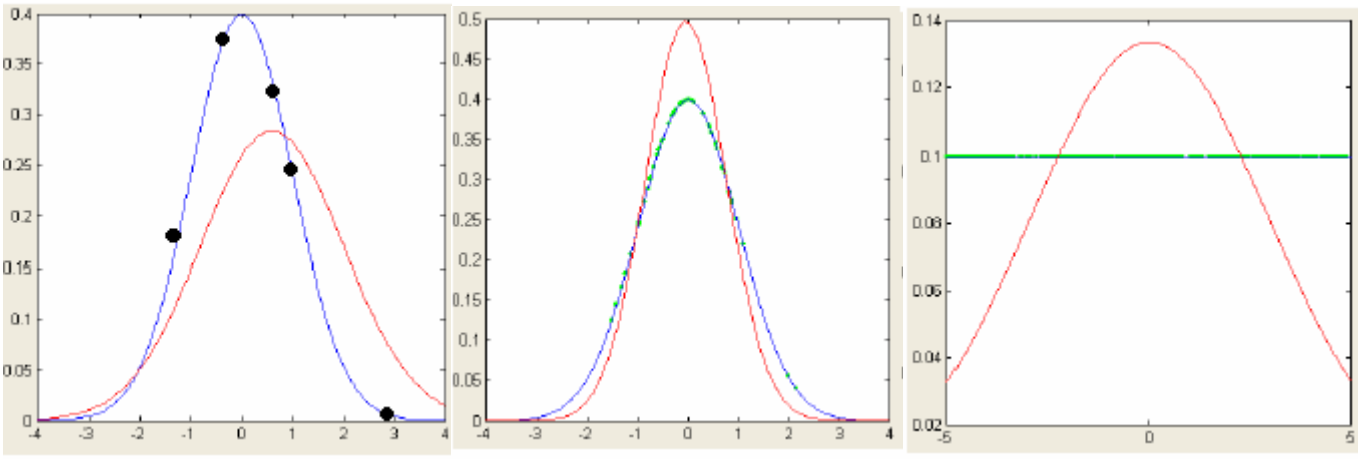
$$\hat{\sigma^2}=\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu})(x_k-\hat{\mu})^T \qquad (3\text{-}24)$$

其中 x_k 为多元正态分布中第 k 个样本，是 d 维向量，$\hat{\mu}$ 是均值向量 μ 的最大似然估计，$\hat{\Sigma}$ 是协方差矩阵 Σ 的最大似然估计。从以上结果可以得出结论：均值向量 μ 的最大似然估计是样本均值。协方差矩阵 Σ 的最大似然估计是 N 个矩阵 $(x_k-\hat{\mu})(x_k-\hat{\mu})^T$ 的算术平均。

可以知道，$E\hat{\mu}=\mu$，所以 $\hat{\mu}$ 是无偏的。但是 $\hat{\Sigma}$ 不是无偏的，而是渐近无偏的，即 $\lim_{N\to\infty}E\hat{\Sigma}=\Sigma$。$\hat{\Sigma}$ 的无偏估计为 $\frac{1}{N-1}\sum_{i=1}^{N}(x_k-\hat{\mu})(x_k-\hat{\mu})^T$，这些证明并不困难，我们把它作为习题留给读者。因此，由于 $\hat{\Sigma}$ 是 Σ 的渐近无偏估计，当样本数很大的时候，$\hat{\Sigma}$ 与 Σ 还是很接近的。但是，当样本数比较小的时候，即使估计 $\hat{\mu}$ 是无偏的，其估计值与真实值仍然可能有较大的差异。例如：我们在一个均值为 80 的正态分布上采样一次，虽然得到 $\hat{\mu}$=80 的概率最大，但是我们也常常会得到其他的数值，如 70。因此，采用最大似然估计，就会得到均值为 70 这样一个估计结果。这与真实值之间有较大差异。统计学在这方面有很多讨论，有兴趣的读者可以阅读统计学方面的教科书和相关研究文献。

当我们有了一定量的样本之后，最大似然估计通常可以给出不错的估计结果，这里还有一个前提条件就是我们给出的模型是正确的。当我们给出的模型有错误时，即使使用了大量的样本，最后的估计结果也往往是非常错误的。图 3.2 给出了这一讨论的一个示例。图 3.2（a）是利用从一个 $N(0,1)$ 正态分布中采样得到的 5 个样本来估计这个正态分布的结果。蓝色曲线是 $N(0,1)$ 正态分布，圆点表示的是采样得到的 5 个样本，红色表示的是估计出的正态分布 $N(0.59,1.97)$。图 3.2（b）是利用从这个 $N(0,1)$ 正态分布中采样得到的 50 个样本来估计这个正态分布的结果。估计结果为 $N(0.06,1.10)$。可以知道，这时的估计结果已经不错了。如果我们采用更多的样本，如 500，结果会与真实结果非常接近。图 3.2（c）是利用了从[-5，+5]这个区间的均匀

分布中采样得到的 500 个样本。这里我们用这些样本来估计一个正态分布的结果。很显虽然，这个正态分布和一个均匀分布会有很大的差异。



（a）使用 5 个点估计的结果　　（b）使用 50 个点估计的结果　　（c）使用从均匀分布中采样得到的 500 个点估计的结果

图 3.2 从不同分布采样之后估计正态分布的结果

在最大似然估计时，选择正确的模型是一个模型选择问题。关于这些问题，我们还会在后面的章节讨论。

3.3 贝叶斯估计与贝叶斯学习

3.3.1 贝叶斯估计

与最大似然估计不同的是，贝叶斯估计把要估计的参数 θ 看作是随机变量。因此利用贝叶斯定理有：

$$p(\theta \mid X) = \frac{p(X \mid \theta)\, p(\theta)}{\int_{\Theta} p(X \mid \theta)\, p(\theta) d\theta} \qquad (3\text{-}25)$$

式（3-25）中分母是一个和 θ 无关的量，这里可以不必考虑。 $p(X \mid \theta)$ 是在参数 θ 情况下样本集 X 出现的概率，也就是似然。 $p(\theta)$ 是随机变量 θ 的先验。 $p(\theta \mid X)$ 被称作是后验概率。

从式（3-25）很容易看出贝叶斯估计和最大似然估计之间的差异：贝叶斯估计把参数看作随机变量，因此，把这个随机变量的先验和似然函数综合考虑，进而对参数做估计。

因此，采用贝叶斯估计，首先需要确定参数的先验分布，然后计算参数的似然，最后计算参数的后验概率。

3.3.2 贝叶斯学习

本章的目的是通过样本集推断总体分布 $p(x \mid X)$。因为已假定概率分布的模型是已知的，所以问题就转化为模型参数 θ 的估计问题。上面我们介绍了两种参数估计方法。此外，我们还可以直接推断总体分布 $p(x \mid X)$。其思路中的前几步都与求贝叶斯估计方法相同，当利用贝叶斯公式求出 θ 的后验密度 $p(\theta \mid X)$ 后，

就直接通过联合概率计算概率分布函数

$$p(x|X)=\int p(x,\theta|X)d\theta=\int p(x|\theta)p(\theta|X)d\theta \qquad (3\text{-}26)$$

上式中用到这样一个结论： $p(x|\theta,X)=p(x|\theta)$ 。这是因为在参数 θ 控制的模型中样本之间是独立抽取的。式(3-26)中参数 θ 的后验密度 $p(\theta|X)$ 可以根据（3-25）计算。

在最大似然方法的讨论中，我们已经知道，似然函数 $p(X|\theta)$ 在 $\theta=\hat{\theta}$ 处很可能有一个尖锐的峰。现在参数 θ 是随机变量，如果其先验概率密度 $p(\theta)$ 在 $\theta=\hat{\theta}$ 处不为零，而且在附近又变化不大，则根据式(3-25)，$p(\theta|X)$ 在 $\theta=\hat{\theta}$ 处也将有尖锐的峰，这样式(3-26)可以近似为

$$p(x|X)\approx p(x|\hat{\theta}) \qquad (3\text{-}27)$$

$\hat{\theta}$ 是 θ 的最大似然估计值，也就是说贝叶斯解的结果 $p(x|X)$ 与最大似然估计的结果近似相等。

如果 $p(\theta|X)$ 的峰不尖锐，贝叶斯解将不能用最大似然估计来代替，这时需要通过式(3-26)和式(3-25)来计算 $p(x|X)$ 。

另外一个要讨论的问题就是： $p(x|X)$ 是否收敛于 $p(x)$，其中 $p(x)$ 是 x 的真实分布，它的参数为真实参数 θ 。为了明确表示 X 中样本的个数，我们记 X^N 为由 N 个样本组成的样本集，即

$$X^N=\{x_1,x_2,\cdots,x_N\}$$

根据式(3-2)，当 $N>1$ 时，有

$$p(X^N|\theta)=p(x_N|\theta)p(X^{N-1}|\theta)$$

把它代入式(3-25)，并利用贝叶斯公式，可得

$$p(\theta|X^N)=\frac{p(x_N|\theta)p(\theta|X^{N-1})}{\int p(x_N|\theta)p(\theta|X^{N-1})d\theta} \qquad (3\text{-}28)$$

设 $p(\theta|X^0)=p(\theta)$ 为无样本条件下 θ 的条件概率密度，也就是 θ 的先验概率密度。反复使用式(3-28)，可以得到一个概率密度函数序列 $p(\theta),p(\theta|x_1),p(\theta|x_1,x_2)$ ，等。如果这个密度序列收敛到真实概率密度函数，就把这种性质称为贝叶斯学习。从后面的例子会看到，对于正态分布来说这种性质是存在的。实际上，对很多的典型概率密度函数 $p(x|\theta)$ ，后验概率密度序列也具有这样的性质。

如果分布具有贝叶斯学习性质，那么当样本数 $N\to\infty$ 时，式(3-27)的近似相等就变为严格相等，且此时 θ 的估计量 $\hat{\theta}$ 就等于真实参数 θ，而 $p(x|\theta)$ 也收敛于真实分布函数 $p(x)$，即

$$\lim_{N\to\infty} p(x \mid X^N) = p(x \mid \hat{\theta} = \theta) = p(x)$$

以上只是原则性地讨论了贝叶斯学习问题。下一小节将结合正态分布的例子进行讨论。

刚才已经讨论了反复使用式(3-28)时，可以得到一个概率密度函数序列 $p(\theta), p(\theta \mid x_1), p(\theta \mid x_1, x_2) \ldots$。

注意这是一种迭代方法，因此这被称为参数估计的递推贝叶斯方法。这一方法在增量学习中非常有用。增量学习是要研究这样情况下的学习问题：数据是序贯到达的，并且我们不需要或者无法保存所有数据。例如：在处理某些视频流数据时就是这样。保存视频流数据可能需要使用大量的存储空间。一种实际需求就是仅仅从这些视频流中提取需要的信息来更新一个模型，每次更新模型后就扔掉视频数据。在类似的应用中，我们需要根据已经获得的 N 个数据学习一个模型。当一个新的数据到达的时候，我们只需要适当地更新和修改已经学习到的模型。由于不需要利用所有数据重新学习该模型，因此这样的算法速度通常更快。当模型以一种紧凑的方式表示时，如参数模型，我们还可以不必保存原始数据。

3.3.3 正态分布时参数的贝叶斯估计与贝叶斯学习

为简单起见，这里只考虑单变量正态分布，并假定方差 σ^2 已知，待估计的只是均值 μ。我们以这个例子来说明贝叶斯估计和贝叶斯学习的运用。

1. 贝叶斯估计

设概率分布密度函数为

$$p(x \mid \mu) \sim N(\mu, \sigma^2) \qquad (3\text{-}29)$$

式中只有参数 μ 是未知的。假定我们掌握的关于 μ 的先验知识可用一个已知的先验密度函数 $p(\mu)$ 来表示，并进一步假定 μ 服从均值为 μ_0 方差为 σ_0^2 的正态分布，即

$$p(\mu) \sim N(\mu_0, \sigma_0^2) \qquad (3\text{-}30)$$

其中 μ_0 和 σ_0^2 是已知的。粗略地说，μ_0 表示我们对 μ 的最好的先验推测，σ_0^2 度量了对这个推测的不确定性。假设 μ 的先验分布为正态分布，这将简化后面的数学运算。该问题可以概括为：

设 $X = \{x_1, x_2, \cdots, x_N\}$ 是取自正态分布 $N(\mu, \sigma^2)$ 的样本集，其中 μ 为未知参数，且假定未知参数 μ 是随机变量，它服从先验分布 $N(\mu_0, \sigma_0^2)$，要求我们用贝叶斯估计方法求出 μ 的估计量 $\hat{\mu}$。

根据贝叶斯定理有

$$p(\mu \mid X) = \frac{p(X \mid \mu)\, p(\mu)}{\int p(X \mid \mu)\, p(\mu) d\mu} = \alpha \prod_{k=1}^{N} p(x_k \mid \mu) p(\mu)$$

其中

$$\alpha = 1 \Big/ \int p(X \mid \mu)\, p(\mu) d\mu$$

是一个比例因子，它仅与 X 有关而与 μ 无关。根据（3-29）和（3-30），有

$$p(\mu \mid X)=\alpha\prod_{k=1}^{N}\frac{1}{\sqrt{(2\pi)}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right\}\frac{1}{\sqrt{(2\pi)}\sigma_0}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$$

$$=\alpha^{'}\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{N}\left(\frac{\mu-x_k}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]\right\} \qquad (3\text{-}31)$$

$$=\alpha^{''}\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{N}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}x_k+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

式中和 μ 无关的因子已全部被吸收到 $\alpha^{'}$ 和 $\alpha^{''}$ 中，这样 $p(\mu \mid X)$ 是 μ 的二次函数的指数函数。由于 $p(\mu \mid X)$ 是一个概率密度函数，所以它仍是一个正态密度函数，可以把 $p(\mu \mid X)$ 写成 $N(\mu_N,\sigma_N^2)$，即

$$p(\mu \mid x)=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right] \qquad (3\text{-}32)$$

用待定系数法，令式(3-31)和式(3-32)两式对应的系数相等，即可求得 μ_N 和 σ_N^2

$$\begin{cases}\dfrac{1}{\sigma_N^2}=\dfrac{N}{\sigma^2}+\dfrac{1}{\sigma_0^2}\\[2ex]\dfrac{\mu_N}{\sigma_N^2}=\dfrac{N}{\sigma^2}m_N+\dfrac{\mu_0}{\sigma_0^2}\end{cases} \qquad (3\text{-}33)$$

其中

$$m_N=\frac{1}{N}\sum_{k=1}^{N}x_k \qquad (3\text{-}35)$$

是样本均值。进一步求解得

$$\begin{aligned}\mu_N&=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0\\ \sigma_N^2&=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\end{aligned} \qquad (3\text{-}36)$$

至此，我们已求得 μ 的后验密度 $p(\mu \mid X)$。这样我们有

$$p(x \mid X)=\int p(x \mid \mu)p(\mu \mid X)d\mu$$

$$=\int\frac{1}{\sqrt{(2\pi)}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\frac{1}{\sqrt{(2\pi)}\sigma_N}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right\}d\mu \qquad (3\text{-}37)$$

$$=\frac{1}{2\pi\sigma\sigma_N}\exp\left\{-\frac{1}{2}\frac{(x-\mu_N)^2}{\sigma^2+\sigma_N^2}\right\}f(\sigma,\sigma_N)$$

其中

$$f(\sigma,\sigma_N)=\int \exp\left\{-\frac{1}{2}\frac{\sigma^2+\sigma_N^2}{\sigma^2\sigma_N^2}\left(\mu-\frac{\sigma_N^2 x+\sigma^2\mu_N}{\sigma^2+\sigma_N^2}\right)^2\right\}d\mu \qquad (3\text{-}38)$$

也就是说，$p(x\mid X)$正比于$\exp\left\{-\frac{1}{2}\frac{(x-\mu_N)^2}{\sigma^2+\sigma_N^2}\right\}$。因此$p(x\mid X)$是一个正态分布，即

$$p(x\mid X)\sim N(\mu_N,\sigma^2+\sigma_N^2) \qquad (3\text{-}39)$$

我们看到，当有了N个数据时，该正态分布的均值和方差分别为（3-36）中定义的μ_N和$\sigma^2+\sigma_N^2$。从（3-36）可以知道，μ_N取决于两个量m_N和μ_0。当$\sigma_0=0$时，表示我们对于先验μ_0非常肯定，无论提供多少数据都不能改变我们对于均值的估计。当m_N和μ_0前面的系数都为正数时，由于其和为1，所以，当样本N小的时候，先验μ_0起了较大的作用。而当N变大的时候，先验μ_0的作用越来越小，而m_N的作用越来越大。当N很大的时候，对于μ_N的估计主要由数据决定，也就是由m_N决定。实际上在解决其他问题的贝叶斯方法中，也存在类似的规律。同样我们可以研究一下方差和N的关系。（3-39）中的方差为$\sigma^2+\sigma_N^2$，这是因为当只有N个有限数量的样本时，会给概率分布函数的估计带来不确定性，因此方差增加了σ_N^2。而当N很大时，σ_N^2接近于零。这时，原来的方差σ^2起了主导作用。图3.3给出的是对一维正态分布的均值进行贝叶斯学习的过程。每个后验分布的估计曲线旁边都标记有使用的训练样本的个数。

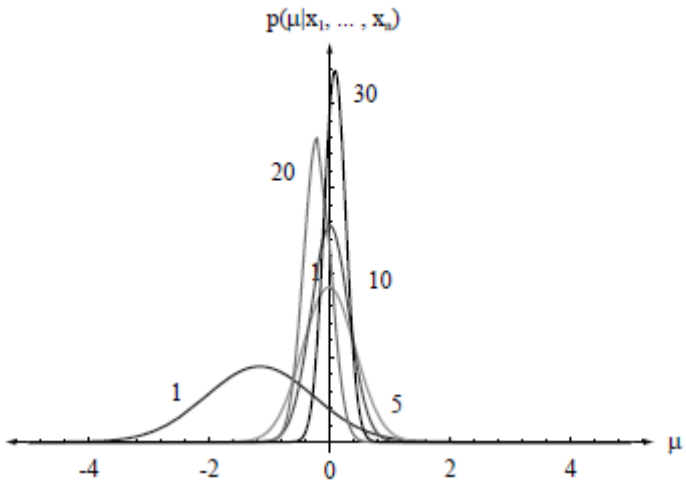


图3.3 对一维正态分布的均值进行贝叶斯学习的过程

3.3 最大似然估计与贝叶斯方法之间的关系

一般来说，当训练样本数无穷多的时候，可以从贝叶斯估计中得到最大似然估计结果。但在实际问题中，训练样本数是有限的，有时候数量是很少的，因此我们应该选用哪一种方法？

总的来说，贝叶斯估计由于使用了先验概率，利用了更多的信息。如果这些信息是可靠的，那么有理由认为贝叶斯估计比最大似然估计的结果更可靠。因此，问题的关键在于先验的获取和表示是否可靠。对于 $p(\theta)$ 的先验知识来自于设计人员对于具体问题的理解和掌握。例如：在考虑一个班学生成绩的分布时，大量的平实观察和经验告诉我们，该班的成绩的平均分更可能出现在 80 分左右，出现在 70，或者 90 的可能性就会比较小，而平均分在 60 以下是几乎不可能的。这时，我们就可以用一个均值在 80 的正态分布来近似这个先验概率。当我们仅仅用采样得到的一个样本 60 去估计该班的平均分时，最大似然估计认为是 60 分，这个结果和我们的经验相差很远，因此，可能带来很大误差。而采用贝叶斯估计结果就会好很多。

贝叶斯估计有时也带来一些困难：有时候先验概率很难设计。例如：估计一个物体在一个区域出现的位置。在没有特别先验知识的时候，我们只好假设该物体在这个区域内处处出现的可能性相同。也就是说，先验概率是这个区域中的均匀分布，或称为“平”的分布。这种情况下贝叶斯估计不会比最大似然估计结果更好。这样的先验常常被称为是无信息先验。采用无信息先验有时候是有问题的，例如，当参数可能取值为一维空间的均匀分布时，有

$$\int p(\theta)d\theta = \infty \quad (3\text{-}40)$$

因此，在实际中，选取先验概率是一个需要研究的问题。

最大似然估计是实际中经常采用的一种方法。最大似然估计方法有几个优点。第一，计算简单。贝叶斯估计方法通常要计算复杂的积分。而相对来说，最大似然估计比较简单。第二，易于理解。最大似然估计给出的是参数的一个最佳估计结果。而贝叶斯估计给出的是结果是一些可行解的加权平均，反映出对于各种可行解的不确定程度。例如：我们日常生活经常使用“平均分”这个术语，但很少使用“平均分的分布”这样的术语。因此，最大似然估计方法更容易被理解和接受。

3.4 高斯混合模型与期望最大算法

3.4.1 高斯混合模型及其参数估计方法

我们看下面这个有多个高斯分布混合在一起构成的一个概率密度函数

$$p(x|\theta) = \sum_{j=1}^{c} p(x|\omega_j, \theta_j)P(\omega_j) \qquad (3\text{-}41)$$

其中 $p(x|\omega_j,\theta_j) \sim N(\mu_i, \Sigma_i)$，$\theta = (\theta_1, \theta_2, ..., \theta_c), \theta_j = (\mu_i, \Sigma_i)$，$P(\omega_j) \geq 0, j = 1, ..., c$，$\sum_{j=1}^{c} P(\omega_j) = 1$，因此有

$$p(x|\theta) = \sum_{j=1}^{c} N(x|\mu_j, \Sigma_j)P(\omega_j) \qquad (3\text{-}42)$$

其中 $N(x|\mu_j,\Sigma_j)$ 表示参数为 μ_j 和 Σ_j 的正态分布函数，$P(\omega_j)$ 称为混合系数。这样的模型被称作高斯混合模型。高斯混合模型在实际中广泛存在。高斯混合模型中每一个高斯密度函数也被称作一个成分，可以代表一个类别，或者一个类别中的一个子类别。而我们观察到的仅仅是这些类别混合在一起的情形。也就是说，我们观察到的样本是没有类别标签的，只是一些不同类别的样本混合在一起。

现在我们讨论这样一个问题：假设具有从式（3-42）表示的模型上独立采样得到一组样本 $\chi=\{x_1,x_2,\cdots,x_N\}$，如何根据这些样本估计高斯混合模型的参数 $\theta_j=(\mu_i,\Sigma_i)$ 和 $P(\omega_j)$。采用最大似然的思路，这时的似然函数为

$$p(X\mid\theta)=\prod_{i=1}^{N}\sum_{j=1}^{c}N(x_i\mid\mu_j,\Sigma_j)P(\omega_j) \qquad (3\text{-}43)$$

取其对数，有

$$\ln p(X\mid\theta)=\sum_{i=1}^{N}\ln\sum_{j=1}^{c}N(x_i\mid\mu_j,\Sigma_j)P(\omega_j) \qquad (3\text{-}44)$$

根据最大似然原则，我们先对参数 μ_k 求偏导，并令其为零，即

$$\frac{\partial\ln p(X\mid\theta)}{\partial\mu_k}=0 \qquad (3\text{-}45)$$

得

$$\sum_{i=1}^{N}\frac{N(x_i\mid\mu_k,\Sigma_k)P(\omega_k)}{\sum_{j=1}^{c}N(x_i\mid\mu_j,\Sigma_j)P(\omega_j)}\Sigma_k^{-1}(x_i-\mu_k)=0 \qquad (3\text{-}46)$$

因为

$$P(\omega_k\mid x_i,\mu_k,\Sigma_k)=\frac{N(x_i\mid\mu_k,\Sigma_k)P(\omega_k)}{\sum_{j=1}^{c}N(x_i\mid\mu_j,\Sigma_j)P(\omega_j)} \qquad (3\text{-}47)$$

方程两边乘以 Σ_k，整理得

$$\mu_k=\frac{\sum_{i=1}^{N}P(\omega_k\mid x_i,\mu_k,\Sigma_k)x_i}{\sum_{i=1}^{N}P(\omega_k\mid x_i,\mu_k,\Sigma_k)} \qquad (3\text{-}48)$$

我们可以采用类似的过程对参数 Σ_k， $P(\omega_k)$ 分别求偏导，并令其为零，整理后得：

$$P(\omega_k)=\frac{1}{N}\sum_{i=1}^{N}P(\omega_k\mid x_i,\mu_k,\Sigma_k) \qquad (3\text{-}49)$$

$$\Sigma_k=\frac{\sum_{i=1}^{N}P(\omega_k\mid x_i,\mu_k,\Sigma_k)(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{N}P(\omega_k\mid x_i,\mu_k,\Sigma_k)} \qquad (3\text{-}50)$$

这样的结果直观上很容易理解。以式（3-48）为例，第 k 类参数均值 μ_k 的估计值就是所有样本 x_i 的加权平均。权重就是每个样本属于第 k 类的可能性。可以类似地解释式（3-49）和（3-50）。

需要说明的是，式（3-48）、（3-49）和（3-50）并没有给出估计混合模型参数的闭式解。以式（3-48）为例，在估计 μ_k 时，样本 x_i 的权重系数依赖于 μ_k 和其他参数的。不过，我们可以根据这些结果给出一种简单的迭代算法来求解这个最大似然问题。我们首先为均值，方差，混合系数指定初始值。然后在轮换执行下面两个被称为 E 步骤和 M 步骤的更新过程。在 E 步骤中，我们用当前的参数值来估计后验概率。然后在 M 步骤中，用这些概率通过（3-51）~（3-53）来重新估计均值 ，协方差和混合系数。如果每次更新后似然值的改变小于某个阈值我们就认为算法已经收敛。

算法：高斯混合模型参数估计的 EM 算法

给定高斯混合模型和在该模型上独立采样得到的一组样本，任务是通过最大化似然函数来估计模型参数（包括均值，协方差以及混合系数）。

1. 初始化均值 μ_k^{new}，协方差 Σ_k^{new}，混合系数 $P^{new}(\omega_k)$，并计算初始对数似然值。

2. E 步骤。 $\mu_k = \mu_k^{new}$， $\Sigma_k = \Sigma_k^{new}$， $P(\omega_k) = P^{new}(\omega_k)$，并计算后验概率

$$P(\omega_k \mid x_i, \mu_k, \Sigma_k)$$

3. M 步骤。用新的后验概率重新估计参数

$$P^{new}(\omega_k) = \frac{1}{N}\sum_{i=1}^{N} P(\omega_k \mid x_i, \mu_k, \Sigma_k) \qquad (3\text{-}51)$$

$$\mu_k^{new} = \frac{\sum_{i=1}^{N} P(\omega_k \mid x_i, \mu_k, \Sigma_k) x_i}{\sum_{i=1}^{N} P(\omega_k \mid x_i, \mu_k, \Sigma_k)} \qquad (3\text{-}52)$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N} P(\omega_k \mid x_i, \mu_k, \Sigma_k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} P(\omega_k \mid x_i, \mu_k, \Sigma_k)} \qquad (3\text{-}53)$$

4. 计算似然函数，并对似然函数值或参数值检查是否收敛。如果收敛则算法停止，输出均值 μ_k^{new}，协方差 Σ_k^{new} 和混合系数 $P^{new}(\omega_k)$。否则回到步骤 2。

从下面一小节的讨论可以知道，该算法就是期望最大（EM）算法在高斯混合模型上的一个应用，这是该算法名字的由来。根据 EM 算法的性质可以知道，该算法通过每次的 E 步骤和 M 步骤更新之后其对数似然函数一定是递增的。

需要注意的是，用最大似然方法来估计高斯混合模型的参数可能存在奇异性的问题。为简单起见，我们假定高斯混合中的每个成分高斯函数的协方差矩阵都是单位阵，即 $\Sigma_j = \sigma_j^2 I$，其中 I 是单位阵。当然后面的

结论对任意的协方差阵都是成立的。假定混合模型中的第 j 个成分的均值正好等于某个数据点，也就是对于某个 i 有 $\mu_j = x_i$。那么这个数据点对似然函数的贡献项为

$$N(x_i \mid x_i, \sigma_j^2 I) \propto \frac{1}{\sigma_j} \qquad (3\text{-}54)$$

考虑 σ_j 趋于 0，那么这一项，也就是说整个似然函数将趋于无穷。当某个高斯成分退化到某个数据点上时，这种奇异性就会发生。也就是说，如果高斯混合模型中的成分数多于一个，只要其中一个的方差为某个非 0 有限值，它就会给每个数据点都分配一个有限概率值，这时如果另一个成分退化到了某个数据点上且方差趋近于 0，总似然函数就会不断增大。图 3.4 说明了这一问题。这表明最大似然估计可能导致严重的过拟合问题。关于过拟合问题，我们会在本章最后一节讨论。我们可以用适当的启发信息来避免这个现象的出现，比如在某个成分退化的时候将它的均值重新设到某个随机位置同时将它的方差重新设到某个较大值。当然，如果采用贝叶斯方法也可以避免这个问题。

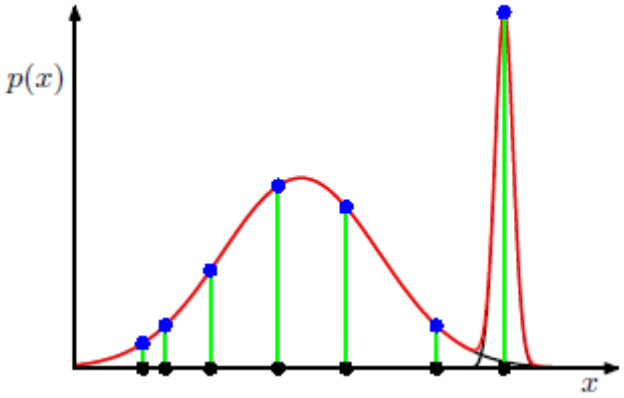
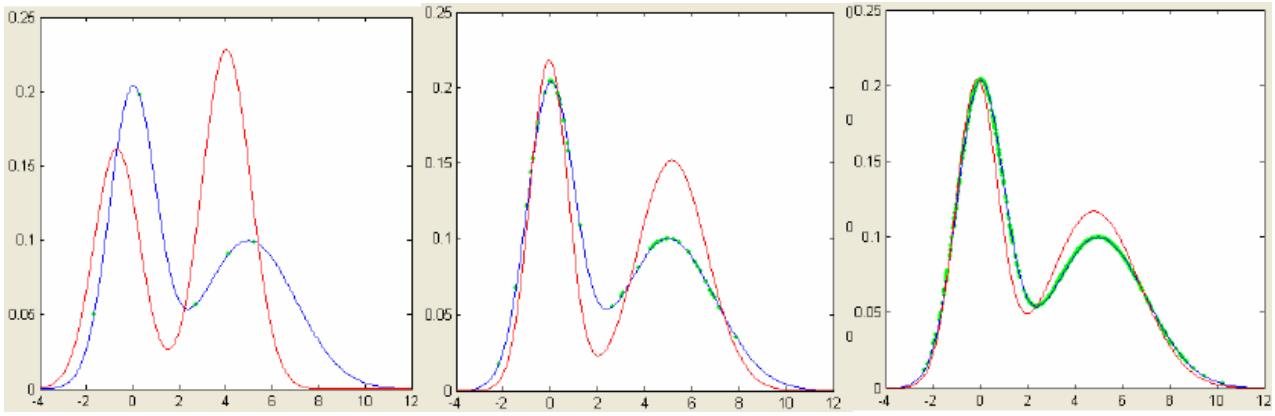


图 3.4 高斯混合模型参数估计中的奇异性。圆点表示采样得到的数据。右边的高斯成分由于只包含了一个数据点，因此，出现了奇异性。

需要注意的另一个问题是，对于成分数为 C 的混合模型的某个最大似然解，因为我们需要将 C 组参数赋给 C 个成分，一共就有 $C!$ 个等价解。也就是说，对于解空间中的一个（非退化）点，都存在 $C!-1$ 个其他点使得模型具有相同的分布。这称为可识别性问题。这会在本节最后一个小节讨论这个问题。该问题在需要对模型参数解释时很重要。不过，在很多情况下这个问题无关紧要，因为所有这些解都一样好。

图 3.5 给出了使用 EM 算法估计两个高斯成分混合时的参数的示例。图 3.5（a）是利用从一个混合高斯分布中采样得到的 5 个样本来估计这个分布的结果。蓝色曲线是原来的混合分布，圆点表示的是采样得到的 5 个样本，红色表示的是估计出的分布。图 3.5（b）和（c）分别是利用了 50 和 500 个这样的样本来估计这个分布的结果。可以看到，使用 50 个点时的结果已经不错。



（a）使用 5 个点估计的结果　　（b）使用 50 个点估计的结果　　（b）使用 500 个点估计的结果。

图 3.5 用 EM 算法估计两个高斯成分混合时的参数的结果。

3.4.2 期望最大算法

本章的任务是利用观察样本估计概率密度函数。在实际中我们得到的观察数据往往是不完整的。一种情况是有些数据是无法、或者是没有直接被观察到，例如，一些特殊场合的温度无法直接测量。在对一个人讲话进行录音时，讲话人的情感状态一般也不会直接得到。这样的一类变量（如温度，情感状态）被称作是隐变量。另一种情况是在采集数据的过程中，由于一些原因有些数据没有采集到。例如，在问卷调查时，有的调查问卷的某项被忘记填写，或者有的传感器出现故障而没有数据输出。我们把这些没有观察到的数据统称为缺失数据。

期望最大算法就是利用不完整数据实现最大似然估计的一种方法。该方法用隐变量对缺失数据建模来求解。当然，有时候引入隐变量是出于技术上的考虑，也就是说，在使用隐变量后，一些实际问题的求解变得更简单了。

假设我们得到的观察数据 $\chi=\{x_1,x_2,...,x_N\}$ 是不完整的，Y 是没有被观察到的变量，完整的变量应该是 $Z=(X,Y)$。我们假设 Y 是离散变量。对于连续变量我们也可以类似地推导。因此有

$$p(X\mid\theta)=\sum_{Y}p(X,Y\mid\theta)\qquad(3\text{-}55)$$

对式（3-55）两边取对数，有

$$L(\theta)=\log p(X\mid\theta)=\log\sum_{Y}p(X,Y\mid\theta)\qquad(3\text{-}56)$$

我们假设直接优化 $p(X\mid\theta)$ 是很困难的，但是优化完全数据的似然函数 $p(X,Y\mid\theta)$ 要简单得多。下面引入隐变量分布 $q(Y)$，对于任意分布 $q(Y)$，有

$$L(\theta)\geq\sum_{y}q(y)\log p(x,y\mid\theta)-\sum_{y}q(y)\log q(y)=F(q,\theta)\qquad(3\text{-}57)$$

这是因为

$$
\begin{aligned}
L(\theta) &= \log \sum_y p(x,y\mid\theta) = \log \sum_y q(y)\frac{p(x,y\mid\theta)}{q(y)} \\
&\geq \sum_y q(y)\log\frac{p(x,y\mid\theta)}{q(y)} \\
&= \sum_y q(y)\log p(x,y\mid\theta) - \sum_y q(y)\log q(y) \\
&= F(q,\theta)
\end{aligned}
\qquad (3\text{-}58)
$$

其中第二行利用了 Jensen 不等式。这样我们就得到了 $L(\theta)$ 的一个下界。

由于直接优化 $L(\theta)$ 是很困难的，我们转而优化 $L(\theta)$ 的下界函数 $F(q,\theta)$。优化函数 $F(q,\theta)$ 可能也是比较困难的。因此，我们采用一种简单的迭代算法对 $F(q,\theta)$ 寻优。即首先对变量 q，θ 初始化，然后通过先固定变量 $\theta_{[k]}$，寻找能够最大化函数 F 的参数 $q_{[k+1]}$。然后固定参数 $q_{[k+1]}$，寻找能够最大化函数 F 的参数 $\theta_{[k+1]}$。即反复执行下面的两个步骤：

$$
q_{[k+1]} \leftarrow \arg\max_q F(q,\theta_{[k]}) \qquad (3\text{-}59)
$$

$$
\theta_{[k+1]} \leftarrow \arg\max_\theta F(q_{[k+1]},\theta) \qquad (3\text{-}60)
$$

可以证明当 $q_{[k+1]}(y) = p(y\mid x,\theta_{[k]})$ 时，式（3-59）中的函数 F 达到最大。这是因为：

$$
\begin{aligned}
F(q,\theta) &= \sum_y q(y)\log\frac{p(y,x\mid\theta)}{q(y)} \\
&= \sum_y p(y\mid x,\theta)\log\frac{p(y\mid x,\theta)p(x\mid\theta)}{p(y\mid x,\theta)} \\
&= \sum_y p(y\mid x,\theta)\log p(x\mid\theta) \\
&= \log p(x\mid\theta) = L(\theta)
\end{aligned}
\qquad (3\text{-}61)
$$

当 $q_{[k+1]}(y) = p(y\mid x,\theta_{[k]})$ 时，$F(q_{[k+1]},\theta) = \sum_y q_{[k+1]}(y)\log p(x,y\mid\theta) - \sum_y q_{[k+1]}(y)\log q_{[k+1]}(y)$。由于第二项不包含需要优化的变量，所以这时的算法就可以简化如下。

算法：期望最大（EM）算法

1. 初始化变量 q，θ。

2. E 步骤。计算函数

$$
Q(\theta_{[k]},\theta) = \sum_y p(y\mid x,\theta_{[k]})\log p(x,y\mid\theta) = E[\ln p(x,y\mid\theta)\mid x,\theta_{[k]}] \qquad (3\text{-}62)
$$

3. M 步骤。

$$\theta_{[k+1]} \leftarrow \arg\max_{\theta} Q(\theta_{[k]}, \theta) \qquad (3\text{-}63)$$

4. 如果算法收敛则停止，否则回到步骤 2。

由于式（3-62）是在计算期望，所以这一步叫做 E 步骤（Expectation Step），式（3-62）是最大化一个函数，所以叫做 M 步骤。这是该算法名称的由来。在本教材的后面的部分，简称该算法为 EM 算法。

在 EM 算法中的 M 步骤，计算最大化 Q 有时是困难的。一种可以替代的方法是在迭代过程中不是求取最大值，而是只保证函数 Q 递增就可以了。而这有时更容易实现。这种方法称作是广义 EM（Generalized EM）算法。

算法的收敛性是一个迭代算法的重要性质。可以证明 EM 算法（包括广义 EM 算法）的每一次迭代都使得似然函数不减少，并且可以证明算法可以收敛到似然函数的一个局部极大值。

可以从不同的角度理解 EM 算法。下面从参数空间的角度来看 EM 算法中的操作。如图 3.6 所示。这里红色曲线为我们想要优化的（不完全数据的）对数似然函数。我们从某个参数初始值 θ^{old} 开始，在 E 步骤中计算下界 F，如蓝色曲线所示。这个下界在 θ^{old} 处和对数似然函数相等，而且和对数似然函数相切。所以它们在此处有相同的梯度。在 M 步骤中，我们最大化这个界并得到 θ^{new}，此处的对数似然值比 θ^{old} 处的大。在接下来的 E 步骤中再构建新的界（绿色曲线）使得这个界与似然函数在 θ^{new} 处相切。

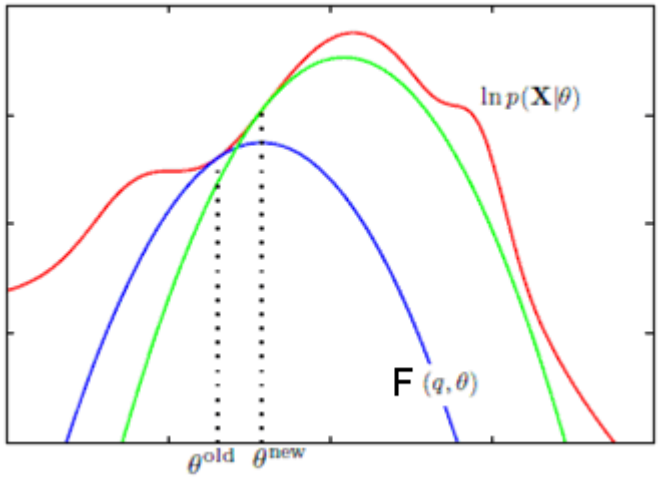


图 3.6 从参数空间看 EM 算法

在对高斯混合模型的参数进行最大似然求解时，可以采用 EM 算法。这时我们引入一个潜变量 $y_i \in 1,...,C$，该变量表示第 i 个样本 x_i 是从第 y_i 个高斯成分中产生的。这时有

$$\log p(X,Y|\theta)=\sum_{i=1}^{N}\log(p(x_i|y_i)P(y_i)) \qquad (3\text{-}64)$$

通过数学推导，可得 E 步骤的 $Q(\theta_{[k]},\theta)$ 函数为

$$\begin{aligned}Q(\theta_{[k]},\theta)&=\sum_{l=1}^{C}\sum_{i=1}^{N}In(\alpha_l p_l(x_i|\theta_l))p(l|x_j,\theta_{[k]})\\&=\sum_{l=1}^{C}\sum_{i=1}^{N}\ln(\alpha_l)p(l|x_i,\theta_{[k]})+\sum_{l=1}^{C}\sum_{i=1}^{N}\ln(p_l(x_i|\theta_l))p(l|x_i,\theta_{[k]})\end{aligned} \qquad (3\text{-}65)$$

其中 $\alpha_l=P(\omega_l)$，l 是一个类别变量。下面计算 M 步骤。我们把高斯密度函数的形式带入（3-65），通过对相应参数求导，令导数为零，就可以得到（3-51）、（3-52）和（3-53）的更新公式。这些数学推导不在此赘述，我们把它做为习题留给读者。

可以得到前面得到的“高斯混合模型参数估计的 EM 算法”。从 EM 算法的角度很容易得到这样的结论：在每一次迭代中，高斯混合模型的似然函数是不降低的，算法会收敛到一个局部极值。

EM 算法经常用于一些混合模型的参数估计，如高斯混合模型，伯努利混合模型，以及其他的一些有潜变量的模型的参数估计中，如贝叶斯网络中的参数估计。需要说明的是，由于在推导式（3-65）的过程中我们没有使用函数 $p(x|\theta)$ 的具体形式，因此式（3-65）给出的结果也适合于其他的混合模型，如伯努利混合模型。

3.4.3 可识别性问题

我们的基本任务是利用从一个混合密度中抽取的样本来估计未知参数向量 θ。我们会问，从混合密度中是否有可能唯一地确定参数 θ。这就是我们现在要讨论的可识别性问题。如果能产生混合密度 $p(x|\theta)$ 的 θ 值只有一个，那么原则上存在唯一解。如果 θ 的几个不同值都能产生相同的函数 $p(x|\theta)$，那么要得到 θ 的唯一解就不可能了。对于可识别的函数的定义为：如果有两个参数 $\theta_1\neq\theta_2$，则一定存在一个 x，使得 $p(x|\theta_1)\neq p(x|\theta_2)$，则称函数 $p(x|\theta)$ 是可识别的。需要说明的是，可识别性是函数本身的性质，和参数的估计方法无关。

下面举一个不可识别的函数的例子。如果 x 是取 0 或 1 的离散随机变量，$p(x|\theta)$ 是如下混合概率密度函数

$$p(x|\theta)=\frac{1}{2}\theta_1^x(1-\theta_1)^{1-x}+\frac{1}{2}\theta_2^x(1-\theta_2)^{1-x}=\begin{cases}\frac{1}{2}(\theta_1+\theta_2),\text{当}x=1\\1-\frac{1}{2}(\theta_1+\theta_2),\text{当}x=0\end{cases} \qquad (3\text{-}66)$$

如果我们知道 $p(x=1|\theta)=0.6$，$p(x=0|\theta)=0.4$，这样就知道了混合概率 $p(x|\theta)$。但我们无法确定 θ 的分量，即无法把 θ 唯一地分解为确定的 θ_1 和 θ_2。这是因为在 $p(x=1|\theta)=0.6$ 和 $p(x=0|\theta)=0.4$ 情况下，由式(3-66)所能得到的独立方程只有一个，即 $\theta_1+\theta_2=1.2$，而未知数却有两个。这时，就出现了对不

同的 $\theta=[\theta_1,\theta_2]^T$，可使 $p(x|\theta)=p(x|\theta')$。这时，混合分布的参数估计就不可能实现。

幸运的是，大部分常见的连续随机变量的概率密度函数都是可识别的，而离散随机变量的混合概率函数则往往是不可识别的。在混合分布中，当未知参数 θ 的个数多于独立方程的数目时就会出现不可识别的情况。

3.5 非参数估计方法

上面讨论的参数估计方法要求已知概率分布函数的形式。然而在很多实际问题中并不知道函数的形式，或函数的形式不是一些通常遇到的概率分布，不能写成某些参数的函数。这时，就可以直接用样本来估计概率分布函数。这样的方法称之为概率密度函数的非参数方法。

3.5.1 基本方法

我们首先不很严格地举例说明这种估计方法的思路。假如样本集由三个一维样本组成，即 $X=\{x_1,x_2,x_3\}$，每个样本 x_i 在以 x_i 为中心，宽度为 h 的范围内，对分布有贡献，贡献量为 a。若要估计 x 点的密度 $p(x)$，可把每个样本在 x 点的贡献相加作为该点的密度 $p(x)$ 的近似。对一维上所有的实数点都这样计算，就可以得到密度函数 $p(x)$ 的估计 $\hat{p}(x)$。这个例子的 $\hat{p}(x)$ 如图 3.7 中虚线所示。每个阴影的矩形面积等于 a，相当于每个样本对分布所作出的贡献。当样本数 N 很大时，估计结果将会很好。当然也可以认为每个样本对于自己所在位置的分布贡献最大，而离得越远，则贡献越小。所以贡献也可以表示为在 x_i 处最大，而往两边越来越小的函数形式，如图 3.8 所示。

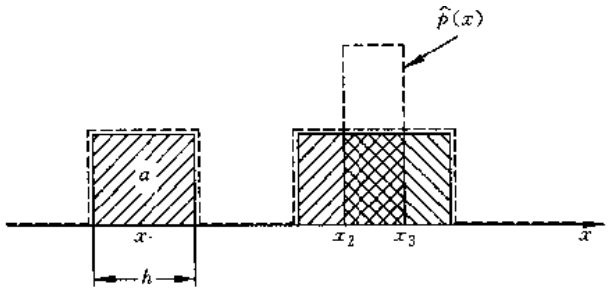


图 3.7 非参数估计的基本思路

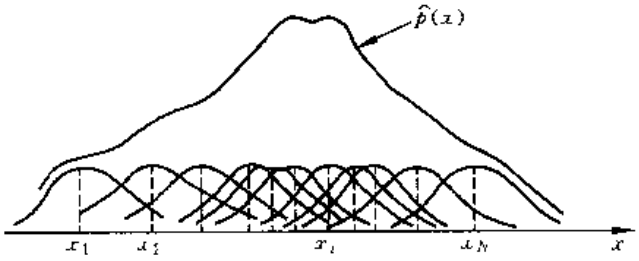


图 3.8 非参数估计中每个样本的贡献随距离而发生变化

下面讨论一般的非参数估计方法。我们的目的是从样本集 X 估计样本空间任何一点的概率密度 $\hat{p}(x)$。

如果样本集 X 来自某一类别(如 ω_j 类)，则估计结果为类条件概率密度 $\hat{p}(x|\omega_j)$。如果样本集 X 来自 c 个类别，而又分不清那个样本来自哪一类，则估计结果为混合密度 $\hat{p}(x)$。

估计概率分布函数有许多方法，但它们的基本思想都很简单，但是要严格证明这些估计的收敛性有时是不容易的。假设 R 为密度函数 $p(x)$ 定义域中的一个区域，那么随机向量 x 落入 R 的概率 P 为

$$P=\int_R p(x)dx \qquad (3\text{-}67)$$

若有 N 个样本 $x_1,x_2,\cdots,x_N$ 是从密度函数 $p(x)$ 中独立抽取的，则 N 个样本中有 k 个落入区域 R 中的概率 P_k 服从二项分布，有

$$P_k=C_N^k P^k(1-P)^{N-k} \qquad (3\text{-}68)$$

其中 P 为样本 x 落入区域 R 的概率， P_k 为 k 个样本落入区域 R 的概率，

$$C_N^k=\frac{N!}{k!(N-k)!} \qquad (3\text{-}69)$$

可以知道 k 的期望为

$$E[k]=NP \qquad (3\text{-}70)$$

根据二项分布的性质可知 k 的众数（定义：对于有频数分布的变量，它的众数为频数最大的变量的值）m 为 $(N+1)P$ 的整数部分，即

$$m=[(N+1)P] \qquad (3\text{-}71)$$

也即 $k=m$ 时， P_k 有最大值

$$P_m=\max P_k \qquad (3\text{-}72)$$

根据众数的定义说明， $k=m$ 个落入区域 R 的概率最大。可取

$$k=m\approx(N+1)\hat{P}\approx N\hat{P} \qquad (3\text{-}73)$$

由式(3-73)得

$$\hat{P}\approx\frac{k}{N} \qquad (3\text{-}74)$$

我们认为 kN 是 P 的一个很好的估计，也就是密度函数 $p(x)$ 在区域 R 上的一个很好的估计。而我们实际上要估计的是密度函数 $p(x)$ 在点 x 处的值 $\hat{p}(x)$。为此假设 $p(x)$ 连续，并且区域 R 足够小，以至使 $p(x)$ 在区域 R 中近似相同，那么我们可得

$$P=\int_R p(x)dx=p(x)V \qquad (3\text{-}75)$$

其中V是区域R的体积，x是R中的点。将式(3-74)和式(3-75)结合，则有

$$\frac{k}{N} \approx \hat{P} = \int_R \hat{p}(x)dx = \hat{p}(x)V \qquad (3\text{-}76)$$

因此

$$\hat{p}(x) = \frac{k/N}{V} \qquad (3\text{-}77)$$

式(3-77)就是概率密度$p(x)$在x点处的估计值，它与样本数N、包含x的区域R的体积V及落入V中的样本数k有关。

现在需要说明几个理论上和实际中的问题。如果把体积V固定，样本取得越来越多，则比值$\frac{k}{N}$会收敛，但我们只能得到$p(x)$在区域空间V上的一个平均估计

$$\frac{\hat{P}}{V} = \frac{\int_R \hat{p}(x)dx}{\int_R dx} \qquad (3\text{-}78)$$

要想得到$p(x)$在x点处的估计值$\hat{p}(x)$，则必须让体积V趋于零。但若把样本数目固定，而令V趋于零，就会使区域R不断缩小以致于最后可能不包含任何样本，这样就会得出$\hat{p}(x) \approx 0$这种没有什么价值的估计。如果碰巧有一个或几个样本同x点重合，则估计值就会为无穷大，这同样也是没有意义的。

在实际中样本数总是有限的，所以体积V不允许任意小。因此若采用这种估计的话，$\frac{k}{N}$和$\hat{p}(x)$将存在随机性，也就是说$\frac{k}{N}$和$\hat{p}(x))$都有一定的方差。

但如果只从理论上考虑，假定有无限多的样本可供利用，那么情况会怎样呢?比如我们采用下面的步骤进行：为了估计x点的密度，我们构造一串包括x的区域序列$R_1, R_2, \cdots, R_N, \cdots$，然后对$R_1$采用一个样本进行估计，对$R_2$采用两个样本，…。设$V_N$是$R_N$的体积，$k_N$是落入在$R_N$中的样本数，$\hat{p}_N(x)$是$p(x)$的第$N$次估计，则

$$\hat{p}_N(x) = \frac{k_N/N}{V_N} \qquad (3\text{-}79)$$

若满足以下三个条件，则$\hat{p}_N(x)$收敛于$p(x)$。

(1) $\lim_{N\to\infty} V_N = 0$ （3-80）

(2) $\lim_{N\to\infty} k_N = \infty$ （3-81）

(3) $\lim_{N\to\infty} \frac{k_N}{N} = 0$ （3-82）

第二个条件表明只要区域平滑地缩小，同时$p(x)$在x点连续，就可以使空间平均P/V收敛于$p(x)$。

第二个条件对 $p(x)\neq 0$ 的点有意义，此时可使频率之比收敛于概率 P 。另外，如果要使式(3-97)中的 $\hat{p}_N(x)$ 收敛，则第三个条件显然是必要的，它表明，尽管在一个小区域 R_N 内最终落入了大量样本，但同样本总数相比仍然是很少的。

满足上述三个条件的区域序列一般有两种选择方法，从而得到两种非参数估计方法。

(1) Parzen 窗法。使区域序列 V_N 根据 N 的某个函数(例如 $V_N=\frac{1}{\sqrt{N}}$)不断缩小。但这时对 k_N 和 $\frac{k_N}{N}$ 都要加些限制条件以使 $\hat{p}_N(x)$ 收敛于 $p(x)$ 。

(2) k_N 近邻估计法。让 k_N 根据 N 的某个函数(例如 $k_N=\sqrt{N}$)增大，而 V_N 的选取是使相应的 R_N 正好包含 x 的 k_N 个近邻。下面我们分别讨论这两种方法。

3.5.2 Parzen 窗法

1. Parzen 窗估计的概念

我们暂时假设区域 R_N 是一个 d 维超立方体。如果 h_N 是超立方体的棱长，则该超立方体的体积为

$$V_N=h_N^d \qquad (3\text{-}83)$$

定义窗函数 $\varphi(u)$

$$\varphi(u)=\begin{cases}1, 当|u_j|\leq\frac{1}{2}, j=1,2,\cdots,d\\ 0, 其他\end{cases} \qquad (3\text{-}84)$$

利用式(3−84)可把落入该超立方体内的样本数解析地表示出来。由于 $\varphi(u)$ 是以原点为中心的一个超立方体，所以当 x_i 落在以 x 为中心、体积为 V_N 的超立方体内时，$\varphi(u)=\varphi(\frac{x-x_i}{h_N})=1$，否则为 0。因此落入该超立方体内的样本数为

$$k_N=\sum_{i=1}^{N}\varphi\left(\frac{x-x_i}{h_N}\right) \qquad (3\text{-}85)$$

将式(3−113)代入式(3−107)，得

$$\hat{p}_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{x-x_i}{h_N}\right) \qquad (3\text{-}86)$$

这就是 Parzen 窗法估计的基本公式。它表示 $p(x)$ 的估计可以看作是 x 和 x_i 的函数的一种平均。实质上，窗函数的作用是插值，每一样本对估计所起的作用依赖于它到 x 的距离。

2. 估计量 $\hat{p}_N(x)$ 为密度函数的条件

估计量 $\hat{p}_N(x)$ 为一个合理的密度函数的条件是：它处处非负且积分为 1。为此，我们需要限制窗函数满足下面两个条件：

(1) $\varphi(u) \geq 0$ （3-87）

(2) $\int \varphi(u)du = 1$ （3-88）

即如果窗函数本身具有密度函数的形式，则 $\hat{p}_N(x)$ 一定为密度函数。从式(3-86)可以看出，在 $\varphi(u) \geq 0$ 时，自然有 $\hat{p}_N(x)$ 非负。利用条件式(3-88)可以证明 $\int \hat{p}_N(x)dx = 1$，因为

$$\int \hat{p}_N(x)dx = \int \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{x-x_i}{h_N}\right)dx = \frac{1}{N}\sum_{i=1}^{N}\int \frac{1}{V_N}\varphi\left(\frac{x-x_i}{h_N}\right)dx$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(u)du = \frac{1}{N}\cdot N = 1 \left(\text{其中} u = \frac{x-x_i}{h_N}\right)$$

从而证明了 $\hat{p}_N(x)$ 确实是密度函数。

3. 窗函数的选择

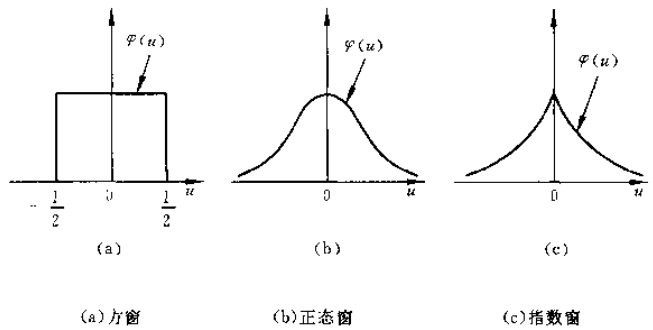
式（3-84）的超立方体窗函数一般称为方窗，见图 3.9（a）。除此以外，还可选择其他的窗函数。举几个一维例子如下：

(1) 正态窗函数(见图 3.7(b))

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}$$

(2) 指数窗函数(见图 3.7(c))

$$\varphi(u) = \frac{1}{2}\exp\{-|u|\}$$



(a)方窗 (b)正态窗 (c)指数窗

图 3.9 几种窗函数

总之，只要所选择的函数满足条件式(3-87)和式(3-88)，都可以作为窗函数使用，但最终估计效果的好坏

则与实际所采样本情况和窗函数及其参数的选择有关。

4.窗宽 h_N 对估计量 $\hat{p}_N(x)$ 的影响

在样本数 N 有限时，窗宽 h_N 对估计量会有很大影响。现在让我们分析一下原因。

如果定义函数 $\delta_N(x)$ 为

$$\delta_N(x)=\frac{1}{V_N}\varphi\left(\frac{x}{h_N}\right) \qquad (3\text{-}89)$$

则可以把 $\hat{p}_N(x)$ 看作一个平均值

$$\hat{p}_N(x)=\frac{1}{N}\sum_{i=1}^{N}\delta_N(x-x_i) \qquad (3\text{-}90)$$

因为 $V_N=h_N^d$，所以 h_N 影响了 $\delta_N(x)$ 的幅度。若 h_N 很大，则 δ_N 的幅度就很小，同时只有 x_i 离 x 较远时才能使 $\delta_N(x-x_i)$ 同 $\delta_N(0)$ 相差得多一些。这时 $\hat{p}_N(x)$ 变成 N 个宽度较大且函数值变化缓慢的函数的叠加，从而它是 $p(x)$ 的一个平均估计，使估计的分辨力降低。如图 3.8 第 3 列。

反之，若 h_N 很小，则 $\delta_N(x-x_i)$ 的幅值较大，这时 $\hat{p}_N(x)$ 就成了 N 个以样本为中心的尖峰函数的叠加，使估计的函数变动很大，如图 3.8 第 1 列。在 $h_N\to 0$ 的极端情况，$\delta_N(x-x_i)$ 趋于一个以 x_i 为中心的 δ 函数，从而使 $\hat{p}_N(x)$ 趋于以样本为中心的 δ 函数的叠加。

在实际问题中样本都是有限的，因此如何选取 h_N 需要一定的经验，一般要作适当折中考虑。

5.估计量 $\hat{p}_N(x)$ 的统计性质

估计量 $\hat{p}_N(x)$ 的性能可以用估计量的统计性质来表示。可以证明在一定条件下，估计量 $\hat{p}_N(x)$ 收敛于 $p(x)$。这些条件除了式（3-87）和（3-88）外，还包括

(1) 总体密度 $p(x)$ 在 x 点连续；

(2) 窗函数要满足下列条件

$$\sup_u \varphi(u)<\infty \qquad (3\text{-}91)$$

$$\lim_{\|u\|\to\infty}\varphi(u)\prod_{i=1}^{d}u_i=0 \qquad (3\text{-}92)$$

(3) 窗宽受下列条件约束

$$\lim_{N\to\infty} V_N = 0 \qquad (3\text{-}93)$$

$$\lim_{N\to\infty} NV_N = \infty \qquad (3\text{-}94)$$

式(3-91)说明 $\varphi(u)$ 是有界的，不能为无穷大；式(3-92)说明 $\varphi(u)$ 随 u 的增加将较快趋于零。它们都保证窗函数 $\varphi(u)$ 有较好的性质。式(3-93)和式(3-94)要求体积随 N 的增大而趋于零，但缩减的速度又不要太快，其速率要低于 $\frac{1}{N}$。下面我们分别考虑均值和方差的收敛性。

首先考虑均值 $\overline{p}_N(x)$ 的收敛性。由于样本 x_i 是从 $p(x)$ 中以独立同分布方式得到的，所以有

$$\begin{aligned}\overline{p}_N(x) &= E[p_N(x)] = \frac{1}{N}\sum_{i=1}^{N} E[\frac{1}{V_N}\varphi\left(\frac{x-x_i}{h_N}\right)] \\ &= \int \frac{1}{V_N}\varphi\left(\frac{x-v}{h_N}\right)p(v)dv = \int \delta_N(x-v)p(v)dv\end{aligned} \qquad (3\text{-}95)$$

式（3-95）表明均值是概率密度函数 $p(x)$ 的平均，也就是 $p(x)$ 被窗函数平滑后的结果。当 V_N 趋于零时，$\delta_N(x-v)$ 趋近于一个中心在 x 狄拉克函数，即在 x 处为 1，其他地方为 0 的函数。这样，如果 $p(x)$ 在 x 处连续，则（3-93）保证了 $\overline{p}_N(x)$ 在 N 趋于无穷大时收敛于 $p(x)$。

其次我们关心每一次的估计 $p_N(x)$ 和 $p(x)$ 可能相差多少。为此考虑 $p_N(x)$ 的方差。我们有

$$\begin{aligned}\sigma_N^2(x) &= \sum_{i=1}^{N} E[(\frac{1}{NV_N}\varphi\left(\frac{x-x_i}{h_N}\right) - \frac{1}{N}\overline{p}_N(x))^2] \\ &= NE[\frac{1}{N^2V_N^2}\varphi^2\left(\frac{x-x_i}{h_N}\right)] - \frac{1}{N}\overline{p}_N^2(x) \\ &= \frac{1}{NV_N}\int \frac{1}{V_N}\varphi^2\left(\frac{x-v}{h_N}\right)p(v)dv - \frac{1}{N}\overline{p}_N^2(x)\end{aligned} \qquad (3\text{-}96)$$

使用式（3-95），有

$$\sigma_N^2(x) \le \frac{\sup(\varphi)\overline{p}_N(x)}{NV_N} - \frac{1}{N}\overline{p}_N^2(x) \qquad (3\text{-}97)$$

因此，为了减小方差，就需要 V_N 较大。直观来看，V_N 较大时可以把概率密度函数局部的变化平滑掉。只有当 N 趋于无穷大时，才可能让 V_N 逐渐趋于零，只要 NV_N 趋于无穷，方差仍然可以很小。

上面这些结论都是样本数趋于无穷时的结果。而在有限样本情况下如何选择窗函数和 V_N，还是一个问题。

6.Parzen 窗法应用示例

下面通过两个简单的例子来看一下 Parzen 窗方法。在例子中，我们用已知的密度函数产生一系列样本，然后分析考查用 Parzen 窗法估计出的密度与真实密度的关系。

先考虑 $p(x)$ 是均值为零，方差为 1 的一维正态分布，设窗函数也选择正态窗函数即

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}$$

再设 $h_N=h_1/\sqrt{N}$，h_1 在例子中是可调节的参量，我们将对 h_1 取不同值以比较它的影响。这样，$\hat{p}_N(x)$ 就是一个以样本为中心的正态密度函数的平均

$$\hat{p}_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N}\varphi\left(\frac{x-x_i}{h_N}\right)$$

一旦正态分布随机样本产生出来后，就可以估计 $\hat{p}_N(x)$，所得结果示于图 3.8，这些结果依赖于 N 和 h_1。当 $N=1$时，$\hat{p}_N(x)$ 是一个以第一个样本为中心的正态形状的小丘；当 $N=16$，$h_1=\frac{1}{4}$ 时，仍可看出各个单个样本所起的作用；但对于 $h_1=1$和 $h_1=4$，各个单个样本所起的作用就模糊了。随着 N 的增加，估计量 $\hat{p}_N(x)$ 越来越好。在样本数不很多，采样又不很规则的情况下，也会在 $\hat{p}_N(x)$ 中出现一些不规则的扰动。但当样本数 N 趋于无穷时，$\hat{p}_N(x)$ 就会收敛于平滑的正态曲线。这说明要想得到较精确的估计，就需要有大量的样本。

在第二个例子里，$\varphi(u)$ 和 h_N 与第一个例子一样，但未知密度假定是两个均匀分布密度的混合密度

$$p(x)=\begin{cases}1, & -2.5<x<-2\\ 0.25, & 0<x<2\\ 0, & \text{其他}\end{cases}$$

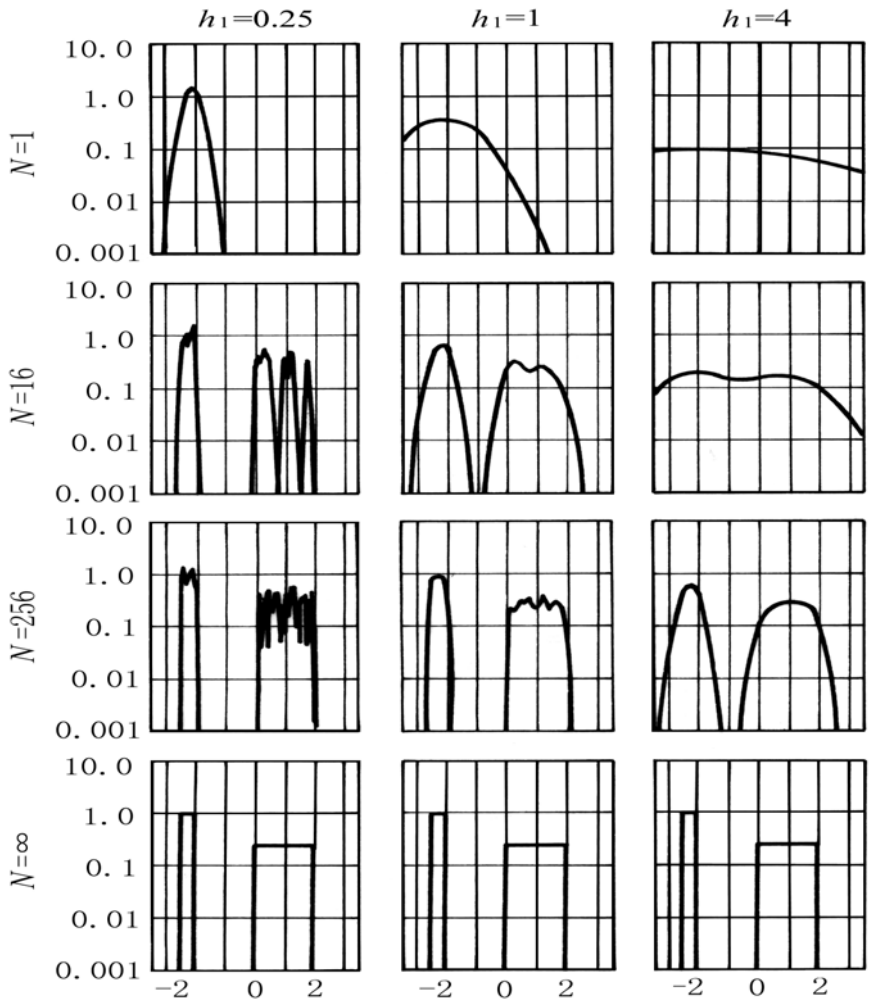


图 3.8 用 Parzen 窗法估计单一正态分布的实验，图左边给出了相应的样本数，图上边给出了窗宽值

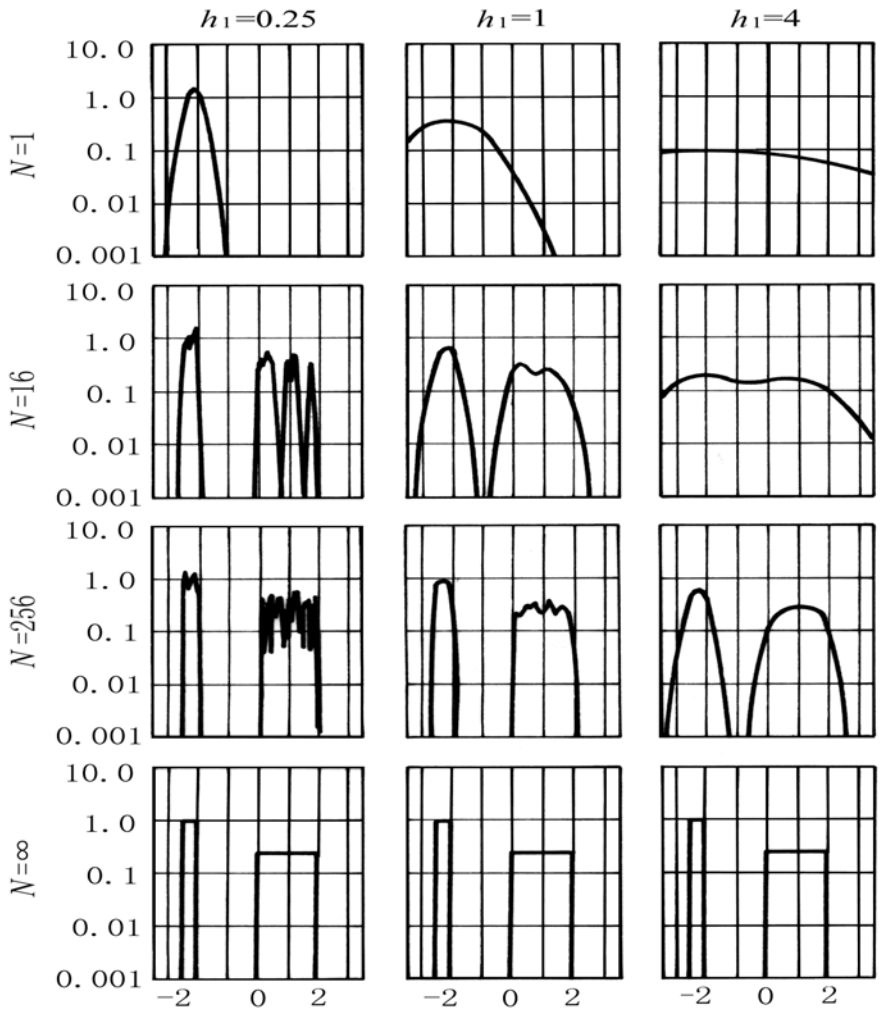


图 3.9 用 Parzen 窗法估计两个均匀分布的实验，图左边给出了相应的样本数，图上边给出了窗宽值

图 3.9 示出了对这个密度函数用 Parzen 窗估计的情况。当 $N=1$时，我们可以清晰地看到窗函数本身；当 $N=16$时，无法说明哪个估计更好；当 $N=256$ 及 $h_1=1$，估计结果与真实分布就较为接近了。

这些例子反映了非参数估计的一些性质及存在的问题。其优点是它的普遍性，即对规则分布或不规则分布，单峰或多峰分布都可以用这个方法得到密度估计。而且只要样本足够多，总可以保证收敛于任何复杂的未知密度。其问题是要想得到较为满意的结果，就需要远比参数估计方法所要求的样本数多得多的样本，因此就需要大量的计算时间和存储量。这个问题我们在本章的下一节会继续讨论。

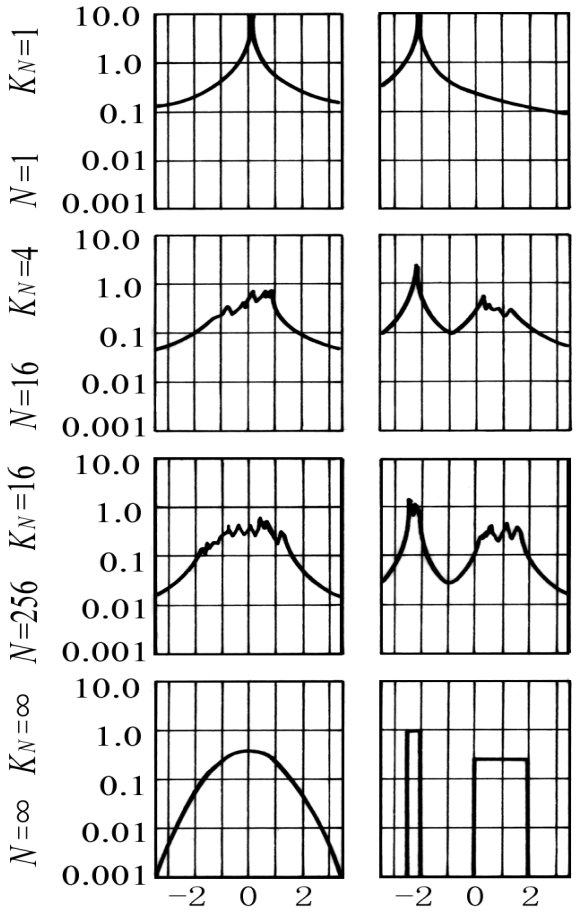
3.5.3 k_N-近邻估计

在 Parzen 窗估计中如何选择体积序列$V_1,V_2,\cdots,V_N,\cdots$是一个重要的问题。例如，当$V_N=V_1/\sqrt{N}$ 时，对任何有限的 N，得到的结果对初值V_1的选择很敏感。若V_1选得太小，则大部分体积将是空的，从而使 $\hat{p}_N(x)$ 估计不稳定；而若V_1选得太大，则 $\hat{p}_N(x)$ 估计较平坦，从而反映不出真实总体分布的变化。为解决这一问题就提出了 k_N 近邻估计法。

k_N 近邻法的基本思想是让体积成为数据的函数，而不是样本数 N 的函数。比如说为了从 N 个样本中估计 $p(x)$，我们可以预先确定 N 的某个函数 k_N，然后在 x 点周围选择一个体积，并让它不断增长直至捕获 k_N 个样本为止，这些样本为 x 的 k_N 个近邻。如果 x 点附近的密度比较高，则包含 k_N 个样本的体积自然就相对比较小，从而可以提高分辨力。如果 x 点附近的密度比较低，则体积就较大，但它一进入高密度区就会停止增大。 k_N 近邻估计仍用基本估计公式（3-79），假设条件也仍然是式(3-80)～式(3-82)

k_N 可以取为 N 的某个函数，如取 $k_N = k_1\sqrt{N}$ ， k_1 可以选择为某个大于零的常数，但至少要选取 k_1 使 $k_N \geq 1$。在样本数 N 有限时，与 Parzen 窗类似，k_1 的选择也会影响到估计的结果。但当 $N \to \infty$ 时，$\hat{p}_N(x)$ 将收敛于密度函数 $p(x)$，图 3.10 示出对正态分布及双峰密度函数的 k_N-近邻估计结果。

最后还应指出， k_N-近邻估计也存在一般非参数估计的问题，即所需样本数很多，因而计算量、存储量很大。尤其是在高维情况下更是如此。我们会在下一节讨论这个问题。



（教材 P72 的图 3.8）

图 3.10 用 k_N-近邻估计对单一正态分布和两个均匀分布估计的实验，图左边给出了相应的样本数和近邻数。

我们的实验表明，一维时用数百个样本一般可以得到较好结果，而两维估计就要数千个样本，随着维数增加，样本数将急剧增多。迄今已有很多人研究了不少解决办法，如用正交级数展开逼近上面的非参数估计等，其目的都在于减少计算量和存储量，这方面的文章已发表很多，有兴趣的读者可以参考有关资料。

3.6 概率密度函数估计中的一些问题

在这一小节我们讨论和概率密度估计相关的一些重要概念和重要问题。

3.6.1 估计的准确性与分类器性能的关系

在模式识别中，估计概率分布函数的目的是为了设计一个好的分类器。对概率分布函数的估计的准确性当然直接影响到分类器的性能。但是，影响分类器性能的不仅仅是这一个因素，一般来说，导致分类器产生误差的因素有如下几个：

贝叶斯误差 这种误差是由于不同的类条件概率分布函数之间的相互重叠引起的。这种分类误差是问题本身所有固有的，在分类器设计阶段是无法消除的。只有在特征的提取阶段才有可能减小或者消除这一误差。对于这个问题，在特征提取和选择部分还会讨论这个问题。

模型误差 这是由于选择了不正确的模型所导致的分类误差。例如，如果数据是从一个均匀分布中采样得到的，但是选择一个正态分布函数去拟合这些数据，得到的结果就会很差。如图 3.2(c)。在设计分类系统时，只有选择了正确的模型这一误差才可能消除。然而，系统的设计人员通常是根据对问题的先验知识和理解来选择模型，这和最大似然估计、贝叶斯估计这些具体的估计方法没有关系。由此出现这样一个问题：是否可以让算法自动选择合适的模型？这就是模型选择问题。需要说明的是，模型选择也是模式识别中一个重要的问题。关于这个问题有一些研究成果。模型的选择关系到设计人员对问题的先验知识和理解、模型的复杂程度和可计算性，样本的数量等问题。在本教材中在其他地方还会对于这个问题展开讨论。

估计误差 这是由于采用有限样本进行估计所带来的误差。例如，利用 5 个样本和 50 个样本估计一个正态分布，其误差会很不一样。如图 3.2(c)。一般来说，要减小估计误差需要增加训练样本数。当样本数无穷多的时候，这个误差会消除，这时最大似然估计和贝叶斯估计的结果是一样的。然而实际中的样本总是有限的，因此，这个问题会涉及到采用的估计方法，以及计算的复杂性和精度之间的折中等问题。

3.6.2 维数问题

我们从一个例子来考虑数据的维数可能带来的问题。

在上一小节中我们给出了用非参数方法估计概率密度函数的例子。看图 3.8，图 3.9 和图 3.10 可以知道，在估计一维概率密度函数时用数百个样本一般可以得到较好的结果，假设需要 100 个样本。这是因为在每一个点附近应该有一定量的样本（也就是说，数据要具有一定的密度）才能得到好的估计结果。简单说，我们可以把一维数轴等分成一些小区间，需要让每一个区间有足够多的样本来保证好的估计结果。这样一来，如果要估计的是一个二维概率密度函数，就需要把二维空间等分成一些小网格，需要每一个小网格中有足够多的样本。因此，可能需要$100^2 = 10000$个样本。当维数增加的时候，空间中小的格子的数量随维数的增加而指数上升。假设维数为d，需要的样本数是100^d。在很多实际的分类问题中，提取的特征个数要多于 10，而在一些文本分类，图像分类问题中，提取的特征数会超过 1000。因此，和特征数相比，我们所拥有的样本数太少。这个现象被称作“维数灾难”。维数灾难的核心问题是，高维函数实事上远比低维函数复杂，而我们对其还没有有效的方法进行分析。

利用我们对于具体分类问题的先验知识，或者利用得到的训练数据可能在一定程度上减少维数灾难带来

的问题。例如，如果我们知道两组特征 x 和 y 之间是独立的，那么就有 $p(x,y)=p(x)p(y)$。因此，对于 $p(x,y)$ 的估计就可以通过分别对 $p(x)$ 和 $p(y)$ 进行估计来完成。由于单独的 x 或 y 的维数要小于它们联合在一起的 (x,y) 的维数，因此，所需要的样本数就会少很多。

另外，对实际问题中的数据分析表明，大量的高维数据实际上嵌入在一个低维的流形上。也就是说，数据并没有充满整个高维空间。其主要原因就是各个特征之间存在很强的相关性。因此，我们实际上并不需要那么多的数据来估计概率密度函数。

上面讨论的都是非参数估计方法中遇到的维数灾难问题。当然，在参数估计中所需要的样本数会远远小于非参数方法所需要的样本数，但样本数仍然随维数的增加而指数上升。因此，在估计概率密度函数的不同方法中都存在维数灾难。实际上，维数灾难不仅仅出现在概率密度函数的估计中。上面谈到的特征独立和流形分布的知识同样有助于缓解在其他方面遇到的维数灾难问题。我们会在本教材的相关部分继续讨论维数灾难问题。

3.6.2 过拟和

我们首先考虑一个简单的回归问题。假设我们观测到一个实值输入变量 x，并希望使用该变量的观测数据来预测一个实值目标变量 y 的取值。为此，我们考虑一个人工合成的数据的例子，因为这样我们就知道产生这些数据的准确过程，并可以将其和其他模型进行比较。这个例子的数据由抛物线函数 $y=f(x)=ax^2+bx+c+\varepsilon$ 生成，其中，a，b，c 是常数，ε 是随机变量，服从一个正态分布 $N(0,\sigma^2)$。现在假设给定了一个训练集，其包含 x 的 10 个观测，以及相应的观测值 y。如图 3.11 所示。这样我们知道了该数据集合的真实特性，即它们的本质规律，这些规律是我们希望通过数据学习得到的。我们可以采用曲线拟合技术得到两条曲线。一条就是该抛物线函数，另一条是一条 10 阶的多项式曲线。我们看到，就这 10 个点来说，这条 10 阶的多项式曲线对这些点拟合得更好。但是这条 10 阶的多项式曲线与这条抛物线相差太远。除了这 10 个点外，如果再对实数轴上的其他点 x 处的 y 预测，这条 10 阶的多项式曲线会给出相当糟糕的结果。对于训练样本以外的其他样本的预测能力被称作做泛化能力，或推广能力。我们说，这条 10 阶多项式曲线出现了过拟合，因为它的泛化能力太差。由于我们对要拟合的模型的先验信息很少，因此，仅仅根据很少的样本点往往会出现过拟合。

根据一些数据点估计一个概率分布函数也可以看作是对真实分布函数的拟合过程。因此，当训练数据太少的时候，也会出现过拟合现象。图 3.12 给出的是一个例子。图 3.12 中的右图是一个真实的正态分布的等高线图，左图是根据在该分布上采样 5 个点后利用最大似然估计得到的正态分布的等高线图。可以知道，这个估计结果并不好。已经有分析告诉我们，当样本数目很小的时候，最大似然估计容易出现过拟合现象。

如果没有先验知识，并且也不能增加样本数量，那么要避免过拟合现象可以降低问题的维数，或采用更为简单的模型，或使用更少的参数这些方法。例如，要估计一个正态分布的参数，当样本数目很少时，可以假设协方差矩阵是一个对角矩阵。这样一来，原来需要估计协方差矩阵的 $d\times d$ 个参数，现在就降低为只估计 d 个参数了。这实际上假定了特征之间是彼此独立的。因而得到的分类器几乎肯定是次优的。因为只有这些特征真正独立时该分类器才是最优的。但是有意思的是，这样得到的结果有时候比估计 $d\times d$ 个参数的协方差矩阵得到的分类器的性能更好。出现这样的结果的一个原因可能是当样本很少时，估计 $d\times d$ 个参数的协方差矩阵会出现严重的过拟合。因此虽然消除了模型误差，但是严重的过拟合使得估计误差变得很大，导致了分类器准确性很低。当限制正态分布为对角阵时，避免了估计出过于“畸形”的正态分布。虽然出现了模型误差，但是由于减小了估计误差，因此最后的分类器性能反而提高了。

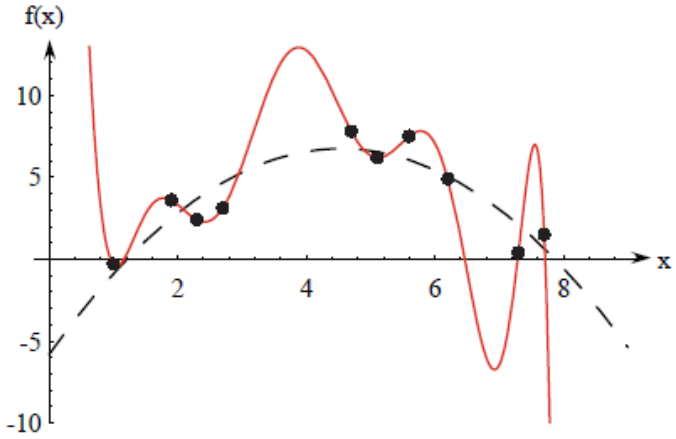


图 3.11 对 10 个数据点拟合的两条曲线

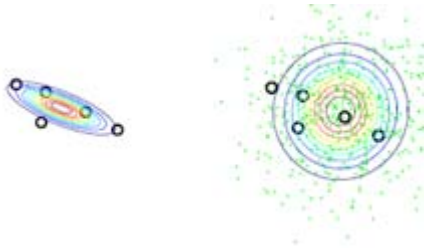


图 3.12 利用 5 个点估计一个正态分布

本小节讨论的问题虽然都是以概率密度函数估计为中心讨论的。但是在模式识别的其他一些内容中也会涉及这些问题。我们在教材的后面的部分还会讨论一些相关的问题。

3.7 文献与讨论

上一章已经指出，应用贝叶斯决策理论设计分类器时需要已知类概率密度函数。因此需要首先估计概率密度函数，然后再设计分类器。

概率密度函数的估计已经有了很长的历史，建立了一套较完整的理论和方法。最初把贝叶斯方法引入模式识别的领域的是文献[6]。贝叶斯方法强调了先验信息的作用。文献[15][21]对不同的先验概率有详细的分析。文献[4]中更多地列举了这方面的文献。

EM 算法是由 Dempster[11]等人提出的。文献[23]对这一方法及其发展历史进行了详细论述。文献[17, 31]描述了 EM 算法的在线学习（增量学习）方式。文献[27]专门讨论了丢失数据情况下的 EM 算法。EM 算法与模式识别、机器学习中使用的很多模型和算法都有很强的联系，关于这方面的内容请参考[34]。EM 算法也广泛用于图像、文本、语音等领域的信息处理问题。

文献[35]首次提出使用窗函数方法来估计概率密度函数。Specht[36]把这一方法用于解决模式分类问题。文献[37]则研究了如何对分布于流形上的数据采用窗函数方法。

已有的研究告诉我们，当样本数趋于无穷时，概率密度函数的估计方法具有很多好的性质。但在实际应用中，我们得到的样本数往往是比较有限的，有时候是非常少的。这样估计出的概率密度函数不能很好地反映真实情况，因此在此基础上设计出的分类器的性能也会比较差。实际上，当样本数有限时，概率密度函数估计问题是一个比分类器设计问题更难的一般性问题[38][39]。我们试图通过解决这个更难的一般问题来解决分类问题显然是不合理的。但是需要指出的是，概率密度函数的估计是一个基本问题，有广泛的用途，也在很多问题的研究中取得了好的结果。

[1] Pierre Baldi, SØren Brunak, Yves Chauvin, Jacob Engelbrecht, and Anders Krogh. Hidden Markov models for human genes. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Neural Information Processing Systems*, volume 6, pages 761-768, San Mateo, CA, 1994. Morgan Kaufmann.

[2] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554-1563,1966.

[3] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164-171, 1970.

[4] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley, New York, NY, 1996.

[5] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK, 1995.

[6] David Braverman. Learning filters for optimum pattern recognition. *IRE Transactions on Information Theory*, IT-8:280-285, 1962.

[7] Wray L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159-225, 1994.

[8] Wray L. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195-210, 1996.

[9] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, 1993.

[10] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[11] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1-38, 1977.

[12] G. David Forney, Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61:268-278, 1973.

[13] Peter E. Hart and Jamey Graham. Query-free information retrieval. *IEEE Expert:Intelligent Systems and Their Application*, 12(5):32-37, 1997.

[14] David Heckerman. *Probabilistic Similarity Networks*. ACM Doctoral Dissertation Award Series. MIT Press, Cambridge, MA, 1991.

[15] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 1961 reprint edition, 1939.

[16] Michael I. Jordan, editor. *Learning in Graphical Models*. Kluwer, Dortrecht, Netherlands, 1998.

[17] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181-214, 1994.

[18] Donald E. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley, Reading, MA, 1 edition, 1973.

[19] Gary E. Kopec and Phil A. Chou. Document image decoding using Markov source models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):602-617, 1994.

[20] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov models in computational biology: Applications to protein modelling. *Journal of Molecular Biology*, 235:1501-1531, 1994.

[21] Dennis Victor Lindley. The use of prior probability distributions in statistical inference and decision. In Jerzy Neyman and Elizabeth L. Scott, editors, *Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, 1961. U. California Press.

[22] Andrei Andreivich Markov. Issledovanie zamechatelnogo sluchaya zavisimykh ispytanii (investigation of a remarkable case of dependant trials). Izvestiya Petersburgskoi akademii nauk, 6th ser., 1(3):61-80, 1907.

[23] Geoffrey J. McLachlan and Thiriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley Interscience, New York, NY, 1996.

[24] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[25] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[26] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2):257-286, 1989.

[27] Donald B. Rubin and Roderick J. A. Little. *Statistical Analysis with Missing Data*. John Wiley, New York, NY, 1987.
[28] Jürgen Schürmann. *Pattern Classiffcation: A unified view of statistical and neural approaches*. John Wiley and Sons, New York, NY, 1996.
[29] Padhraic Smyth, David Heckerman, and Michael Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9:227-269, 1997.
[30] Charles W. Therrien. *Decision Estimation and Classiffcation: An Introduction to Pattern Recognition and Related Topics*. Wiley Interscience, New York, NY, 1989.
[31] D. Michael Titterington. Recursive parameter estimation using incomplete data. *Jounal of the Royal Statistical Society series B*, 46:257-267, 1984.
[32] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13:260-269, 1967.
[33] Sewal Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557-585, 1921.
[34] **http://www.vision.caltech.edu/welling/class/LearningSystemsB.html**
[35] Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065-1076, 1962.
[36] Donald F. Specht. Generation of polynomial discriminant functions for pattern recognition. *IEEE Transactions on Electronic Computers*, EC-16:308-319, 1967.
[37] http://www.vision.caltech.edu/welling/class/LearningSystems156B.html
[38] Vladimir N. Vapnik, The Nature of Statistical Learning Theory, New York: Springer-Verlag, 1995.
[39]Vladimir Vapnik, Estimation of Dependences Based on Empirical Data, Springer-Verlag NY,1982