第三章 概率密度函数的估计

3.1 引言

这一章中主要讲的是概率密度函数的估计方法，主要分类两大类分别是：参数估计和非参数估计。

参数估计是知道概率密度的分布形式，但其中的参数部分未知或全部未知时，用来估计未知部分参数的方法。例如我们获得了一个一维样本，只知其服从高斯分布 $N(\mu,\sigma^2)$，但不知道 μ 与 σ 的具体取值时，我们就可以用本章的知识通过样本去估计 μ 与 σ 的取值。本章介绍了最大似然估计、贝叶斯估计和 EM 估计方法。

参数估计是一个点估计问题，点估计问题就需要构造一个统计量 $d(x_1,\cdots,x_N)$ 作为参数 θ 的估计 $\hat{\theta}$。将样本 $x_1,\cdots,x_N$ 的具体数值带入统计量公式 d 中获得的 $\hat{\theta}$ 的具体数值就是参数 θ 的估计值。

非参数估计是既不知道分布形式，也不知道分布的参数，只能通过样本的分布把概率密度函数值数值化估计出来。本章介绍的是 Parzen 窗法和 k_N 近邻法。

3.2 最大似然估计

3.2.1 最大似然估计

最大似然估计（maximum likelihood estimation）的思想是：随机试验有若干个可能的结果，如果在一次试验中某一结果出现了，我们就认为这一结果出现的概率比较大，从而可以假设这一结果是所有可能出现结果中最大的一个。

在最大似然估计中，我们做如下假设：

（1）我们把要顾及的参数记做 θ，它是确定但是未知的量（多个参数时是向量）。

（2）一共有 c 类，每类的样本集记作 X_i，$i=1,2,\cdots\cdots,c$，其中的样本都是从密度为 $p(x|\omega_i)$ 的总体中独立抽取出来的，即所谓满足独立同分布条件。

（3）类条件概率密度 $p(x|\omega_i)$ 具有某种确定的函数表达式，只是其中的参数 θ 未知。为了强调概率密度中待估计的参数，也可以把 $p(x|\omega_i)$ 写作 $p(x|\omega_i,\theta_i)$ 或 $p(x|\theta_i)$ 。

（4）各类样本只包括本类的分部信息，也就是说，不同类别的参数是独立的，这样就可以分别对每一类单独处理。

在这些假设的前提下，我们就可以分别处理 c 个独立的问题，即，在一类中独立的按照概率密度 $p(x|\theta)$ 抽取样本集 X ，用 X 来估计位置参数 θ 。

因为样本集 X 是独立的从概率密度为 $p(x|\theta)$ 中随机取的，所以样本的联合概率密度可以写为：

$$L(\theta)=p(X|\theta)=p(x_1,x_2,\cdots,x_N|\theta)=\prod_{i=1}^{N}p(x_i|\theta) \tag{3-1}$$

这个概率反映了参数为 θ 时得到样本集 X 的概率，其中样本 $x_1,x_2,\cdots,x_N$ 是已知的，参数 θ 是未知的，式(3-1)就变成了参数 θ 的函数，因此这个函数叫做 θ 相对于 X 的似然函数（likelihood function）。其中 $p(x_i|\theta)$ 可以被看作参数 θ 相对于样本 x_i 的似然函数。

我们先假定参数 θ 已知，从一个分布函数和参数的都已知分布中抽取一个样本，例如从 $N(6,1)$ 中抽取一个样本，样本最有可能的值是 $x_1=6$ ，也仅仅在 $x_1=6$ 时似然函数 $L(6,1)=p(x|6,1)$ 取到最大值。我们再假定分布函数已知，参数 θ 未知，而我们通过抽样得到了 N 个样本 $x_1,x_2,\cdots,x_N$ ，想知道这个样本集来自于哪个密度函数（参数 θ 的取值）的可能性最大。

根据最大似然估计的思想，样本集 X 是所有可能出现的结果中最大的一个，即此时 $L(\theta)$ 取到最大值。我们将 $L(\theta)$ 取到最大值的 $\hat{\theta}=d(x_1,x_2,\cdots,x_N)$ 称为 θ 的最大似然估计量，记为 $\hat{\theta}=\arg\max L(\theta)$ 。通常为了便于分析，还定义了对数似然函数

$$H(\theta)=\ln L(\theta)=\ln\prod_{i=1}^{N}p(x_i|\theta)=\sum_{i=1}^{N}\ln p(x_i|\theta) \tag{3-2}$$

对似然函数取对数后不改变函数的单调性，所以使对数似然函数$H(\theta)$最大的值θ也使似然函数$L(\theta)$最大。

在式(3-2)中，函数形式$p(\bullet)$是已知的，样本x_i也是已知的，未知量仅有参数θ。通常情况下在似然函数满足连续、可微的条件下，如果θ是一维变量，其最大似然估计量就是如下微分方程的解：$\frac{dL(\theta)}{d\theta}=0$或$\frac{dH(\theta)}{d\theta}=0$。如果$\theta$是多个未知参数组成的向量时，例如有$s$个未知参数，求解似然函数的最大值就需要对$\theta$的每一维分别求偏导，获得$s$个方程并令其等于零，方程组的解就是$\theta$的最大似然估计量，若偏导方程组有多个解，其中使似然函数最大的那个解才是最大似然估计量。

并不是所有概率密度形式都可以用上述方法求得其最大似然估计，比如均匀分布：

$$p=\frac{1}{\theta_2-\theta_1}\qquad \theta_1<x<\theta_2 \tag{3-3}$$

其中分布的参数θ_1,θ_2未知。从总体分布中独立抽取了N个样本$x_1,x_2,\cdots,x_N$，则似然函数为：

$$L(\theta)=p(X\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)=\frac{1}{(\theta_2-\theta_1)^N} \tag{3-4}$$

对数似然函数为：

$$H(\theta)=-N\ln(\theta_2-\theta_1) \tag{3-5}$$

对似然函数求偏导得到：

$$\frac{\partial H}{\partial\theta_1}=N\frac{1}{(\theta_2-\theta_1)},\quad \frac{\partial H}{\partial\theta_2}=-N\frac{1}{(\theta_2-\theta_1)} \tag{3-6}$$

令偏导等于零，解得$\theta_2-\theta_1=\infty$，结果无意义。像这种似然函数在最大值的地方没有零斜率的情况，只能通过别的方法寻找最大值。由公式(3-4)可以看出θ_2与θ_1越接近的时候，似然函数的值就越大。而在有给定观测值的样本集中，θ_1不能小于最小的观测值，θ_2不能大于最大的观测值。因此θ的最大似然估计为$\theta_1=\min(x_1,x_2,\cdots,x_N)$，$\theta_2=\max(x_1,x_2,\cdots,x_N)$。

3.2.2 正态分布下的最大似然估计

首先我们来看在正态分布下，仅有一个参数未知的情况，假设参数μ未知，对于单变量（样本特征仅有一维）的正态分布来说，其分布密度函数为

$$p(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] \tag{3-7}$$

在这样的条件下，我们假设一个样本点x_k，有下面的式子成立

$$\ln p(x_k|\mu)=-\frac{1}{2}\ln[2\pi\sigma^2]-\frac{1}{2\sigma^2}(x_k-\mu)^2 \tag{3-8}$$

对上述对数似然函数进行求导得到

$$\frac{d\ln p(x_k|\mu)}{d\mu}=\frac{(x_k-\mu)}{\sigma^2} \tag{3-9}$$

对于 N 个样本点的样本集来说，对μ的似然估计值$\hat{\mu}$的最大似然估计必须满足

$$\sum_{k=1}^{N}\frac{(x_k-\hat{\mu})}{\sigma^2}=0 \tag{3-10}$$

整理可得到

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{N}x_k \tag{3-11}$$

$\hat{\mu}$就是μ的最大似然估计值。

我们再来考虑两个参数都未知的情况：对于单变量（样本特征仅有一维）的正态分布来说，其分布密度函数如式(3-7)其中的均值μ和方差σ^2都是未知参数。求解多参数似然函数最大值时需要对每个参数都求偏导。

本文中使用对数似然函数求导可得：

$$\frac{\partial H(\theta)}{\partial\mu}=\sum_{k=1}^{N}\frac{1}{\sigma^2}(x_k-\mu),\quad \frac{\partial H(\theta)}{\partial\sigma^2}=-\sum_{k=1}^{N}\frac{1}{\sigma^2}+\sum_{k=1}^{N}\frac{(x_k-\mu)^2}{(\sigma^2)^2} \tag{3-12}$$

令上面两个偏导等于零可以解得

$$\hat{\mu}=\frac{1}{N}\sum_{k=1}^{N}x_k,\quad \hat{\sigma}^2=\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu})^2 \tag{3-13}$$

同理可得，对于多元正态分布的均值和方差为

$$\hat{\boldsymbol{\mu}}=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k\ ,\quad \hat{\boldsymbol{\Sigma}}^2=\frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}})^T \tag{3-14}$$

3.3 贝叶斯估计与贝叶斯学习

3.3.1 贝叶斯估计

贝叶斯估计（Bayesian Estimation）是概率密度估计中另一种主要的参数估计方法，其结果在很多情况下与最大似然法相同或几乎相同，但是两种方法对问题的处理视角是不同的，在应用上也各有各自的特点。与最大似然估计根本的区别是，似然估计是把参数当作未知但固定的量，要做的是根据观测数据估计这个量的取值；而贝叶斯估计是把未知参数看作一个随机的变量，根据观测数据和参数的先验分布来估计参数的分布。

在用于分类的贝叶斯决策中，最优的条件是最小错误率或者最小风险，在贝叶斯估计中，我们假定把连续变量θ估计成$\hat{\theta}$的损失为$\lambda(\hat{\theta},\theta)$，也称作损失函数。

设样本取值空间为E^d，参数θ的取值空间是Θ，那么，当用$\hat{\theta}$来作为估计时总期望风险就是

$$\begin{aligned}R&=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(x,\theta)\mathrm{d}\theta\mathrm{d}x\\&=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,x)\,p(x)\mathrm{d}\theta\mathrm{d}x\end{aligned} \tag{3-15}$$

我们定义在样本x下的条件风险为：

$$R(\hat{\theta}\,|\,x)=\int_{\Theta}\lambda(\hat{\theta},\theta)\,p(\theta\,|\,x)\mathrm{d}\theta \tag{3-16}$$

那么式(3-15)可以写为

$$R=\int_{E^d}R(\hat{\theta}\,|\,x)\,p(x)\mathrm{d}x \tag{3-17}$$

现在的目标是对期望风险求最小。与贝叶斯分类决策时相似，这里的期望风险也是在所有可能的x情况下的条件风险的积分，而条件风险又都是非负的，所以求期望风险最小就等价与对所有的样本求条件风险最小，即

$$\theta^*=\arg\min R(\hat{\theta}\,|\,x)=\int_{\Theta}\lambda(\hat{\theta},\theta)\,p(\theta\,|\,X)\mathrm{d}\theta \tag{3-18}$$

通常情况下我们使用的损失函数为平方误差损失函数 $\lambda(\hat{\theta},\theta)=(\hat{\theta}-\theta)^2$，在平方误差损失函数与样本集 X 下，θ 的贝叶斯估计量 θ^* 为 θ 在 X 下的条件期望，即：

$$\theta^* = E[\theta \mid X] = \int_{\Theta} \theta p(\theta \mid X) d\theta \tag{3-19}$$

在最小平方误差损失函数下，贝叶斯估计的步骤是：

1.根据对问题的认识，或者猜测确定 θ 的先验分布 $p(\theta)$。

2.由于样本是独立同分布，而且已知样本密度函数的形式 $p(x \mid \theta)$，可以求出样本集的联合分布为 $p(X \mid \theta)=\prod_{i=1}^{N} p(x_i \mid \theta)$，其中 θ 为变量。

3.利用贝叶斯公式 $p(\theta \mid x)=\frac{p(x \mid \theta)p(\theta)}{p(x)}$，求 θ 的后验概率分布：

$$p(\theta \mid X)=\frac{p(X \mid \theta)p(\theta)}{\int_{\Theta} p(X \mid \theta)p(\theta)d\theta}$$

4.根据结论(3-19)，θ 的贝叶斯估计量是 $\theta^* = \int_{\Theta} \theta p(\theta \mid X) d\theta$。

3.3.2 正态分布下的贝叶斯估计

我们以一维正态分布模型为例来说明贝叶斯估计的应用。假设 σ^2 已知且均值 μ 的先验分布为正态分布 $N(\mu_0,\sigma_0^2)$。x 的分布密度可以写为

$$p(x \mid \mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3-20}$$

μ 的分布密度为

$$p(\mu)=\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}) \tag{3-21}$$

求得 μ 的后验概率分布为：

$$p(\mu \mid X)=\frac{p(X \mid \mu)p(\mu)}{\int_{\Theta} p(X \mid \mu)p(\mu)d\mu} \tag{3-22}$$

上式的分母部分为归一化的常数项，将 $p(x \mid \mu)=\prod_{n=1}^{N} p(x_i \mid \mu)$，带入分子，

可得

$$p(x|\mu)p(\mu)=\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})\prod_{i=1}^{N}p(x|\mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

将与μ无关的分量写成一个常数项，上式可整理为另一个正态分布$N(\mu_N,\sigma_N^2)$。其中$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}$，$\mu_N=\sigma_N^2(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2})$。

所以，用式(3-19)可以得到参数μ的估计值，即

$$\hat{\mu}=\int\mu p(\mu|X)d\mu=\int\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(\frac{\mu-\mu_N}{\sigma_N})^2)d\mu=\mu_N \tag{3-23}$$

整理μ_N可以得到

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 \tag{3-24}$$

其中$m_N=\frac{1}{N}\sum_{i=1}^{N}x_i$为所有样本的参数平均。

3.3.3 贝叶斯学习

贝叶斯学习和贝叶斯估计的前提条件是相同的，详细如下：

（1）已知各个类型的训练样本子集$X=\{x_1,x_2,\cdots,x_N\}$，每次训练实验都是独立进行的，类型ω_i的参数与类型ω_j的样本无关。因此，训练过程可以逐个类型的进行，而不用保留类型标记。

（2）已知类概率分布密度函数$p(x|\theta)$，但是参数向量θ未知（θ是属于某一类型的）。

（3）关于未知参数θ的一般性信息包含在他的先验分布密度$p(\theta)$中。

（4）关于未知参数θ的其余信息要从训练样本集X中提取。

贝叶斯学习和贝叶斯估计有着密切的联系，但是贝叶斯学习最关心的并不是某个具体参数的估计，而是获得后验分布密度$p(x|X)$。具体的讲，在贝叶斯估计的四个步骤中，贝叶斯学习要执行前三个步骤，得到未知参数的后验分布

$p(\theta|x)$之后，不必真正的求出$\hat{\theta}$，而是直接求后验分布密度$p(x|X)$。

下面就是要研究如何得到$p(x|X)$。为此我们利用联合概率分布密度在参数空间Θ里积分，即

$$p(x|X)=\int_{\Theta}p(x,\theta|X)d\theta=\int_{\Theta}p(x|\theta,X)p(\theta|X)d\theta \tag{3-25}$$

因为在θ被确定后，x仅与θ有关，所以上式变为

$$p(x|X)=\int_{\Theta}p(x|X)p(\theta|X)d\theta \tag{3-26}$$

不妨再列出后验分布密度$p(\theta|X)$的计算公式，它由贝叶斯公式得到，即

$$p(\theta|X)=\frac{p(X|\theta)p(\theta)}{\int_{\Theta}p(X|\theta)p(\theta)d\theta} \tag{3-27}$$

根据独立实验的假设，可有

$$p(X|\theta)=\prod_{k=1}^{N}p(x_k|\theta) \tag{3-28}$$

而$p(x|\theta)$的函数形式是给定的。所以，一般来说，我们可以通过计算上面三个式子来确定后验分布密度$p(x|X)$。

现在做两点深入的分析。

第一，假设已经得到未知参数θ的估计$\hat{\theta}$，因此也就确定了后验分布密度$p(x|\hat{\theta})$。那么，如果从所研究的类型中任取一个样本x，它最有可能在$\hat{\theta}$处出现，采集的样本越多，在$\hat{\theta}$处出现的概率也就越大。当$\hat{\theta}$为均值向量时，如果在该类型中任取的样本足够多，$p(X|\theta)$就会在$\hat{\theta}$处出现一个尖峰；在这种情况下，如果先验密度$p(\theta)$在$\hat{\theta}$处不为零且比较平坦由式(3-25)可以看出，后验分布密度$p(\theta|X)\Rightarrow p(X|\theta)$，即$p(X|\theta)$也会在$\hat{\theta}$处出现尖峰，当$N$趋于无穷大时，$p(\theta|X)$在$\theta$处逼近$\delta$函数，代入(3-24)，得到

$$p(x|X)\approx p(x|\hat{\theta}) \tag{3-29}$$

上式表明，在上述条件下，后验分布$p(x|\hat{\theta})$可以近似的作为真实概率分布。

第二，我们研究在训练样本数目 N 趋于无穷大时，后验分布密度 $p(x|X)$ 是否收敛于真实分布密度 $p(x)$。请注意，这里同样省略了类型的限制条件。

我们把 N 个训练样本组合的训练样本子集记为 $X^N=\{x_1,x_2,\cdots,x_N\}$，当 $N>1$ 时，有

$$p(X^N|\theta)=p(x_N|\theta)p(X^{N-1}|\theta) \tag{3-30}$$

根据贝叶斯公式

$$p(\theta|X^N)=\frac{p(x_N|\theta)p(\theta|X^{N-1})}{\int_\Theta p(x_N|\theta)p(\theta|X^{N-1})d\theta} \tag{3-31}$$

令先验分布 $p(\theta)=p(\theta|X^0)$ 为无样本条件下的后验分布密度，重复使用式(3-30)，就得到一个密度函数序列 $p(\theta)$、$p(\theta|x_1)$、$p(\theta|x_1,x_2)$、…这称为参数估计的递推贝叶斯方法。如果这个密度序列收敛于以真实的均值参数 θ_t 为中心的 δ 函数 $\delta(\theta-\theta_t)$，就把具有这种性质的递推过程乘坐贝叶斯学习。正态分布具有这种性质，而对于大多数典型的概率分布，也都具有这种性质。

如果给定的概率分布密度 $p(x|\theta)$ 具有贝叶斯学习的性质，当训练样本数目 N 趋于无穷大时，显然，式(3-29)描述的近似式就变成确切的等式了。并且此时的估计 $\hat{\theta}$ 就是真实参数 θ，而后验分布就是真实分布，即

$$p(x|X^{N\to\infty})=p(x|_{\hat{\theta}\to\theta})=p(x) \tag{3-32}$$

3.4 EM 估计方法

3.4.1 EM 算法

期望最大化（Expectation Maximization，EM）算法是当数据存在缺失时，极大似然估计的一种常用迭代算法，因为其操作简便，收敛稳定，具有很强的适用性。它主要应用于以下常见的两种情况下的参数估计：1）观测到的数据不完整，这是因为数据丢失或者观测条件受限；2）似然函数不是显然的，或者函数

的形式非常复杂导致难以用极大似然传统方法进行估计。

记 $\boldsymbol{Z}=\{\boldsymbol{X},\boldsymbol{Y}\}$ 为完全数据，由于数据缺失，其中包括观测到的数据和未观测到的潜在数据。已知 $\boldsymbol{X}=\{x_1,x_2,\cdots x_n\}$ 为观测数据，$\boldsymbol{Y}$ 为未观测到的潜在数据，θ 为参数。EM 算法的目标是关于 θ 最大化似然函数 $L(\theta|\boldsymbol{X})$。设 $\theta^{(k)}$ 表示在第 k 次迭代时估计得到的最大值点，定义 $Q(\theta|\theta^{(k)})$ 为观测数据 $\boldsymbol{X}=\{x_1,x_2,\cdots x_n\}$ 条件下完全数据的联合对数似然函数的期望，即：

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E\{\log L(\theta|\boldsymbol{Z})|x,\theta^{(k)}\} \\ &= E\{\log p(\mathrm{z}|\theta)|x,\theta^{(k)}\} \\ &= \int[\log p(\mathrm{z}|\theta)]p(\mathrm{y}|x,\theta^{(k)})dy \end{aligned} \tag{3-33}$$

EM 算法从 $\theta^{(0)}$ 开始，是一个寻找参数的最大似然解的两阶段迭代优化技术，第一段求期望（E 步）和第二段最大化（M 步），该算法步骤可概括如下：

E 步：在给定的观测数据 $\boldsymbol{X}=\{x_1,x_2,\cdots x_n\}$ 和已经知道的参数 $\theta^{(k)}$ 的条件下，求“缺失数据 $\boldsymbol{Y}$”的条件期望，即计算上面提到的对数似然函数的条件期望 $Q(\theta|\theta^{(k)})$。

M 步：针对完全数据下的对数似然函数的期望进行极大化估计，即求关于 θ 的似然函数 $Q(\theta|\theta^{(k)})$ 的最大化，更新 $\theta^{(k)}$：

$$\theta^{(k+1)}=\max_{\theta} Q(\theta|\theta^{(k)},\boldsymbol{X}) \tag{3-34}$$

以上 E 步和 M 步即为一次完整的迭代过程，之后返回 E 步继续迭代，直到满足停止条件。

为详细的说明 EM 算法的理论以及计算方法，来看一个例子：

例 1（基因环模型） 假设进行一个试验中会出现四种结果，每种结果发生的概率分别为 $\frac{1}{2}+\frac{\theta}{4}$，$\frac{1}{4}(1-\theta)$，$\frac{1}{4}(1-\theta)$，$\frac{\theta}{4}$，其中 $\theta\in(0,1)$，试验进行了 197 次，四种结果分别发生了 125，18，20，34 次，即此时得到观察数据 $X=(x_1,x_2,x_3,x_4)=(125,18,20,34)$。

为了估计参数，我们可以取 θ 的先验分布 $p(\theta)$ 为 $U(0,1)$，由贝叶斯公式可知，θ 的后验分布为：

$$
\begin{aligned}
p(\theta \mid X) &= p(\theta)\, p(X \mid \theta) \\
&= (\frac{1}{2}+\frac{\theta}{4})^{x_1}[\frac{1}{4}(1-\theta)]^{x_2}[\frac{1}{4}(1-\theta)]^{x_3}(\frac{\theta}{4})^{x_4} \\
&\propto (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}
\end{aligned}
\tag{3-35}
$$

把第一种结果分成发生概率分别为$\frac{1}{2}$和$\frac{\theta}{4}$的两部分，令Y和x_1 - Y分别表示这两部分实验成功的次数(Y为缺失数据)。则θ的后验分布为：

$$
\begin{aligned}
p(\theta \mid X,Y) &= p(\theta)\, p(X,Y \mid \theta) \\
&= (\frac{1}{2})^{Y}(\frac{\theta}{4})^{x_1-Y}[\frac{1}{4}(1-\theta)]^{x_2}[\frac{1}{4}(1-\theta)]^{x_3}(\frac{\theta}{4})^{x_4} \\
&\propto \theta^{x_1-Y+x_4}(1-\theta)^{x_2+x_3}
\end{aligned}
\tag{3-36}
$$

直接用式(3-23)求θ的极大似然估计是比较麻烦的，所以考虑用 EM 算法添加数据，迭代得到(3-24)式的后验分布函数要简单得多。在上面计算过程中，$\propto$表示符号两端的式子成比例，而且比例与θ无关，这个比例不会影响到 EM 迭代算法的估算结果，因为它在后面的极大化过程中可以约去。

假设在第 i+1 次迭代中，有估计值$\theta^{(i)}$，则可通过 EM 算法的 E 步和 M 步得到一个新的估计。在 E 步中，由(3-23)式得到：

$$
\begin{aligned}
Q(\theta \mid \theta^{(i)}) &= E[(x_1 - Y + x_4)\log(\theta) + (x_2+x_3)\log(1-\theta) \mid X, \theta^{(i)}] \\
&= [x_1 - E(Y \mid x, \theta^{(i)}) + x_4]\log(\theta) + (x_2+x_3)\log(1-\theta)
\end{aligned}
$$

在x和$\theta^{(i)}$给定的情况下，Y服从二项分布，即$Y \mid x, \theta^{(i)} \sim b(x_1, p_i)$，其中

$$
p_i = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta^{(i)}}{4}} = \frac{2}{2+\theta^{(i)}}
$$

因此，$E(Y \mid x, \theta^{(i)}) = \frac{2x_1}{2+\theta^{(i)}}$便有：

$$
Q(\theta \mid \theta^{(i)}) = [x_1 - \frac{2x_1}{2+\theta^{(i)}} + x_4]\log(\theta) + (x_2+x_3)\log(1-\theta)
$$

在 M 步中，对上式求导并令其为零，可以得到迭代公式如下：

$$
\theta^{(i+1)} = \frac{159\theta^{(i)} + 68}{197\theta^{(i)} + 144}
\tag{3-37}
$$

从$\theta^{(0)} = 0.5$开始，经过计算 EM 算法经过四次迭代收敛 0.6268。

3.4.2 混合正态分布的 EM 估计

混合正态分布（Gaussian Mixture Distribution）是指 $p(x)=\sum_{k=1}^{K}\pi_k N(x\mid\mu_k,\Sigma_k)$，其中 K 可以理解为是这个混合正态分布中正态分布的个数。如果直接对其对数似然函数求导来寻求极值是不可行的。但是如果我们知道每一个观测值具体是来自哪一个正态分布，则问题的难度就会下降很多。因此，从这个想法出发，我们引进隐含变量 Y，其分布为 $p(y)=\prod_{k=1}^{K}\pi_k^{y_k}$。其中 $\sum_{k=1}^{K}\pi_k=1$，$0\le\pi_k\le 1$。

假设有条件分布：$p(x\mid y_k=1)=N(x\mid\mu_k,\Sigma_k)$。也就是说 x 关于 y 的条件分布为

$$p(x\mid y)=\prod_{k=1}^{K}N(x\mid\mu_k,\Sigma_k)^{y_k} \tag{3-38}$$

条件概率为 $P(y_k=1)=\pi_k$ 可以理解为第 k 个正态分布占总体的大小为 π_k。

对似然函数

$$\ln p(X\mid\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(x_n\mid\mu_k,\Sigma_k)\} \tag{3-39}$$

关于 μ_k 求偏导并令其等于零，可以得到

$$0=-\sum_{n=1}^{N}\underbrace{\frac{\pi_k N(\mathrm{x}_n\mid\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_j N(\mathrm{x}_n\mid\mu_j,\Sigma_j)}}_{\gamma_{y_{nk}}}\Sigma_k(\mathrm{x}_n-\mu_k) \tag{3-40}$$

解得 $\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(y_{nk})\mathrm{x}_n$ 其中 $N_k=\sum_{n=1}^{N}\gamma(y_{nk})$ 可以理解为分配到的第 k 个分布中的有效点的个数。

同理，对似然函数求关于 Σ_k 偏导并令其等于零可得

$$\Sigma_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(y_{nk})(\mathrm{x}_n-\mu_k)(\mathrm{x}_n-\mu_k)^T \tag{3-41}$$

最后，在求π_k的最大似然估计时需要考虑到$\sum_{k=1}^{K}\pi_k=1$的约束。似然函数变为

$$\ln p(X\mid\pi,\mu,\Sigma)+\lambda(\sum_{k=1}^{K}\pi_k-1) \tag{3-42}$$

对其求π_k的偏导得到：

$$0=\sum_{n=1}^{N}\frac{N(\mathrm{x}_n\mid\mu_k,\Sigma_k)}{\sum_j\pi_j N(\mathrm{x}_n\mid\mu_j,\Sigma_j)}+\lambda \tag{3-43}$$

利用隐函数的条件概率可以得到$\lambda=-N$。化简得$\pi_k=\frac{N_k}{N}$。

混合正态分布模型下的 EM 算法：

1.初始化均值μ_k，协方差Σ_k，以及混合系数π_k，并估计初始对数似然函数值。

2.E 步。计算在这组参数下 Y 的后验概率下y_{nk}的期望$\gamma(y_{nk})$。

3.M 步。使用y_{nk}的期望$\gamma(y_{nk})$重新估计参数的最大值

$$\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(y_{nk})\,\mathrm{x}_n$$

$$\Sigma_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(y_{nk})(\mathrm{x}_n-\mu_k)(\mathrm{x}_n-\mu_k)^T \tag{3-44}$$

$$\pi_k^{new}=\frac{N_k}{N}\text{ 这里 }N_k=\sum_{n=1}^{N}\gamma(y_{nk})$$

4.估计对数似然函数$\ln p(X\mid\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(x_n\mid\mu_k,\Sigma_k)\}$，并检测是否达到收敛条件，若没有达到则继续执行第二步。

3.5 非参数估计方法

前面的三种方法都是参数化的估计方法，要求分布类型已知，只是利用观测到的样本数据来估计分布函数的参数。但在很多情况下我们无法知道密度函数的形式，而且样本分布也很难用简单的函数来描述。在这种情况下就需要用到非参

数估计，即不对概率密度做任何假设而是用样本估计出整个函数。当然这种估计只能用数值方法取得，无法得到完美的封闭函数形式。从另一角度看，概率密度函数的参数估计都是在指定函数形式中对函数的估计，而非参数估计郑则仕从所有可能的函数中进行一种选择。

直方图（histogram）是最简单也是最直观的非参数估计方法。

直方图估计的做法是

（1）将样本 x 在其取值范围内分成 k 个等间隔的小窗

（2）并统计落入每个小窗中样本的数目 q_i

（3）每个小窗的概率密度为 $\frac{q_i}{NL}$，其中 N 为样本总数 L 为小窗长度（ x 为一维的情况）。若 x 为 d 维向量则分成 k^d 个小舱，每个小舱的概率密度为 $\frac{q_i}{NV}$ 其中 V 是小舱的体积。

所以非参数估计是在已知样本集 $X=\{x_1,x_2,\cdots,x_N\}$ 中，样本是服从 $p(x)$ 的总体中独立抽取出来的，求 $p(x)$ 的估计 $\hat{p}(x)$。

考虑在样本所在空间的某个小区域 R，某个随机向量落入这个小区域的概率为：

$$P_R=\int_R p(x)dx \tag{3-45}$$

根据二项分布，在样本集 X 中，恰好有 k 个落入小区域 R 的概率是

$$P_k=C_N^k P_R^k(1-P_R)^{N-k} \tag{3-46}$$

其中 C_N^k 表示在 N 个样本中取 k 个的组合数，k 的期望值是

$$E[k]=NP_R \tag{3-47}$$

而且 k 的众数是

$$m=[(N+1)P_R] \tag{3-48}$$

其中[]表示向下取整。因此当小区域实际落入 k 个样本时，P_k 的一个很好的估计是

$$\hat{P}_R=\frac{k}{N} \tag{3-49}$$

当 $p(x)$ 连续、且小区域 R 的体积 V 足够小时，可以假定在该小区域范围内

$p(x)$ 是个常数，则式(3-45)可近似为

$$P_R = \int_R p(x)dx = p(x)V \tag{3-50}$$

用式(3-49)的估计带入(3-50)中，可得，在小区域 R 的范围内

$$\hat{p}(x) = \frac{k}{NV} \tag{3-51}$$

这就是在上面直方图中使用的对小舱内概率密度的估计。

3.5.1 Parzen 窗法

我们也可以这么理解直方图的概率密度公式。

假设 x 是一个 d 维的向量，并假设每个小舱是一个超立方体，它的每一维棱长为 h 。

定义如下 d 维单位方窗函数

$$\varphi([u_1,u_2,\cdots,u_d]) = \begin{cases} 1 & 若 \left|u_j\right| \le \frac{1}{2}, j = 1,2,...,d \\ 0 & 其他 \end{cases} \tag{3-52}$$

该函数在超正方体内取值为 1，其他地方取值为 0。对于每个样本 x_i 要考察其是否在以 x 为中心的小舱中就可以通过计算 $\varphi(\frac{x-x_i}{h})$ 来进行。任意一点落入以 x 为中心的小舱的个数为 $k = \sum_{i=1}^{N} \varphi(\frac{x-x_i}{h})$ 。则对任意一点 x 的密度估计表达式为

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V}\varphi(\frac{x-x_i}{h}) \tag{3-53}$$

其中 V 是小舱的体积。

从另一角度理解上式：定义核函数（也叫窗函数）

$$K(x,x_i) = \frac{1}{V}\varphi(\frac{x-x_i}{h}) \tag{3-54}$$

它反映了一个观测样本 x_i 对在 x 处的概率密度估计的贡献。概率密度估计就是在每一点上把所有观测样本的贡献的平均

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}K(x,x_i) \tag{3-55}$$

这种用核函数估计概率密度的方法叫做 Parzen 窗法。

例：选择正态窗函数

$$\varphi(\mu)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\mu^2} \tag{3-56}$$

由式(3-53)，对总体概率密度估计为：

$$\hat{p}(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi(\frac{x-x_i}{h_N}) \tag{3-57}$$

式中，N 为训练样本数目，$h_N=h_1/\sqrt{N}$，h_1 为可调节的参量。我们将研究 h_1 对估计结果的影响。

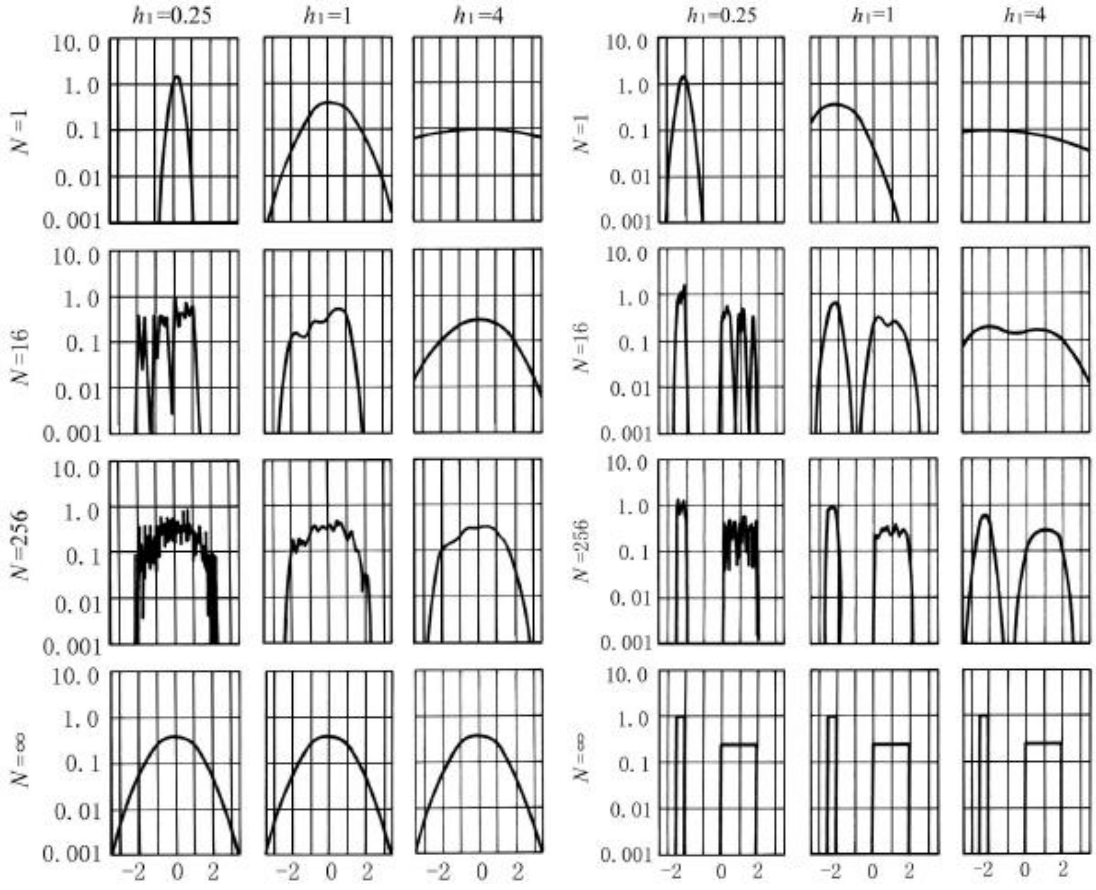

图 3-1　不同样本数和不同参数下 Parzen 窗估计的效果实例

在此例中，使用正态分布的随机样本，下图描绘了在不同 h_1 和 N 状态下的估计结果，当 $N=1$ 时，$\hat{p}(x)$ 是以第一个样本为中心的小丘，若选 $h_1=0.25$，小丘比较陡峭，若选 $h_1=4$，小丘比较平坦；当 $N=16$ 时，对 $h_1=0.25$，$\hat{p}(x)$ 仍然

清楚的体现各个样本的作用，而对 $h_1=1$， $h_1=4$ 各个样本的作用就变得模糊了；随着 N 的增加 $\hat{p}(x)$ 的曲线会变得越来越平滑， h_1 的影响变得越来越小，结果越来越真实。在样本数目不太多的情况下，例如 $N=256$，还会在 $\hat{p}(x)$ 中出现一些不规则的扰动，特别是在 h_1 较小时，例如对 $h_1=0.25$ 的曲线扰动就比较大。但是当训练样本数目趋于无限多时，不管 h_1 取多少， $\hat{p}(x)$ 都会收敛于平滑的正态分布曲线。

Parzen 窗方法的好处是不需事先知道概率密度函数的参数形式，比较通用，可以应对不同的概率密度分布形式。在处理有监督学习过程的时候，现实世界的情况往往是我们不知道概率密度函数形式。就算能够给出概率密度函数的形式，这些经典的函数也很少与实际情况符合。所有经典的概率密度函数的形式都是单模的（只有单个局部极大值），而实际情况往往是多模的。非参数方法正好能够解决这个问题，所以从这个意义上来讲，Parzen 窗方法能够更好地对应现实世界的概率密度函数，而不必实现假设概率密度函数的形式是已知的。Parzen 窗方法能处理任意的概率分布，不必假设概率密度函数的形式已知，这是非参数化方法的基本优点。

Parzen 窗方法的一个缺点是它需要大量的样本。在样本有限的情况下，很难预测它的收敛性效果如何。为了得到较精确的结果，实际需要的训练样本的个数是非常惊人的。这时要求的训练样本的个数比在知道分布的参数形式下进行估计所需要的训练样本的个数要多得多。而且，直到今天人们还没有找到能够有效的降低训练样本个数的方法。这也导致了 Parzen 窗方法对时间和存储空间的消耗特别大。更糟糕的是，它对训练样本个数的需求，相对特征空间的维数呈指数增长。这种现象被称为“维数灾难（curse of dimensionality）”，严重制约了这种方法的实际应用。Parzen 窗方法的另外一个缺点是，它在估计边界区域的时候会出现边界效应。

Parzen 窗方法的一个问题是，窗宽度的选择难以把握。下图是一个二维 Parzen 窗的两类分类器的判决边界。其中窗宽度 h 不相同。左边的图中的窗宽度 h 较小，右边的图中的窗宽度 h 较大。所以左侧的 Parzen 窗分类器的分类界面比右边复杂。这里给出的训练样本的特点是，上半部分适合用较小的窗宽度 h，而

下半部分适合用较大的窗宽度 h。所以，这个例子说明没有一个理想的固定的 h 值能够适应全部区域的情况。这算是 Parzen 窗方法的一个不足之处。

3.5.2 Kn 近邻估计方法

在 Parzen 窗算法中，我们固定了窗口的大小，即把体积V_N作为N的函数，例如$V_N = V_1 / \sqrt{N}$，导致了V_1的选择对估计结果的影响很大。在K_N近邻中，我们采用可变大小的小舱的密度估计方法，即选择K_N是N的函数，例如$K_n = K_1\sqrt{N}$，基本做法是：根据总样本，确定一个参数K_N，即在总样本为 N 时我们要求每个小舱内拥有同样的样本个数。在求x处的密度估计$\hat{p}(x)$时，我们调整包含x的小舱的体积，直到小舱内恰好有K_N个样本，并用下式来估算$\hat{p}(x) = \frac{k_N / N}{V}$。

这样在样本密度比较高的区域小舱的体积会比较小，在样本密度比较低的区域小舱体积会变大。这样就比较好的兼顾在高密度区域估计的分辨率和低密度区域估计的连通性。

K_N近邻估计与简单的直方图方法相比还有一个不同，就是K_N近邻估计不是把x的取值范围划分成若干区域，而是在x的取值范围内，以每一小点为小舱中心，用$\hat{p}(x) = \frac{k_N / N}{V}$来估算，如图 3-4 所示。图 3-5 给出了两个一维情况下在不同样本数目时k_N邻近估计效果的例子。

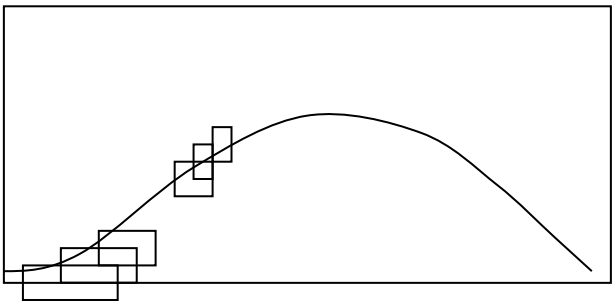

图 3-2　k_N 法的窗口宽度与样本密度的关系示意

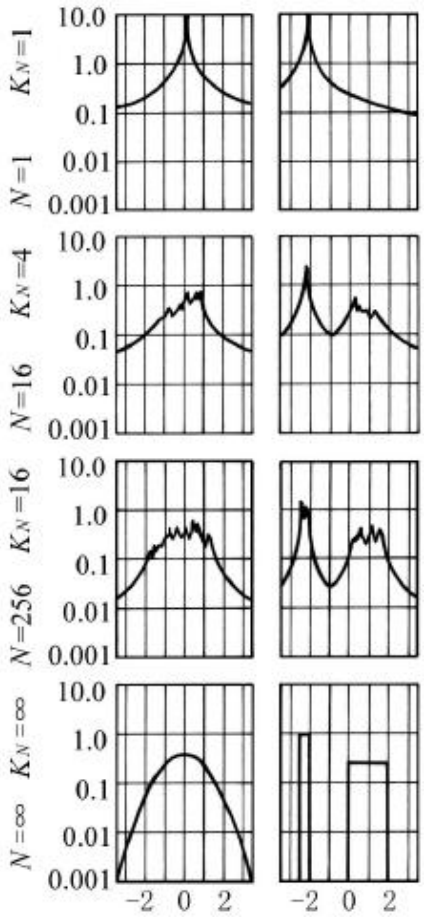

图 3-3 不同样本数和不同参数下 k_N 法的效果举例