第 3 章

CHAPTER 3

概率密度函数的参数估计

第 2 章讨论了贝叶斯决策理论，在采用贝叶斯决策理论进行分类决策时，需要计算后验概率 $P(\omega_i|\boldsymbol{X})$，或者需要事先知道各类的先验概率 $P(\omega_i)$和样本的类条件概率密度函数 $p(\boldsymbol{X}|\omega_i)$，但实际应用中先验概率和类条件概率密度函数往往是未知的。通常，对研究的对象只有一些模糊性的知识，或者通过实验采样而得到的一些样本。这就需要根据已有的样本，利用统计推断中的估计理论对样本的分布做出估计，然后将估计值当成真实值来使用。在模式识别问题中，先验概率的估计相对比较容易，它可以由各类样本在总体样本集中所占的比例进行估计。但类条件概率密度函数的估计却比较困难，从样本出发估计其函数形式和参数，这就是本章要讨论的参数估计问题。

3.1 概率密度函数估计概述

所谓的概率密度函数估计是已知某类别 ω_i 的样本 $\boldsymbol{X}_i(i=1,2,\cdots,N)$，采用某种规则估计出样本所属类的概率函数 $p(\boldsymbol{X}|\omega_i)$。从估计的方法来讲，可分为参数估计和非参数估计。参数估计是先假定样本的类条件概率密度函数 $p(\boldsymbol{X}|\omega_i)$的类型已知，如服从正态分布、二项分布，再用已知类别的学习样本估计函数里的未知参数θ，这项工作也叫训练或学习。参数估计的方法通常采用的是最大似然估计方法和贝叶斯估计方法。非参数估计则是类条件概率密度函数的形式未知，直接用已知类别的学习样本去估计函数的数学模型。非参数估计的方法通常采用的是 Parzen 窗法、k_N-近邻法等。

为了便于理解，首先介绍参数估计中的一些基本概念。

(1) 统计量。假如概率密度函数的形式已知，但表征函数的参数θ 未知，则可将θ 的估计值构造成样本 $\boldsymbol{X}_i(i=1,2,\cdots,N)$的某种函数，这种函数称为统计量。参数估计的任务，就是利用样本求出参数 θ 的估计值$\hat{\theta}=\theta(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N)$。

(2) 参数空间。参数θ 的取值范围称为参数空间，书中用Θ 来表示。

(3) 点估计、估计量和估计值。构造一统计量作为未知参数θ 的估计，称为点估计，由样本$(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N)$作为自变量计算出来的$\hat{\theta}$ 值称为估计值，$\hat{\theta}$ 称为估计量。

(4) 区间估计。通过从总体中抽取的样本，根据一定的正确度与精确度的要求，构造出适当的区间，作为未知参数的真值所在范围的估计。

下面分别介绍最大似然估计、贝叶斯估计、贝叶斯学习三种参数估计方法，以及 Parzen

窗法和 k_N-近邻法两种非参数估计方法。

3.2　最大似然估计

对 c 类问题，设类别 ω_i 的概率密度函数 $p(\boldsymbol{X}|\omega_i)$ 的形式已知，但表征该函数的参数未知，记为 θ_i。从 ω_i 中独立抽取 N 个样本，如果能从这 N 个样本中推断出 θ_i 的估计值 $\hat{\theta}_i$，则完成了概率密度函数 $p(\boldsymbol{X}|\omega_i)$ 的估计。为了强调 $p(\boldsymbol{X}|\omega_i)$ 与参数 θ_i 的关联性，也可把概率密度函数写成 $p(\boldsymbol{X}|\omega_i,\theta_i)$。例如，如果已知某一类别 ω_i 概率密度函数服从正态分布，则未知参数 θ_i 包含了表征该函数的均值 $\boldsymbol{\mu}_i$ 和协方差 $\boldsymbol{\Sigma}_i$ 的全部信息，对参数 θ_i 的估计，实质上就是对正态函数的均值 $\boldsymbol{\mu}_i$ 和协方差 $\boldsymbol{\Sigma}_i$ 的估计。下面首先给出似然函数的定义，然后从似然函数出发，讨论最大似然估计的原理。

1. 似然函数

从 ω_i 类中抽取 N 个样本 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，由于这 N 个样本均来自 ω_i 类，因此可将其概率密度函数 $p(\boldsymbol{X}|\omega_i,\theta_i)$ 简化为 $p(\boldsymbol{X}|\theta)$，则称这 N 个样本的联合概率密度函数 $p(\boldsymbol{X}^{(N)},\theta)$ 为相对于样本集 $\boldsymbol{X}^{(N)}$ 的 θ 的似然函数。由于 θ 是概率密度函数的一个确定性参数集，因此概率密度函数 $p(\boldsymbol{X}^{(N)},\theta)$ 实际上就是条件概率 $p(\boldsymbol{X}^N|\theta)$。如果 N 个样本为独立抽取，似然函数可表示为

$$p(\boldsymbol{X}^{(N)}\mid\theta)=p(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\mid\theta)=\prod_{k=1}^{N}p(\boldsymbol{X}_k\mid\theta)\tag{3-1}$$

第 5 集
微课视频

式(3-1)是在参数 θ 下观测到的样本集 $\boldsymbol{X}^{(N)}$ 的概率(联合分布)密度。

2. 最大似然估计

从 ω_i 类中独立抽取 N 个样本 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，那么这 N 个样本最有可能来自哪个概率密度函数，或者说与这 N 个样本最匹配的未知参数 θ 是什么。这是最大似然估计要解决的问题，它的主要思想是，给定样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，通过极大化似然函数 $p(\boldsymbol{X}^{(N)}|\theta)$ 去求与样本匹配的参数 θ，θ 的最大似然估计量 $\hat{\theta}$ 就是使似然函数达到最大的估计量，图 3-1 所示为 θ 为一维时的最大似然估计。由 $\dfrac{\mathrm{d}p(\boldsymbol{X}^{(N)}|\theta)}{\mathrm{d}\theta}=0$ 可求得解。

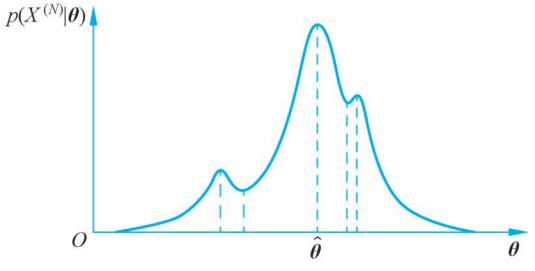


图 3-1　θ 为一维时的最大似然估计

由于对数函数具有单调性，为了便于分析，对似然函数取对数

$$H(\theta)=\ln p(\boldsymbol{X}^{(N)}\mid\theta)\tag{3-2}$$

显然，当估计量$\hat{\theta}$使数函数取最大值时，似然函数达到最大值，θ的最大似然估计是下面微分方程的解：

$$\frac{\mathrm{d}H(\theta)}{\mathrm{d}\theta}=0 \tag{3-3}$$

设ω_i类的概率密度函数包含q个未知参数，则θ为q维向量

$$\theta=[\theta_1,\theta_2,\cdots,\theta_q]^{\mathrm{T}} \tag{3-4}$$

此时

$$H(\theta)=\ln p(\boldsymbol{X}^{(N)}\mid\theta)=\sum_{k=1}^{N}\ln p(\boldsymbol{X}_k\mid\theta) \tag{3-5}$$

式(3-3)可表示为

$$\frac{\partial}{\partial\theta}\left[\sum_{k=1}^{N}\ln p(\boldsymbol{X}_k\mid\theta)\right]=0 \tag{3-6}$$

即

$$\begin{cases}\displaystyle\sum_{k=1}^{N}\frac{\partial}{\partial\theta_1}\ln p(\boldsymbol{X}_k\mid\theta)=0\\ \displaystyle\sum_{k=1}^{N}\frac{\partial}{\partial\theta_2}\ln p(\boldsymbol{X}_k\mid\theta)=0\\ \vdots\\ \displaystyle\sum_{k=1}^{N}\frac{\partial}{\partial\theta_q}\ln p(\boldsymbol{X}_k\mid\theta)=0\end{cases} \tag{3-7}$$

求解式(3-7)微分方程组，可得到θ的最大似然估计值$\hat{\theta}$。

【例 3.1】 设从ω_i中抽取了N个样本，表示为$\boldsymbol{X}^{(N)}$，这N个样本是从一维正态分布概率密度函数$p(\boldsymbol{X}|\omega_i)$[或$p(\boldsymbol{X}^{(N)}|\theta)$]总体中独立抽取的，用最大似然估计方法，估计正态分布的均值和协方差。

【解】 $p(\boldsymbol{X}|\omega_i)$可表示为$p(\boldsymbol{X}|\theta)\sim N(\mu,\sigma^2)$，其中，$\theta=[\theta_1,\theta_2]^{\mathrm{T}}$，$\theta_1=\mu$，$\theta_2=\sigma^2$。因为$\boldsymbol{X}^{(N)}$是从$\omega_i$中独立抽取的$N$个样本，则$\theta$的似然函数为

$$p(\boldsymbol{X}^{(N)}\mid\theta)=\prod_{k=1}^{N}p(\boldsymbol{X}_k\mid\theta)$$

式中，$p(\boldsymbol{X}_k|\theta)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left[-\dfrac{(\boldsymbol{X}_k-\mu)^2}{2\sigma^2}\right]$。

取似然函数的对数，得

$$\ln p(\boldsymbol{X}_k\mid\theta)=-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{(\boldsymbol{X}_k-\mu)^2}{2\sigma^2}$$

函数$\ln p(\boldsymbol{X}_k|\theta)$分别对$\theta_1$和$\theta_2$求导，并令导数为零，即

$$\begin{cases}\displaystyle\sum_{k=1}^{N}\frac{\partial}{\partial\theta_1}\ln p(\boldsymbol{X}_k\mid\theta)=\sum_{k=1}^{N}\frac{\boldsymbol{X}_k-\theta_1}{\theta_2}=0\\ \displaystyle\sum_{k=1}^{N}\frac{\partial}{\partial\theta_2}\ln p(\boldsymbol{X}_k\mid\theta)=\sum_{k=1}^{N}\left[\frac{-1}{2\theta_2}+\frac{(\boldsymbol{X}_k-\theta_1)^2}{2\theta_2^2}\right]=0\end{cases}$$

由以上方程组解得均值和方差的估计量为

$$\hat{\mu}=\hat{\theta}_1=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{X}_k \tag{3-8}$$

$$\hat{\sigma}^2=\hat{\theta}_2=\frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{X}_k-\hat{\mu})^2 \tag{3-9}$$

对于一般的多维正态分布情况，用类似例 3.1 的方法，可以求得其最大似然估计值为

$$\hat{\boldsymbol{\mu}}_i=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{X}_k \tag{3-10}$$

$$\hat{\boldsymbol{\Sigma}}_i=\frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \tag{3-11}$$

式(3-10)与式(3-11)表明，在多元正态分布情况下，均值向量的最大似然估计是样本的算术平均值，而协方差矩阵的最大似然估计是 N 个矩阵的$(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)^{\mathrm{T}}$ 的算术平均值。

3.3　贝叶斯估计与贝叶斯学习

1. 贝叶斯估计

贝叶斯估计可描述为给定样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，对样本的概率密度函数的真实参数$\boldsymbol{\theta}$ 进行估计，使其估计值$\hat{\boldsymbol{\theta}}$ 带来的贝叶斯风险最小。回顾上一章的最小风险贝叶斯决策，可以看出贝叶斯决策和贝叶斯估计都是以贝叶斯风险最小为基础，只是要解决的问题不同，前者是要判决样本 $\boldsymbol{X}$ 的类别归属，而后者是估计样本集 $\boldsymbol{X}^{(N)}$ 所属总体分布的参数，本质上二者是统一的。贝叶斯决策和贝叶斯估计各变量的对应关系如表 3-1 所示。

表 3-1　贝叶斯决策和贝叶斯估计各变量的对应关系

决 策 问 题	估 计 问 题
样本 x	样本集 $\boldsymbol{X}^{(N)}$
决策 α_i	估计量$\hat{\boldsymbol{\theta}}$
真实状态 ω_j	真实参数$\boldsymbol{\theta}$
状态空间 A 是离散空间	参数空间$\boldsymbol{\Theta}$
先验概率 $P(\omega_j)$	参数的先验分布 $p(\boldsymbol{\theta})$

在第 2 章研究分类问题时，我们用式(2-11)定义了条件平均风险：

$$R(\alpha_i\mid\boldsymbol{X})=E[L(\alpha_i\mid\omega_j)]=\sum_{j=1}^{c}L(\alpha_i\mid\omega_j)\cdot P(\omega_j\mid\boldsymbol{X}),\quad i=1,2,\cdots,a$$

参考上式，并对照表 3-1 中贝叶斯决策和贝叶斯估计各变量的对应关系，可以定义在观测样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 的条件下，用$\hat{\boldsymbol{\theta}}$ 作为$\boldsymbol{\theta}$ 的估计的期望损失为

$$R(\hat{\boldsymbol{\theta}}\mid\boldsymbol{X}^{(N)})=\int_{\boldsymbol{\Theta}}L(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})p(\boldsymbol{\theta}\mid\boldsymbol{X}^{(N)})\mathrm{d}\boldsymbol{\theta} \tag{3-12}$$

式中，$L(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$为用$\hat{\boldsymbol{\theta}}$ 代替$\boldsymbol{\theta}$ 所造成的损失，$\boldsymbol{\Theta}$ 为参数空间。

考虑 $\boldsymbol{X}^{(N)}$ 的各种取值，应该求 $R(\hat{\boldsymbol{\theta}}|\boldsymbol{X}^{(N)})$在空间$\boldsymbol{\Omega}^N=\boldsymbol{\Omega}\times\boldsymbol{\Omega}\times\cdots\times\boldsymbol{\Omega}$ 中的期望，即

$$R=\int_{\boldsymbol{\Omega}^N}R(\hat{\boldsymbol{\theta}}\mid\boldsymbol{X}^{(N)})p(\boldsymbol{X}^{(N)})\mathrm{d}\boldsymbol{X}^{(N)} \tag{3-13}$$

将式(3-12)代入上式，得

$$R=\int_{\Omega^N}\int_{\Theta}L(\hat{\theta},\theta)p(\theta\mid\boldsymbol{X}^{(N)})p(\boldsymbol{X}^{(N)})\mathrm{d}\theta\mathrm{d}\boldsymbol{X}^{(N)} \tag{3-14}$$

使 R 最小的参数θ 的估计值$\hat{\theta}$ 即贝叶斯估计。显然，损失函数 $L(\hat{\theta},\theta)$对$\hat{\theta}$ 的求解有重要影响，当选用不同形式的损失函数时，所得到的贝叶斯估计值$\hat{\theta}$ 也不同。当损失函数为二次函数时，有

$$L(\hat{\theta},\theta)=(\theta-\hat{\theta})^{\mathrm{T}}(\theta-\hat{\theta}) \tag{3-15}$$

可证明$\hat{\theta}$ 的求解公式如下：

$$\hat{\theta}=\int_{\Theta}\theta\, p(\theta\mid\boldsymbol{X}^{(N)})\mathrm{d}\theta \tag{3-16}$$

上式表明，θ 的最小方差贝叶斯估计是观测样本集 $\boldsymbol{X}^{(N)}$ 条件下的θ 的条件期望。

综上所述，观测到一组样本 $\boldsymbol{X}^{(N)}$，通过似然函数 $p(\boldsymbol{X}^{(N)}|\theta)$并利用贝叶斯公式将随机变量$\theta$ 的先验概率密度 $p(\theta)$转换为后验概率密度，然后根据θ 的后验概率密度求出估计量$\hat{\theta}$，具体步骤如下。

(1) 确定θ 的先验概率密度 $p(\theta)$。

(2) 由样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$求出样本的联合概率密度函数也就是$\theta$ 的似然函数 $p(\boldsymbol{X}^{(N)}|\theta)$。

(3) 利用贝叶斯公式求出θ 的后验概率密度：

$$p(\theta\mid\boldsymbol{X}^{(N)})=\frac{p(\boldsymbol{X}^{(N)}\mid\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{X}^{(N)}\mid\theta)p(\theta)\mathrm{d}\theta} \tag{3-17}$$

(4) 根据式(3-16)求贝叶斯估计量$\hat{\theta}$。

在步骤(2)涉及 $p(\boldsymbol{X}^{(N)}|\theta)$的求解，当样本的类概率密度函数的类型已知时，由于样本 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N$ 为独立抽取，因此有

$$p(\boldsymbol{X}^{(N)}\mid\theta)=p(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\mid\theta)=\prod_{i=1}^{N}p(\boldsymbol{X}_i\mid\theta)$$

2. 贝叶斯学习

贝叶斯学习是指利用θ 的先验概率密度 $p(\theta)$及样本提供的信息递推求出θ 的后验概率密度 $p(\theta\mid\boldsymbol{X}^{(N)})$，根据后验概率密度直接求出类概率密度函数 $p(\boldsymbol{X}\mid\boldsymbol{X}^{(N)})$。因此，贝叶斯学习和贝叶斯估计的前提条件完全相同，区别在于当求出后验概率密度 $p(\theta\mid\boldsymbol{X}^{(N)})$后，贝叶斯学习没有对参数$\theta$ 进行估计，而是直接进行总体概率密度的推断得到 $p(\boldsymbol{X}\mid\boldsymbol{X}^{(N)})$。所以，贝叶斯学习的前三步与贝叶斯估计完全一致，最后 $p(\boldsymbol{X}\mid\boldsymbol{X}^{(N)})$可由迭代计算完成。迭代计算式的推导如下。

$p(\boldsymbol{X}\mid\omega_i)$由未知参数$\theta$ 确定，可写为 $p(\boldsymbol{X}\mid\omega_i)=p(\boldsymbol{X}\mid\theta)$，假定 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$是独立抽取的 ω_i 类的一组样本，设θ 的后验概率密度函数为 $p(\theta\mid\boldsymbol{X}^{(N)})$，根据贝叶斯公式有

$$p(\theta\mid\boldsymbol{X}^{(N)})=\frac{p(\boldsymbol{X}^{(N)}\mid\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{X}^{(N)}\mid\theta)p(\theta)\mathrm{d}\theta} \tag{3-18}$$

由条件独立可知，当 $N>1$ 时

$$p(\boldsymbol{X}^{(N)}\mid\theta)=p(\boldsymbol{X}_N\mid\theta)p(\boldsymbol{X}^{(N-1)}\mid\theta) \tag{3-19}$$

式中，$\boldsymbol{X}^{(N-1)}$ 表示除样本 $\boldsymbol{X}_N$ 以外其余样本的集合。

将式(3-19)代入式(3-18)得

$$p(\theta\mid\boldsymbol{X}^{(N)})=\frac{p(\boldsymbol{X}_N\mid\theta)p(\boldsymbol{X}^{(N-1)}\mid\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{X}_N\mid\theta)p(\boldsymbol{X}^{(N-1)}\mid\theta)p(\theta)\mathrm{d}\theta} \tag{3-20}$$

类似地，由式(3-18)也可推导出

$$p(\theta\mid\boldsymbol{X}^{(N-1)})=\frac{p(\boldsymbol{X}^{(N-1)}\mid\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{X}^{(N-1)}\mid\theta)p(\theta)\mathrm{d}\theta} \tag{3-21}$$

将式(3-21)代入式(3-20)得

$$p(\theta\mid\boldsymbol{X}^{(N)})=\frac{p(\boldsymbol{X}_N\mid\theta)p(\theta\mid\boldsymbol{X}^{(N-1)})}{\int_{\Theta}p(\boldsymbol{X}_N\mid\theta)p(\theta\mid\boldsymbol{X}^{(N-1)})\mathrm{d}\theta} \tag{3-22}$$

式(3-22)就是利用样本集 $\boldsymbol{X}^{(N)}$ 估计 $p(\theta\mid\boldsymbol{X}^{(N)})$ 的迭代计算方法。对于参数估计的递推贝叶斯方法，其迭代过程是贝叶斯学习的过程。下面简述迭代式的使用。

(1) 根据先验知识得到 θ 的先验概率密度函数的初始估计 $p(\theta)$。相当于 $N=0$ 时 $(\boldsymbol{X}^{(N)}=\boldsymbol{X}^{(0)})$ 密度函数的一个估计。

(2) 用 $\boldsymbol{X}_1$ 对初始的 $p(\theta)$ 进行修改，根据式(3-22)，令 $N=1$，得

$$p(\theta\mid\boldsymbol{X}^{(1)})=p(\theta\mid\boldsymbol{X}_1)=\frac{p(\boldsymbol{X}_1\mid\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{X}_1\mid\theta)p(\theta)\mathrm{d}\theta} \tag{3-23}$$

$p(\boldsymbol{X}_1\mid\theta)$ 根据式 $p(\boldsymbol{X}\mid\omega_i)=p(\boldsymbol{X}\mid\theta)$ 计算得到。

(3) 给出 $\boldsymbol{X}_2$，对用 $\boldsymbol{X}_1$ 估计的结果进行修改

$$p(\theta\mid\boldsymbol{X}^{(2)})=p(\theta\mid\boldsymbol{X}_1,\boldsymbol{X}_2)=\frac{p(\boldsymbol{X}_2\mid\theta)p(\theta\mid\boldsymbol{X}^{(1)})}{\int_{\Theta}p(\boldsymbol{X}_2\mid\theta)p(\theta\mid\boldsymbol{X}^{(1)})\mathrm{d}\theta} \tag{3-24}$$

(4) 逐次给出 $\boldsymbol{X}_3,\boldsymbol{X}_4,\cdots,\boldsymbol{X}_N$，每次在前一次的基础上进行修改，$p(\theta\mid\boldsymbol{X}^{(N-1)})$ 可以看成 $p(\theta\mid\boldsymbol{X}^{(N)})$ 的先验概率。最后，当 $\boldsymbol{X}_N$ 给出后，得

$$p(\theta\mid\boldsymbol{X}^{(N)})=\frac{p(\boldsymbol{X}_N\mid\theta)p(\theta\mid\boldsymbol{X}^{(N-1)})}{\int_{\Theta}p(\boldsymbol{X}_N\mid\theta)p(\theta\mid\boldsymbol{X}^{(N-1)})\mathrm{d}\theta}$$

(5) $p(\boldsymbol{X}\mid\omega_i)$ 直接由 $p(\theta\mid\boldsymbol{X}^{(N)})$ 计算得到，此时 $p(\boldsymbol{X}\mid\omega_i)$ 可以写为 $p(\boldsymbol{X}\mid\boldsymbol{X}^{(N)})$，由一般概率公式得

$$p(\boldsymbol{X}\mid\boldsymbol{X}^{(N)})=\int p(\boldsymbol{X},\theta\mid\boldsymbol{X}^{(N)})\,\mathrm{d}\theta=\int p(\boldsymbol{X}\mid\theta)p(\theta\mid\boldsymbol{X}^{(N)})\mathrm{d}\theta \tag{3-25}$$

这就是贝叶斯学习。下面通过两个例子，讨论正态分布密度函数的贝叶斯估计和贝叶斯学习问题。

【例 3.2】 对一个单变量正态分布，已知方差 σ^2，试用贝叶斯估计方法估计均值 μ。

【解】 设 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 是 ω_i 类的 N 个独立抽取的样本，$p(\boldsymbol{X}\mid\mu)\sim N(\mu,\sigma^2)$，

μ为未知随机参数。假定μ服从正态分布，且μ的先验概率密度 $p(\mu)$也为正态分布，即 $p(\mu)\sim N(\mu_0,\sigma_0^2)$。利用贝叶斯公式求$\mu$的后验概率密度函数 $p(\mu|\boldsymbol{X}^{(N)})$，有

$$p(\mu|\boldsymbol{X}^{(N)})=\frac{p(\boldsymbol{X}^{(N)}|\mu)p(\mu)}{\int p(\boldsymbol{X}^{(N)}|\mu)p(\mu)\mathrm{d}\mu}$$

式中，μ的似然函数 $p(\boldsymbol{X}^{(N)}|\mu)$可以表示为 $p(\boldsymbol{X}^{(N)}|\mu)=\prod_{k=1}^{N}p(\boldsymbol{X}_k|\mu)$，且有 $p(\mu|\boldsymbol{X}^{(N)})=\alpha\prod_{k=1}^{N}p(\boldsymbol{X}_k|\mu)p(\mu)$，$\alpha=1/\int p(\boldsymbol{X}^{(N)}|\mu)p(\mu)\mathrm{d}\mu$，$\alpha$ 是与μ无关的比例因子，不影响 $p(\mu|\boldsymbol{X}^{(N)})$的形式。

因为

$$p(\boldsymbol{X}|\mu)\sim N(\mu,\sigma^2)\quad p(\mu)\sim N(\mu_0,\sigma_0^2)$$

所以

$$\begin{aligned}p(\mu|\boldsymbol{X}^{(N)})&=\alpha\prod_{k=1}^{N}p(\boldsymbol{X}_k|\mu)p(\mu)\\&=\alpha\prod_{k=1}^{N}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(\boldsymbol{X}_k-\mu)^2}{2\sigma^2}\right]\cdot\frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]\\&=\alpha'\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{N}\frac{(\mu-\boldsymbol{X}_k)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]\right\}\\&=\alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}\boldsymbol{X}_k+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}\end{aligned}\tag{3-26}$$

式中，α'和α''是与μ无关的项。

将 $p(\mu|\boldsymbol{X}^{(N)})$写为正态分布密度函数的标准形式 $N(\mu_N,\sigma_N^2)$：

$$p(\mu|\boldsymbol{X}^{(N)})=\frac{1}{\sqrt{2\pi}\,\sigma_N}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right\}\tag{3-27}$$

比较式(3-26)与式(3-27)，可求得 μ_N 和σ_N^2 分别为

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\boldsymbol{m}_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0\tag{3-28}$$

$$\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\tag{3-29}$$

式中，$\boldsymbol{m}_N=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{X}_k$。

将所求的μ_N 和σ_N^2 代入式(3-27)就得到了μ的后验概率密度函数 $p(\mu|\boldsymbol{X}^{(N)})$。这时，由式(3-16)计算$\mu$的贝叶斯估计为

$$\hat{\mu}=\int\mu p(\mu|\boldsymbol{X}^{N})\mathrm{d}\mu=\int\mu\frac{1}{\sqrt{2\pi}\,\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right]\mathrm{d}\mu=\mu_N$$

将式(3-28)结果代入上式，得

$$\hat{\boldsymbol{\mu}} = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\boldsymbol{m}_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\boldsymbol{\mu}_0 \tag{3-30}$$

当 $N(\mu_0,\sigma_0^2)=N(0,1)$ 且 $\sigma^2=1$ 时，有

$$\hat{\boldsymbol{\mu}} = \frac{N}{N+1}\boldsymbol{m}_N = \frac{1}{N+1}\sum_{k=1}^{N}\boldsymbol{X}_k \tag{3-31}$$

也就是说，此时μ的贝叶斯估计与最大似然估计有类似的形式，只是分母不同。

【例 3.3】 同例 3.2，用贝叶斯学习方法计算。

【解】 递推求解后验概率密度 $p(\mu|\boldsymbol{X}^{(N)})$：

$$p(\mu|\boldsymbol{X}^{(N)}) = \frac{1}{\sqrt{2\pi}\,\sigma_N}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right\}$$

式中，μ_N 为观察了 N 个样本后对μ的最好估计，σ_N^2 为估计的不确定性，分别由式(3-28)和式(3-29)求得。

由后验概率密度 $p(\mu|\boldsymbol{X}^{(N)})$计算类概率密度函数 $p(\boldsymbol{X}|\boldsymbol{X}^{(N)})$：

$$\begin{aligned}
p(\boldsymbol{X}\mid\boldsymbol{X}^{(N)}) &= \int p(\boldsymbol{X}\mid\mu)p(\mu|\boldsymbol{X}^{(N)})\mathrm{d}\mu \\
&= \int \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(\boldsymbol{X}-\mu)^2}{2\sigma^2}\right]\cdot\frac{1}{\sqrt{2\pi}\,\sigma_N}\exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]\mathrm{d}\mu \\
&= \int \frac{1}{\sqrt{2\pi}\,\sigma\sigma_N}\exp\left[-\frac{(\boldsymbol{X}-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}\right]\cdot\frac{1}{\sqrt{2\pi}\,\sigma_N} \\
&\quad \int\exp\left[-\frac{\sigma^2+\sigma_N^2}{2\sigma^2\sigma_N^2}\left(\mu-\frac{\sigma_N^2\boldsymbol{X}+\sigma^2\mu_N}{\sigma_N^2+\sigma^2}\right)\right]\mathrm{d}\mu \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_N^2}}\exp\left[-\frac{(\boldsymbol{X}-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}\right]
\end{aligned} \tag{3-32}$$

可见 $p(\boldsymbol{X}|\boldsymbol{X}^{(N)})$是正态分布，均值为$\mu_N$，方差为$(\sigma^2+\sigma_N^2)$。均值与贝叶斯估计的结果是相同的，原方差$\sigma^2$ 增加到$(\sigma^2+\sigma_N^2)$，这是由于用 μ 的估计值代替了真实值，引起了不确定性的增加。

对于多维正态分布，可以采用与一维情况类似的方法估计均值向量，但计算比较复杂。

3.4　非参数估计

以上内容讨论了最大似然估计、贝叶斯估计和贝叶斯学习这三种参数估计方法，其共同的特点是样本概率密度函数的分布的形式已知，而表征函数的参数未知，所需要做的工作是从样本估计出参数的最优取值。但在实际应用中，上述条件往往并不能得到满足，人们并不知道概率密度函数的分布形式，或者函数分布并不典型，或者不能写出某些参数的函数。为了设计贝叶斯分类器，仍然需要获取概率密度函数的分布知识，所以非常有必要研究如何从样本出发，直接推断其概率密度函数。于是，人们提出一些直接用样本来估计总体分布的方

法，称为估计分布的非参数法。

非参数估计方法的任务是从样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 中估计样本空间 Ω 中任何一点的概率密度 $p(\boldsymbol{X})$。如果样本集来自某个确定类别（如 ω_i 类），则估计的结果为该类的类条件概率密度 $p(\boldsymbol{X}|\omega_i)$。如果样本集来自多个类别，且不能分清哪个样本来自哪个类别，则估计结果为混合概率密度。

3.4.1 非参数估计的基本方法

下面从一个例子说明非参数估计的基本思想。假如样本集 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 由 N 个一维样本组成，每个样本 $\boldsymbol{X}_i$ 在以 $\boldsymbol{X}_i$ 为中心，宽度为 h 的范围内，对分布的贡献为 a。显然，可以把每个样本在 $\boldsymbol{X}_i$ 点的“贡献”相加作为这点的概率密度 $p(\boldsymbol{X}_i)$ 的估计。对所有的样本 $\boldsymbol{X}$ 都这么做，就可以得到总体分布 $p(\boldsymbol{X})$ 的估计值。通常，采用某种函数表示某一样本对某点概率密度的贡献，则某点概率密度 $p(\boldsymbol{X})$ 的估计为所有样本所做贡献的线性组合。非参数估计的原理如图 3-2 所示。

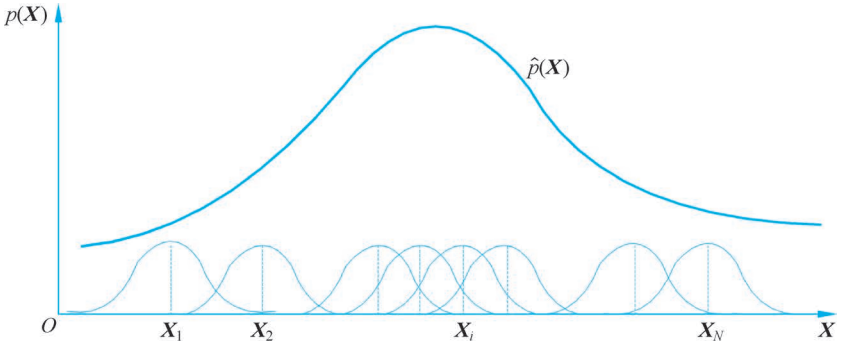


图 3-2 非参数估计的原理

当然，也可以认为每个样本对自己所在位置的分布“贡献”最大，距离越远，贡献越小。下面讨论如何估计概率密度函数。

设一个随机向量 $\boldsymbol{X}$ 落入特征空间区域 R 的概率 P 为

$$P=\int_R p(\boldsymbol{X})\mathrm{d}\boldsymbol{X} \tag{3-33}$$

式中，$p(\boldsymbol{X})$ 是 $\boldsymbol{X}$ 的概率密度函数，P 是概率密度函数的一种平均形式，对 P 做估计就是估计出 $p(\boldsymbol{X})$ 的这个平均值。

设 N 个样本 $\boldsymbol{X}^{(N)}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，它们是从概率密度为 $p(\boldsymbol{X})$ 的总体分布中独立抽取的，则 N 个样本中有 k 个样本落入区域 R 的概率最大，则可以得到

$$k \doteq N\hat{P} \tag{3-34}$$

$$\hat{P} \doteq k/N \tag{3-35}$$

式中，$\hat{P}$ 希望是 $\boldsymbol{X}$ 落入区域 R 中的概率 P 的一个理想的估计，但我们要估计的是类概率密度函数 $p(\boldsymbol{X})$ 的估计 $\hat{p}(\boldsymbol{X})$。为此，设 $p(\boldsymbol{X})$ 连续，且区域 R 足够小，以致 $p(\boldsymbol{X})$ 在这样小的区域中没有什么变化，由此可得

$$P=\int_R p(\boldsymbol{X})\mathrm{d}\boldsymbol{X}\approx p(\boldsymbol{X})V \tag{3-36}$$

式中，$\boldsymbol{X}$ 是 R 中的一个点，V 是 R 的“体积”。

由式(3-35)和式(3-36)可得

$$\frac{k}{N}\doteq\hat{P}=\int_R\hat{p}(\boldsymbol{X})\mathrm{d}\boldsymbol{X}=\hat{p}(\boldsymbol{X})V \tag{3-37}$$

所以

$$\hat{p}(\boldsymbol{X})=\frac{k/N}{V} \tag{3-38}$$

式(3-38)就是 $\boldsymbol{X}$ 点概率密度的估计，它与样本数 N、包含 $\boldsymbol{X}$ 的区域 R 的体积 V 和落入 R 中的样本数有关。

从理论上来讲，如果使 $\hat{p}(\boldsymbol{X})$ 趋近 $p(\boldsymbol{X})$，就必须让体积 V 趋近于零，同时 k、N 趋向于无穷大。但事实上，V 不可能无穷小，样本也总是有限的，N 不可能无穷大，所以 $\hat{p}(\boldsymbol{X})$ 总是存在误差。如果碰巧有一个或几个样本重合于 $\boldsymbol{X}$ 出现在 R，则会使估计发散，甚至到无穷大。因此要是采用这种估计，我们在使用式(3-38)时就必须注意 V，k，k/N 随 N 变化的趋势，使得当 N 适当增大时能保持式(3-38)的合理性。

从理论上考虑，假设有无穷多个样本，可以采取如下措施去提高 $\boldsymbol{X}$ 处的概率密度 $p(\boldsymbol{X})$ 的估计精度。构造一个区域序列 R_1，R_2，…，对 R_1 采用一个样本进行估计，对 R_2 采用两个样本进行估计，以此类推，对 R_N 采用 N 个样本进行估计。设 V_N 是区域 R_N 的体积，k_N 是落入 R_N 的样本个数。第 N 次估计的总体概率密度为

$$\hat{p}_N(\boldsymbol{X})=\frac{k_N/N}{V_N} \tag{3-39}$$

为了保证上述估计的合理性，应满足以下三个条件：

(1) $$\lim_{N\to\infty}V_N=0 \tag{3-40}$$

(2) $$\lim_{N\to\infty}k_N=\infty \tag{3-41}$$

(3) $$\lim_{N\to\infty}\frac{k_N}{V_N}=0 \tag{3-42}$$

此时，总体概率密度 $\hat{p}_N(\boldsymbol{X})$ 的估计值收敛于实际值 $p(\boldsymbol{X})$。

在上述条件中，条件(1)保证了空间平均式的收敛性；条件(2)保证了频数比的收敛性；条件(3)保证了估计式的收敛性。以上三个条件说明当 N 增大时，落入 R_N 的样本数也增加；V_N 不断减少，以使 $\hat{p}_N(\boldsymbol{X})$ 趋于 $p(\boldsymbol{X})$；尽管在一个小区域 R_N 中落入了大量的样本，但它的数目与样本总数相比还是可以忽略的。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法。

(1) Parzen 窗函数法：选定一个中心在 $\boldsymbol{X}$ 处的区域 R_N，其体积 V_N 以 N 的某个函数(如 $V_N=1/\sqrt{N}$)的关系不断缩小，同时需对 k_N 和 k_N/N 加以限制，以使 $\hat{p}_N(\boldsymbol{X})$ 收敛于 $p(\boldsymbol{X})$，然后计算落入 R_N 的样本数 k_N，用来估计局部密度 $\hat{p}_N(\boldsymbol{X})$ 的值。

(2) k_N-近邻法：令 k_N 为 N 的某个函数(例如，$k_N=\sqrt{N}$)，以 $\boldsymbol{X}$ 为中心构造一个体积为 V_N 的区域 R_N，使 R_N 恰好包含 k_N 个样本，用这时的体积来估计 $\hat{p}_N(\boldsymbol{X})$

的值。

3.4.2 Parzen 窗法

假设区域 R_N 为 d 维超立方体，向量 $\boldsymbol{X}$ 为 d 维特征空间中的一个点，超立方体 R_N 以原点为中心，侧棱长为 h_N，则其体积 V_N 为

$$V_N = h_N^d \tag{3-43}$$

为了计算 R_N 中包含的样本数 k_N，定义 d 维空间的基本窗函数：

$$\varphi(u) = \begin{cases} 1, & |u_j| \leqslant \dfrac{1}{2}, j = 1,2,\cdots,d \\ 0, & \text{其他} \end{cases} \tag{3-44}$$

式中，$u = (u_1, u_2, u_3, \cdots, u_d)$，$\varphi(u)$ 称为 Parzen 窗函数，它是以原点为中心的超立方体。

利用函数 $\varphi(u)$ 可以实现对落在区域 R_N 的样本进行计数，当 $\boldsymbol{X}_i$ 落在以 $\boldsymbol{X}$ 为中心、体积为 V_N 的超立方体内时，计数为 1，即

$$\varphi(u) = \varphi\left(\frac{\boldsymbol{X} - \boldsymbol{X}_i}{h_N}\right) = 1 \tag{3-45}$$

否则取值为零。因此，落入该立方体的样本数 k_N 为

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{\boldsymbol{X} - \boldsymbol{X}_i}{h_N}\right) \tag{3-46}$$

将式(3-46)代入式(3-39)，可得概率密度估计：

$$\hat{p}_N(\boldsymbol{X}) = \frac{k_N / N}{V_N} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi\left(\dfrac{\boldsymbol{X} - \boldsymbol{X}_i}{h_N}\right)}{V_N} \tag{3-47}$$

式(3-47)是 Parzen 窗法估计的基本公式。上式表明，Parzen 窗的估计函数实质上由一系列的基函数叠加而成，式(3-44)定义的窗函数即叠加的基函数，每个样本点处作为叠加节点，使用 k_N 个以样本 $\boldsymbol{X}_i$ 为中心的窗函数叠加作为 $\boldsymbol{X}$ 处的概率密度函数的估计，每一个样本对估计所起的作用依赖于它到 $\boldsymbol{X}_i$ 的距离。显然，概率密度函数与样本的密集程度相关，样本在该区域越密集，叠加函数的值越大，也意味着概率密度函数的估计值越大。

式(3-44)定义了 Parzen 窗法估计法的窗函数，从上面的讨论可以看出估计结果和窗函数密切相关。下面讨论窗函数需要满足什么条件，以及如何去选择合适的窗函数。为了使 $\hat{p}_N(\boldsymbol{X})$ 成为一个概率密度函数，则其必须满足概率密度函数的一般要求，即 $\hat{p}_N(\boldsymbol{X})$ 非负且积分为 1，相应地要求窗函数 $\varphi(u)$ 满足下面两个条件：

$$\varphi(u) \geqslant 0 \tag{3-48}$$

$$\int \varphi(u) \mathrm{d}u = 1 \tag{3-49}$$

上述两式表明，窗函数本身满足密度函数的要求。式(3-47)窗函数 $\varphi(u)$ 的非负性保证了 $\hat{p}_N(\boldsymbol{X})$ 的非负性，进一步有

$$\int p_N(\boldsymbol{X})\mathrm{d}\boldsymbol{X}=\int\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)\mathrm{d}\boldsymbol{X}=\frac{1}{N}\sum_{i=1}^{N}\int\frac{1}{V_N}\varphi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)\mathrm{d}\boldsymbol{X}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\int\varphi(u)\mathrm{d}u=1$$

从而证明了 $\hat{p}_N(\boldsymbol{X})$ 是一个概率密度函数。

由此可见，一个函数只要满足式(3-48)和式(3-49)，它就可以作为窗函数。除了上面选择的超立方体窗函数以外，还可以选择其他的窗函数形式。以一维窗函数为例，常用的方窗函数、正态窗函数和指数窗函数的定义如下。

(1) 方窗函数：

$$\varphi(u)=\begin{cases}1, & |u|\leqslant\dfrac{1}{2}\\ 0, & \text{其他}\end{cases}\tag{3-50}$$

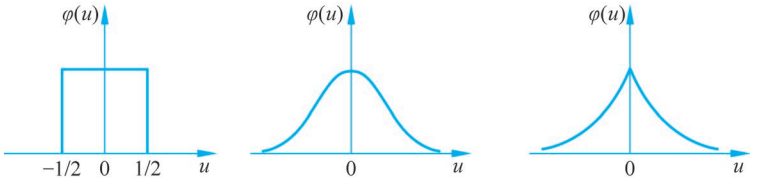
(2) 正态窗函数：

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)\tag{3-51}$$

(3) 指数窗函数：

$$\varphi(u)=\frac{1}{2}\exp(-|u|)\tag{3-52}$$

以上三种窗函数如图 3-3 所示。



(a) 方窗函数　(b) 正态窗函数　(c) 指数窗函数

图 3-3　三种窗函数

对密度函数的估计有影响的另一个因素是窗函数的宽度 h_N，下面分析 h_N 对估计量的影响。定义函数

$$\delta_N(\boldsymbol{X})=\frac{1}{V_N}\varphi\left(\frac{\boldsymbol{X}}{h_N}\right)\tag{3-53}$$

由式(3-47)可以表示为

$$\hat{p}_N(\boldsymbol{X})=\frac{1}{N}\sum_{i=1}^{N}\frac{\varphi\left(\dfrac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)}{V_N}=\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_N(\boldsymbol{X}-\boldsymbol{X}_i)}{V_N}\tag{3-54}$$

由 $V_N=h_N^d$ 可知，h_N 既影响 $\delta_N(\boldsymbol{X})$ 的幅度又影响它的宽度。如果 h_N 较大，则 $\delta_N(\boldsymbol{X})$ 的幅度就较小，从而 $\delta_N(\boldsymbol{X})$ 宽度较大，只有当 $\boldsymbol{X}_i$ 距离 $\boldsymbol{X}$ 较远时才使得 $\delta_N(\boldsymbol{X}-\boldsymbol{X}_i)$ 与 $\delta_N(0)$ 相差较大。此时，$\hat{p}_N(\boldsymbol{X})$ 就变成 N 个宽度较大的缓慢变化函数的叠加，造成估计值 $\hat{p}_N(\boldsymbol{X})$ 较平滑，跟不上函数 $p(\boldsymbol{X})$ 的变化，估计分辨率较低。反过来，如果 h_N 选得较小，则 $\delta_N(\boldsymbol{X}-\boldsymbol{X}_i)$ 的幅度就较大，$\delta_N(\boldsymbol{X})$ 宽度较小。此时 $\hat{p}_N(\boldsymbol{X})$ 就是 N 个以 $\boldsymbol{X}_i$ 为中心的尖脉

冲的叠加，$\hat{p}_N(\boldsymbol{X})$波动较大，从而使估计不稳定。

综上所述，h_N 的选取对 $\hat{p}_N(\boldsymbol{X})$的影响很大，$h_N$ 太大或太小，都对估计的精度不利。如何选择 h_N，需要一定的经验，当样本数目有限时，可做适当的折中，通过试探的方法选出合理的结果；当样本数目无限时，可让 V_N 随 N 的增大而缓慢地趋于零，从而使 $\hat{p}_N(\boldsymbol{X})$收敛于 $p(\boldsymbol{X})$。

【例 3.4】 设待估计的 $p(\boldsymbol{X})$是均值为 0、方差为 1 的正态密度函数，用 Parzen 窗法估计 $p(\boldsymbol{X})$，即求估计式 $\hat{p}_N(\boldsymbol{X})$。

【解】 考虑 $\boldsymbol{X}$ 是一维模式向量的情况。选择正态窗函数：

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

并设 $h_N=h_1/\sqrt{N}$，h_1 为可调节的参数。有

$$\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)^2\right]=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_1/\sqrt{N}}\right)^2\right]$$

这样，估计值 $\hat{p}_N(\boldsymbol{X})$是一个以样本为中心的正态密度的平均值：

$$\hat{p}_N(\boldsymbol{X})=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{N}}{h_1}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)$$

$$=\frac{1}{h_1\sqrt{N}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_1/\sqrt{N}}\right)^2\right]$$

当采集到正态分布样本后，就可求得估计值 $\hat{p}_N(\boldsymbol{X})$，结果如图 3-4 所示。图 3-4 给出了 h_1 分别取 0.25、1 和 4，样本量 N 取 1、16、256 和∞时 $\hat{p}_N(\boldsymbol{X})$的估计情况。从图中可以看出，样本量越大，估计结果越精确；同时，当样本量较小时，窗宽的选择对估计结果也有一定的影响。如 $N=1$ 时，$h_1=1$ 的估计结果明显好于其他两种情况。

【例 3.5】 已知一维随机变量 x，假设待估计的概率密度函数 $p(x)$为两个均匀分布密度函数的混合，用 Parzen 窗法估计 $p(x)$，即求估计式 $\hat{p}_N(x)$。

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

【解】 仍选择正态窗函数，h_N 的定义同例 3.4，估计结果如图 3-5 所示。从图中可以得出，当 N 较小时，估计结果与真实分布相差很大，当 $N=1$ 时，$\hat{p}_N(x)$只是反映出窗函数本身，当 $h_1=0.25$，$N=16$ 时还能看到单个样本的作用，当 $h_1=1$ 和 $h_1=4$ 时，就显得比较平滑了。当 $N=16$ 时，还无法确定哪一种估计比较好，但当 $N=256$ 和 $h_1=1$ 时，估计基本满足要求；当 N 增大时，估计结果与真实分布越来越接近。

通过以上的例子可以归纳出非参数估计的优缺点。其优点是该方法具有一般性，对规则或不规则分布、单峰或多峰分布都可以用这个方法得到概率密度估计，而且只要样本足够多，总可以保证收敛于任何复杂的未知概率密度函数。其缺点是要想得到较为满意的估计结果，就需要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。而且随着样本特征维数的增加，用于估计的样本数量也需要相应地增加。

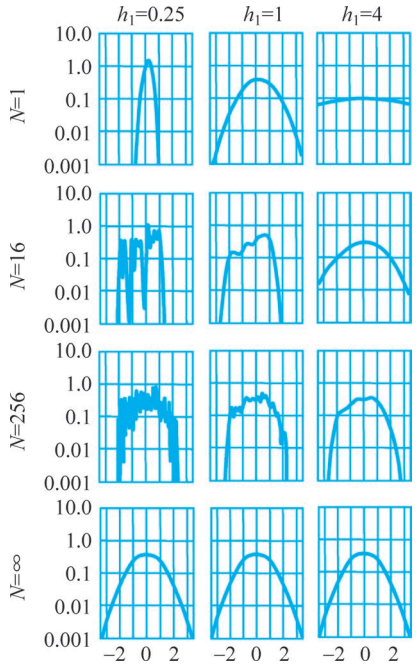


图 3-4　正态分布的 Parzen 窗估计

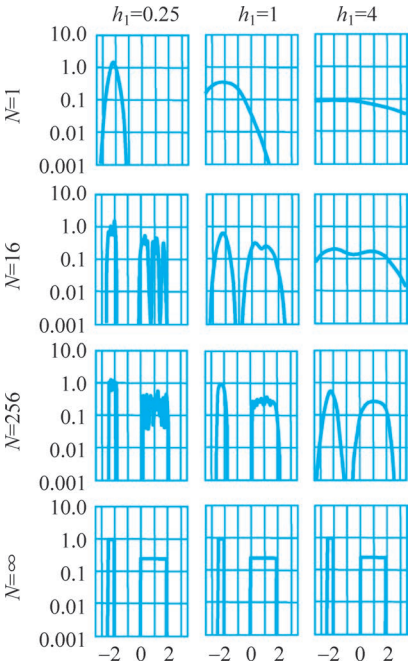


图 3-5　均匀分布的 Parzen 窗估计

3.4.3　k_N-近邻估计法

Parzen 窗估计方法是基于落入体积 V_N 中的样本来做总体估计。由于体积 V_N 中的样本是样本数 N 的函数，故无论窗函数怎样选取，Parzen 窗估计方法都难以是最佳的。此外，体积序列 $V_1, V_2, \cdots, V_N$ 的选择也是一个难题，由前面的分析，我们知道估计的结果对体积的选择比较敏感。为了解决这一问题提出了 k_N-近邻估计方法（简称 k_N-近邻法）。

k_N-近邻估计方法的基本思想是，使包含 $\boldsymbol{X}$ 点的序列体积 $V_1, V_2, \cdots, V_N$ 受落入 R_N 中的样本数 k_N 控制，而不是作为实验样本数 N 的函数。具体来说，可以预先确定 N 的一个函数 k_N，在 $\boldsymbol{X}$ 点近邻选择一个体积，并使其不断增大直到捕获 k_N 个样本为止，这 k_N 个样本就是 $\boldsymbol{X}$ 点的近邻。很显然，如果 $\boldsymbol{X}$ 点附近的概率密度较大，则包含 k_N 体积较小；反之，如果 $\boldsymbol{X}$ 点附近的概率密度较小，则包含 k_N 体积较大。k_N-近邻估计方法使用的基本估计公式仍为

$$\hat{p}_N(\boldsymbol{X}) = \frac{k_N / N}{V_N}$$

当满足式(3-40)、式(3-41)和式(3-42)三个条件时，$\hat{p}_N(\boldsymbol{X})$ 收敛于概率密度 $p(\boldsymbol{X})$。k_N-近邻估计具有以下特点。

(1) k_N 大小的选择会影响估计的结果。k_N 可以选择为样本容量 N 的某种函数，在 $k_N = k_1\sqrt{N}$，$k_N \geqslant 1$ 条件下，当样本容量 $N \to \infty$ 时，可以保证 $\hat{p}_N(\boldsymbol{X})$ 收敛于真实分布 $p(\boldsymbol{X})$。但是在有限样本容量条件下，k_1 的选择也会影响估计结果的正确性。

(2) k_N-近邻估计方法的计算量大。与 Parzen 窗法一样，为保证估计结果的正确性，所

需样本量 N 一般要很大，尤其当样本的特征维数比较高时更是如此，因此存储量和计算量都很大。经验数据表明，当样本的特征维数为一维时，所需样本容量一般为数百个。但是，当样本特征维数为二维时，样本容量就需要几千个。随着样本特征维数的增加，样本容量会急剧增长，带来超大计算量与超大存储量的问题。目前已有一些针对该问题的解决方法，有兴趣的读者可以参考相关资料。

图 3-6 所示为概率密度函数分别为正态分布和双峰均匀分布时用 k_N-近邻法估计 $p(\boldsymbol{X})$的结果。

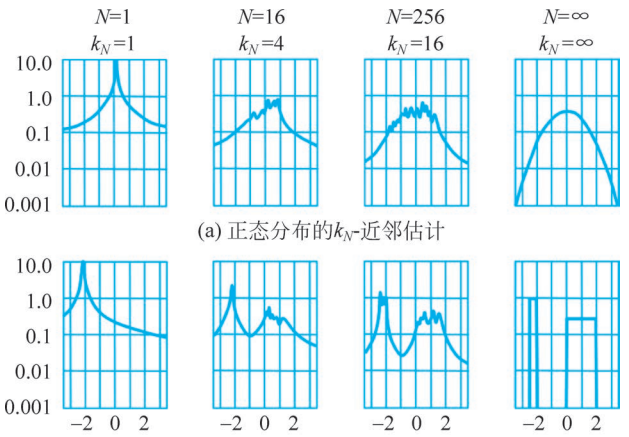


图 3-6 k_N-近邻估计 $p(\boldsymbol{X})$的结果

3.5 Python 示例

【例 3.6】 生成 100 万个服从标准正态分布的随机数，用最大似然估计这些随机数服从的正态分布的参数值。

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from scipy.stats import norm

# 生成正态分布样本
testdata = np.random.normal(loc = 0, scale = 1.0, size = 1000000)
print('样本均值的无偏估计:', np.mean(testdata))
print('样本标准差的无偏估计:', np.std(testdata, ddof = 1))
```

运行结果：

```
样本均值的无偏估计: -3.8854235939791765e-06
样本标准差的无偏估计: 1.0008638816906816
```

程序生成了 100 万个服从标准正态分布的随机数，并直接计算了它们的均值和标准差。然后，绘制出这些样本点的直方图(如图 3-7 所示)，观察其与正态分布函数的逼近程度。

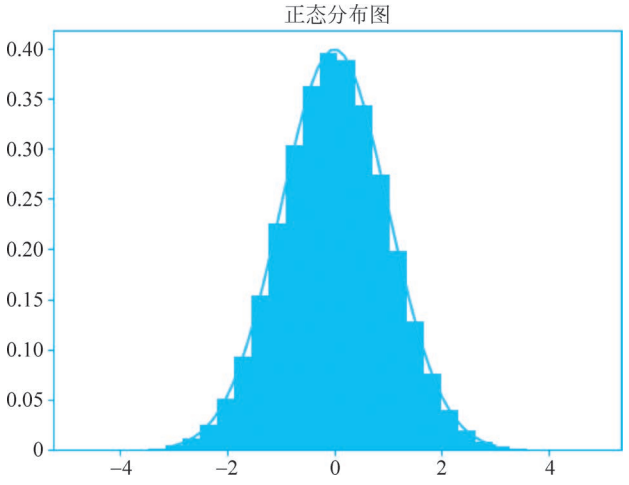


图 3-7　样本点的直方图

```
# 绘制正态分布图
plt.hist(testdata, density = True, bins = 30)   # 归一化直方图(用出现频率代替次数),将划分
                                                # 区间变为 20(默认 10)
x = np.linspace( - 3, 3, 50)                    # 在( - 3,3)返回均匀间隔的 50 个数字
plt.plot(x, norm.pdf(x), 'r - ')
plt.title('正态分布图')
plt.show()
# 正态分布均值区间估计
def confidence_interval_u(data, sigma = - 1, alpha = 0.05, side_both = True):
    xb = np.mean(data) # 均值
    # s = np.std(data, ddof = 1)                # 无偏标准差
    # sigma 已知,服从标准正态分布
    z = stats.norm(loc = 0, scale = 1.0)
    if side_both: # 求双侧置信区间
        tmp = sigma/np.sqrt(len(data)) * z.ppf(1 - alpha/2)
    else: # 单侧置信下限或单侧置信上限
        tmp = sigma/np.sqrt(len(data)) * z.ppf(1 - alpha)
    bottom_limit = xb - tmp
    top_limit = xb + tmp
    return bottom_limit, top_limit

# 正态分布标准差区间估计
def confidence_interval_sigma(data, mu = - 1, alpha = 0.05, side_both = True):
    sum_tmp = 0.0
    for i in data:
        sum_tmp = sum_tmp + (i - mu) ** 2
    if side_both: # 双侧置信区间
        bottom_limit = sum_tmp / stats.chi2.ppf(1 - alpha / 2, df = len(data))
        top_limit = sum_tmp / stats.chi2.ppf(alpha / 2, df = len(data))
    else: # 单侧置信下限或单侧置信上限
        bottom_limit = sum_tmp / stats.chi2.ppf(1 - alpha, df = len(data))
        top_limit = sum_tmp / stats.chi2.ppf(alpha, df = len(data))
    return np.sqrt(bottom_limit), np.sqrt(top_limit)

paramhat = confidence_interval_u(testdata, 1)
```

```
paramint = confidence_interval_sigma(testdata, 0)
print('样本均值的区间估计:', np.round(paramhat, 6))
print('样本标准差的区间估计:', np.round(paramint, 6))
```

运行结果：

```
样本均值的区间估计: [-0.001964 0.001956]
样本标准差的区间估计: [0.999478 1.002252]
```

程序可以获得两个结果：paramhat 和 paramint。它们分别对应均值和标准差的区间估计。从结果来看，均值和标准差的区间估计都很窄，而且很接近生成随机数时使用的参数值。

【例 3.7】 设置两类为 2×60 的样本，第一类样本是一些随机设定的 0.6～1.2 的数，第二类样本是第一类样本的第一行加 0.1，第二行减 0.1 而生成的。用贝叶斯方法分别估计这两类样本的条件概率密度的均值参数 u。其中，设两类样本的先验估计都为[0.93,0.94]，$p(u)$的方差为[0.05,0.05]。

```
import numpy as np
def bayes_estimation(class_id, X, sigma0, mu0):
    # 求样本均值
    m1, m2 = np.mean(X, axis=1)
    print(f'第{class_id}类样本均值:', [m1, m2])

    # 求解样本方差
    s1, s2 = np.var(X, axis=1, ddof=1)
    print(f'第{class_id}类样本方差:', [s1, s2])

    # 利用贝叶斯公式(3-30)
    N = 60
    mu_1 = N * sigma0[0] * m1 / (N * sigma0[0] + s1) + s1 * mu0[0] / (N * sigma0[0] + s1)
    mu_2 = N * sigma0[1] * m2 / (N * sigma0[1] + s2) + s2 * mu0[1] / (N * sigma0[1] + s2)
    print(f'第{class_id}类样本均值估计值:', [mu_1, mu_2])

# 设置方差
c = [0.05, 0.05]
# 设置先验估计参数
d = [0.93, 0.94]

# 生成第一类样本
# 生成范围在(0.6,1.2)的 2*60 的样本,并保留一位小数
X = np.random.uniform(0.6, 1.2, size=(2, 60)).round(1)
print('第一类样本数据:', X)
bayes_estimation(1, X, c, d)

运行结果:
第一类样本数据: [[0.8 0.7 1. 1. 1. 0.6 1.1 1.2 1.1 0.7 0.8 0.8 0.6 0.6 1. 1. 0.7 0.7 0.8 0.7 1.
0.7 1.1 1. 1. 0.7 1. 0.9 0.8 1. 0.8 1. 0.7 1. 1.1 1. 0.8 0.7 1.1 1.2 0.8 0.9 0.9 0.9 1.1 0.9 0.8
0.8 0.6 1.2 0.9 1. 0.8 0.6 1.1 0.7 1. 0.7 1.1 0.8]
[1.1 1.1 0.7 1.1 0.9 0.9 0.6 0.8 1.1 0.9 1. 0.7 0.9 0.6 0.9 1. 1.1 0.8 1. 1.2 0.8 0.6 1.2 1. 0.6
1.1 1. 0.6 1. 1. 0.9 1. 0.8 1.1 0.8 0.9 1.1 0.8 0.9 1.1 0.7 1. 1. 1.1 0.6 1. 0.6 1.1 1. 1.1 0.8
0.8 0.7 1.2 1. 1.1 0.9 0.7 1.2 0.8]]
```

```
第 1 类样本均值: [0.885, 0.9183333333333333]
第 1 类样本方差: [0.029432203389830512, 0.03237005649717515]
第 1 类样本均值估计值: [0.8854371938579977, 0.918564621471337]

# 生成第二类样本
x1 = X[0, :] + 0.1                # 生成第一行数据
x2 = X[1, :] - 0.1                # 生成第二行数据
X2 = np.vstack((x1, x2))          # 将两行数据进行垂直拼接
# print('第二类样本数据:', X2)
bayes_estimation(2, X2, c, d)

运行结果:
第 2 类样本均值: [0.985, 0.8183333333333335]
第 2 类样本方差: [0.029432203389830533, 0.032370056497175136]
第 2 类样本均值估计值: [0.9844656519513362, 0.8196321051852005]
```

【例 3.8】 产生 1、16、256 和 16384 个服从一维标准正态分布的样本。

(1) 用窗宽 h_1 分别为 0.25、1、4，$h_N=h_1/\sqrt{N}$，窗函数为高斯函数的情形估计所给样本的密度函数并画出图形。

(2) 当 $k_N=\sqrt{N}$ 时，用 k_N-近邻法估计所给样本的密度函数并画出图形。

【解】 Python 代码及运行结果如下。

(1) 样本量 N 分别取 1、16、256、16384，h_1 分别取 0.25、1、4 时采用 Parzen 窗法的估计，结果如图 3-8 所示。

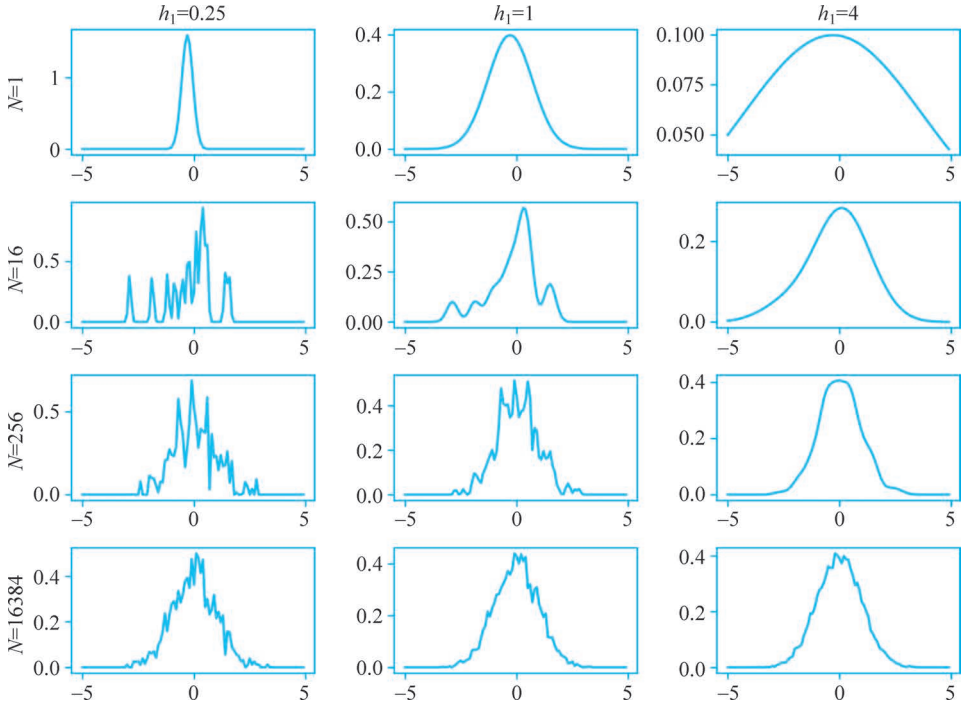


图 3-8　用 Parzen 窗法估计的结果

```
import numpy as np
import math
import matplotlib.pyplot as plt

def Paezen_window(u):                                    # 正态分布窗函数
    s = np.exp( - (u * u) / 2.0) / math.sqrt(2 * math.pi)
    return s

def Parzen(N, x, h1):
    n = len(N)
    hn = h1 / (math.sqrt(n))
    # hn = h1/(math.log(n) + 1)
    Px = []
    for i in x:
        p = Paezen_window((N - i) / hn)
        Px.append(np.sum(p) / (h1 * math.sqrt(n)))     # 概率密度公式
    return np.array(Px)

x = np.arange( - 5, 5, 0.1).reshape([ - 1, 1])
N_number = [1, 16, 256, 16384]
h1 = [0.25, 1, 4]
result = []

for i in range(4):
    N_i = np.random.normal(0, 1, N_number[i]).reshape([ - 1, 1])
    for k in range(0, 3):
        result_i = Parzen(N_i, x, h1[k])
        result.append(result_i)

plt.figure(figsize = (8, 6), dpi = 300)
for i in range(1, 13):
    plt.subplot(4, 3, i)
    plt.plot(x, result[i - 1])
    if i % 3 == 1:
        plt.ylabel('N = {0}'.format(N_number[(i - 1)//3]))
    if i <= 3:
        plt.title('h1 = {}'.format(h1[i - 1]))
plt.tight_layout()
plt.show()
```

(2) 样本量 N 分别取 1、16、256、16384 时采用 k_N-近邻法估计，结果如图 3-9 所示。

```
from scipy.spatial.distance import cdist
import numpy as np
import math
import matplotlib.pyplot as plt

def knn_estimate(n, x, d):
    # n 随机样本, x 横坐标,d 特征维度
    Kn = int(math.sqrt(len(n)))
    Px = []
    q = Kn/len(n)
    for i in x:
        hn = cdist(np.array([i]), n, metric = 'euclidean').reshape( - 1)
```

```
            hn.sort()
            if Kn == 1:
                Vn = (hn[Kn-1] * 2) ** d
            else:
                Vn = (hn[Kn] * 2) ** d
            Px.append(q/Vn)   # 计算公式
        return np.array(Px)

X = np.arange(-4, 4, 0.005).reshape([-1, 1])   # reshape([-1,1])将矩阵转换为一列

result = []
N_number = [1, 16, 256, 16384]

for i in range(4):
    N_i = np.random.normal(0, 1, N_number[i]).reshape([-1, 1])
    result_i = knn_estimate(N_i, X, d=1)
    result.append(result_i)

fig, axes = plt.subplots(2, 2, figsize=(8, 6), dpi=200)
ax = axes.ravel()
for i in range(0, 4):
    ax[i].plot(X, result[i])
    ax[i].set_title('N={0}\nKn={1}'.format(
        N_number[i], math.sqrt(N_number[i])))
plt.tight_layout()
plt.show()
```

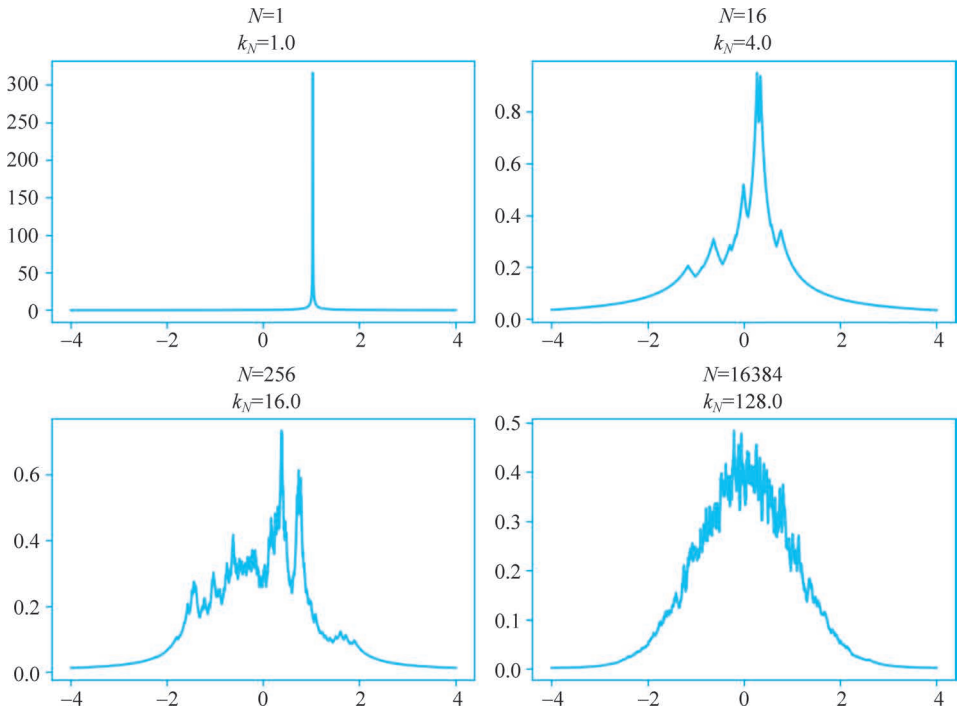


图 3-9　用 k_N-近邻法估计的结果

在实际应用中，类条件概率密度通常是未知的。那么，在先验概率和类条件概率密度都未知或者其中之一未知的情况下，该如何来进行类别判断呢？其实，只要能收集到一定数量的样本，根据统计学的知识，我们是可以从样本集来推断总体概率分布的。

监督参数估计就是由已知类别的样本集对总体分布的某些参数进行统计推断；非监督参数估计是已知总体概率密度函数形式但未知样本所属的类别，要求推断出概率密度函数的某些参数。通常采用最大似然估计方法和贝叶斯估计方法。最大似然估计把参数看成确定(非随机)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；而贝叶斯估计则是把参数当成具有某种分布的随机变量，样本的观察结果使先验分布转换为后验分布，再根据后验分布修正原先对参数的估计。非参数估计是已知样本所属的类别，但未知总体概率密度函数的形式，要求我们直接推断概率密度函数本身。统计学中常见的一些典型分布形式不总是能够拟合实际中的分布。此外，在许多实际问题中经常遇到多峰分布的情况，这就迫使我们必须用样本来推断总体分布，常见的总体类条件概率密度估计方法有 Parzen 窗法和 k_N-近邻法两种。

习题及思考题

3.1 证明：按照贝叶斯决策理论进行分类时，其结果满足分类错误率最小。

3.2 一元正态分布的最大似然估计：假设样本 $\boldsymbol{X}$ 服从正态分布 $N(\mu,\sigma^2)$，已获得一组样本 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N$，用 Python 语言设计一程序片段，计算估计参数 μ,σ^2。

3.3 简述参数估计、非参数估计和非参数分类器等概念间的关系。

3.4 证明：对正态总体的期望 μ 的最大似然估计是无偏的，对方差 σ^2 的最大似然估计是有偏的。

3.5 设样本 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N$ 是来自 $p(\boldsymbol{X}|\theta)$的随机样本，其中 $0\leqslant x\leqslant\theta$ 时，$p(\boldsymbol{X}|\theta)=\frac{1}{\theta}$，否则为 0。证明$\theta$ 的最大似然估计是 $\max\boldsymbol{X}_k$。

3.6 设 $p(\boldsymbol{X})\sim N(\mu,\sigma^2)$，窗函数 $\varphi(\boldsymbol{X})\sim N(0,1)$，指出 Parzen 窗估计 $\hat{p}_N(\boldsymbol{X})=\frac{1}{Nh_N}\sum_{i=1}^{N}\varphi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i}{h_N}\right)$，对于小的 h_N 有如下性质。

(1) $E[\hat{p}_N(\boldsymbol{X})]\sim N(\mu,\sigma^2+h_N^2)$。

(2) $\mathrm{var}[\hat{p}_N(\boldsymbol{X})]=\frac{1}{Nh_N2\sqrt{\pi}}p(\boldsymbol{X})$。

3.7 设总体 $\boldsymbol{X}$ 的概率密度函数为 $f(\boldsymbol{X},\theta)=(\theta\alpha)\boldsymbol{X}^{\alpha-1}\mathrm{e}^{-\theta x^{\alpha}}$，求参数$\theta$ 的最大似然估计。

3.8 在掷硬币的游戏实验中，正面出现的概率是 q，反面出现的概率是 $1-q$。设 $\boldsymbol{X}_i$，$i=1,2,\cdots,N$ 是这个实验的结果，$\boldsymbol{X}_i\in(0,1)$。

(1) 证明 q 的最大似然估计是 $q_{\mathrm{ML}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i$。

提示：似然函数是 $p(\boldsymbol{X},q)=\prod_{i=1}^{N}q^{\boldsymbol{X}_i}(1-q)^{(1-\boldsymbol{X}_i)}$。

（2）证明最大似然估计结果是下列方程的解：

$$q^{\boldsymbol{\Sigma}_i\boldsymbol{X}_i}(1-q)^{(N-\boldsymbol{\Sigma}_i\boldsymbol{X}_i)}\left(\frac{\boldsymbol{\Sigma}_i\boldsymbol{X}_i}{q}-\frac{N-\boldsymbol{\Sigma}_i\boldsymbol{X}_i}{1-q}\right)=0$$

3.9　证明：对于对数正态分布 $p(\boldsymbol{X})=\frac{1}{\sigma\boldsymbol{X}\sqrt{2\pi}}\exp\left[-\frac{(\ln\boldsymbol{x}-\boldsymbol{\theta})^2}{2\sigma^2}\right]$，$\boldsymbol{X}>0$，最大似然估计 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}=\frac{1}{N}\sum_{k=1}^{N}\ln\boldsymbol{X}_k$。