第七章　非参数密度估计

概率分布是统计推断的核心,从某种意义上看, 联合概率密度提供了关于所要分析变量的全部信息, 有了联合密度, 则可以回答变量子集之间的任何问题. 从广义上看, 参数估计是在假定数据总体密度形式下对参数的估计, 比如：我们所熟知的$\bar{X}$是两点分布中p的一致性估计, $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$是一元正态总体方差的极大似然估计等. 而$\boldsymbol{X}_{n\times p}\hat{\boldsymbol{B}}_{p\times q} = \boldsymbol{X}_{n\times p}(\boldsymbol{X}'\boldsymbol{X})^{-1}_{p\times p}\boldsymbol{X}'\boldsymbol{Y}_{n\times q}$是多元正态分布均值的最小二乘估计等. 一旦参数确定, 则分布完全确定, 因而可以说参数统计推断的核心内容就是对密度的估计. 实际中, 很多数据的分布是无法事先假定的, 加上决策的可靠性要求不断提高, 因此需要适应性更广的密度估计方法. 最近几年尤其是随着数据库的广泛应用和数据挖掘技术的兴起, 概率密度估计成为模式分类技术的重要内容得到广泛关注.

§7.1　直方图密度估计

§7.1.1　基本概念

在基础的统计课程中，直方图经常用来描述数据的频率，使研究者对所研究的数据有一个较好的理解。这里，我们介绍如何使用直方图估计一个随机变量的密度。直方图密度估计与用直方图估计频率的差别在于，在直方图密度估计中，我们需要对频率估计进行归一化，使其成为一个密度函数的估计。直方图是最基本的非参数密度估计方法，有着广泛的应用。

以一元为例, 假定有数据$x_1, x_2, \cdots, x_n \in [a, b)$. 对区间$[a, b)$做如下划分，即$a = a_0 < a_1 < a_2 < \ldots < a_k = b$, $I_i = [a_{i-1}, a_i)$，$i = 1, \ldots, k$。我们有$\cup_{i=1}^{k} I_i = [a, b)$, $I_i \cap I_j = \emptyset, i \neq j$. 令$n_i = \#\{x_i \in I_i\}$ 为落在I_i中数据的个数。

我们如下定义直方图密度估计,

$$\hat{p}(x) = \begin{cases} \dfrac{n_i}{n(a_i - a_{i-1})}, & \text{当}x \in I_i; \\ 0, & \text{当}x \notin [a, b) \end{cases}$$

在实际操作中，我们经常取相同的区间，即$I_i (i = 1, 2, \ldots, k)$的宽度均为$h$, 在此情况下，我们有

$$\hat{p}(x) = \begin{cases} \dfrac{n_i}{nh}, & \text{当}x \in I_i; \\ 0, & \text{当}x \notin [a, b) \end{cases}$$

上式中, h既是归一化参数, 又表示每一组的组距, 称为带宽或窗宽. 另外，我们可以看到

$$\int_a^b \widehat{p}(x)dx = \sum_{i=1}^{k} \int_{I_i} n_i/(nh)dx = \sum_{i=1}^{k} n_i/n = 1$$

由于位于同一组内所有点的直方图密度估计均相等, 因而直方图所对应的分布函数$\hat{F}_h(x)$是单调增的阶梯函数. 这与经验分布函数形状类似. 实际上, 当分组间隔h缩小到每组中最多只有一个数据时, 直方图的分布函数就是经验分布函数, 即$h \to 0$,有$\hat{F}_h(x) \to \hat{F}_n(x)$.

定理7.1 固定x和h,令估计的密度是$\widehat{p}(x)$, 如果$x \in I_j, p_j = \int_{I_j} \widehat{p}(x)\ \mathrm{d}x$, 有

$$E\hat{p}(x) = p_j/h, \quad \mathrm{var}\hat{p}(x) = \frac{p_j(1-p_j)}{nh^2}.$$

证明提示：注意到$E\hat{p}_j = n_j/n = \int_{I_j} \hat{p}(x)\ \mathrm{d}x, \mathrm{var}\hat{p}_j = p_j(1-p_j)/n.$

例7.1(见chap7 数据fish.txt) 给出了鲑鱼和鲈鱼两种鱼类长度的观测数据, 共计230条. 在图7.1中, 我们从左到右, 分别采用逐渐增加的带宽间隔:$h_l = 0.75, h_m = 4, h_r = 10$制作了3个直方图. 可以发现当带宽很小的时候, 个体特征比较明显, 从图中可以看到多个峰值; 而带宽过大的最右边的图上, 很多峰都不明显了. 中间的图比较合适, 它有两个主要的峰, 提供了最为重要的特征信息. 实际上, 参与直方图运算的是鲑鱼和鲈鱼两种鱼类长度的混合数据, 经验表明, 大部分鲈鱼具有身长比鲑鱼长的特点, 因而两个峰是合适的. 这也说明直方图的技巧在于确定组距和组数, 组数过多或过少, 都会淹没主要特征. R程序如下：

```
fish=read.table("......../fish.txt", header=T)
length=fish[,1]
par(mfrow=c(1,3))
hist(length,breaks=0:35*0.75, freq=F, xlab="bodysize", main="Bandwidth=0.75")
hist(length,breaks=0:7*4, freq=F, xlab="bodysize", main="Bandwidth=4")
hist(length,breaks=0:3*10, freq=F,  xlab="bodysize", main="Bandwidth=10")
```

§7.1.2 理论性质和最优带宽

由上面的例子，我们可以看出，选择不同的带宽，我们会得到不同的结果。选择合适的带宽，对于得到好的密度估计是很重要的。在计算最优带宽前，我们先定义$\hat{p}$的平方损失风险$R(\hat{p}, p) = \int (\hat{p}(x) - p(x))^2\ \mathrm{d}x$

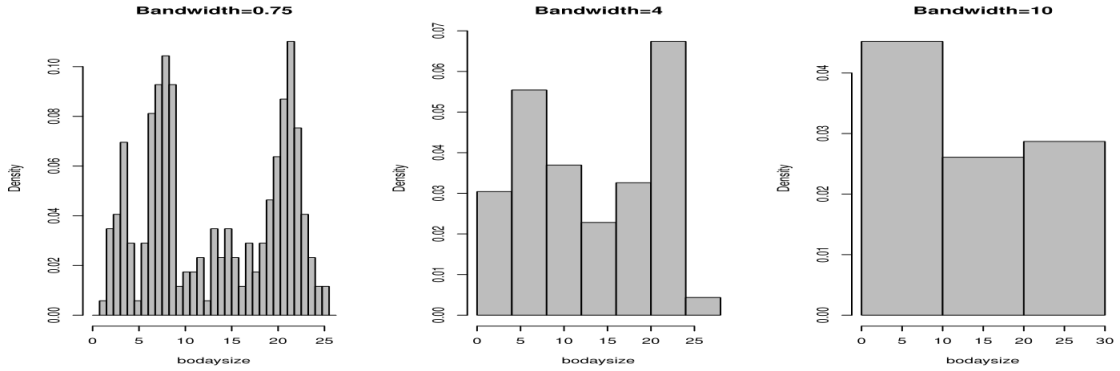

图 7.1 鲑鱼和鲈鱼身长(bodysize)直方图

定理7.2 $\int p'(x)\,\mathrm{d}x < +\infty$,则在平方损失风险下,

$$R(\hat{p},p) \approx \frac{h^2}{12}\int (p'(u))^2\,\mathrm{d}u + \frac{1}{nh}.$$

极小化上式, 得到理想带宽为

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int p'(x)^2\,\mathrm{d}x}\right)^{1/3}.$$

于是理想的带宽为$h = Cn^{-1/3}$.

证明 考察平方损失风险:

$$\begin{aligned} R(\hat{p},p) &= EL(\hat{p}(x),p(x)) \\ &= E\int (\hat{p}(x)-p(x))^2\,\mathrm{d}x \\ &= \int (E\hat{p}(x)-p(x))^2\,\mathrm{d}x + E\int (\hat{p}(x)-E\hat{p}(x))^2\,\mathrm{d}x \\ &= \int \mathrm{Bias}^2(x)\,\mathrm{d}x + \int V(x)\,\mathrm{d}x. \end{aligned}$$

风险分解为两项: 偏差项和方差项. 偏差项用于评价估计量对真实函数估计的精准度, 方差项用于测量估计量本身的波动大小.

先看第一项偏差项：

$$\begin{aligned}\mathrm{Bias}(x)=&E\hat{p}(x)-p(x)=\frac{p_j}{h}-p(x)\\=&\frac{p(x)h+hp'(x)(h/2-x)}{h}-p(x)\\=&p'(x)(h/2-x).\end{aligned}$$

注意到

$$\begin{aligned}\int_{I_j}\mathrm{Bias}^2(x)\mathrm{d}x=&\int_{I_j}\left(p'(x)\right)^2\left(h/2-x\right)^2\ \mathrm{d}x\\\approx&\left(p'(\xi_j)\right)^2\frac{h^3}{12},\end{aligned}$$

于是

$$\begin{aligned}\int\mathrm{Bias}^2(x)\mathrm{d}x=&\sum_{j=1}^m\int_{I_j}\mathrm{Bias}^2(x)\\\approx&\sum_{j=1}^m p'(\xi)^2\frac{h^3}{12}\\\approx&\frac{h^2}{12}\int p'(x)^2\mathrm{d}x.\end{aligned}$$

再看第二项方差项：

$$\begin{aligned}V(x)\approx&\frac{p_j}{nh^2}\\=&\frac{p(x)h+hp'(x)(h/2-x)}{nh^2}\\\approx&p(x)/nh.\end{aligned}$$

一般当h未知的时候, 可以用更实用的方式选择窗宽,

$$\begin{aligned}R(h)=&\int(\hat{p}-p(x))^2\mathrm{d}x\\=&\int\hat{p}^2\mathrm{d}x-2\int\hat{p}p\mathrm{d}x+\int p^2(x)\mathrm{d}x\\=&J(h)+\int p^2(x)\mathrm{d}x.\end{aligned}$$

注意到后面一项与h无关, 第一项可以用交叉验证方法估计：

$$\hat{J}(h)=\int(\hat{p})^2\mathrm{d}x-\frac{2}{n}\sum_{i=1}\hat{p}_{(-i)}(x_i).$$

其中, $\hat{p}_{(-i)}(x_i)$是去掉第i个观测值后对直方图的估计, $\hat{J}(h)$称为交叉验证得分. 证毕(Scott D.W. 2009)。

在大多数情况下，我们不知道密度$p(x)$，因此也不知道$p'(x)$。对于理想带宽$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int p'(x)^2\,\mathrm{d}x}\right)^{1/3}$也无法计算，在实际操作中，经常假设$p(x)$为一个标准正态分布，并进而得到一个带宽$h_0 \approx 3.5n^{-1/3}$.

直方图密度估计的优势在于简单易懂，在计算过程中也不涉及到复杂的模型计算，只需要计算I_j中样本点的个数。另一方面，直方图密度估计只能给出一个阶梯函数，该估计不够光滑。另外一个问题是直方图密度估计的收敛速度比较慢，也就是说，$\hat{p}(x) \longrightarrow p(x)$比较慢。

§7.1.3 多维直方图

直方图的密度定义公式很容易扩展到任意维空间. 设有n个观测点$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$, 将空间分成若干小区域$R$, V是区域R所包含的体积. 如果有k个点落入R, 则可以得到如下密度公式：$p(\boldsymbol{x})$的估计为

$$p(\boldsymbol{x}) \approx \frac{k/n}{V}. \tag{7.1}$$

如果这个体积和所有的样本体积相比很小, 就会得到一个很不稳定的估计, 这时, 密度值局部变化很大, 呈现多峰不稳定的特点; 反之, 如果这个体积太大, 则会圈进大量样本, 从而使估计过于平滑. 在稳定与过度光滑之间寻找平衡就引导出下面两种可能的解决方法.

(1) 固定体积V不变, 它与样本总数呈反比关系即可. 注意到, 在直方图密度估计中, 每一点的密度估计只与它是否属于某个I_i有关, 而I_i是预先给定的与该点无关的区域. 不仅如此, 区域I_i中每个点共有相等的密度, 这相当于待估点的密度取邻域R的平均密度. 现在以待估点为中心, 作体积为V的邻域, 令该点的密度估计与纳入该邻域中的样本点的多少呈正比, 如果纳入的点多, 则取密度大, 反之亦然. 这一点还可以进一步扩展开去, 将密度估计不再局限于R内的带内, 而是将体积V合理拆分到所有样本点对待估计点贡献的加权平均, 同时保证距离远的点取较小的权, 距离近的点取较大的权, 这样就形成了核函数密度估计法的基本思想. 后面我们将看到, 这些方法都可能获得较为稳健而适度光滑的估计.

(2) 固定k值不变, 它与样本总数呈一定关系即可. 根据数据之间的疏密情况调整V, 这样就导致了另外一种密度估计方法——k近邻法.

下面介绍核估计和k近邻估计两种非参数方法.

§7.2 核密度估计

§7.2.1 核函数的基本概念

在上节中，我们介绍了直方图密度估计。但是通过直方图得到密度估计不是一个光滑函数。为了克服这个缺点，我们介绍核函数密度估计。核函数密度估计有着广泛的应用，其理论性质也已经得到了很好的研究。这里我们首先介绍一维的情况。

定义7.1 假设数据$x_1, x_2, \cdots, x_n$取自连续分布$p(x)$, 在任意点x处的一种核密度估计定义为

$$\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n}\omega_i = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right), \tag{7.2}$$

其中$K(\cdot)$称为核函数(kernel function). 为保证$\hat{p}(\boldsymbol{x})$作为概率密度函数的合理性, 既要保证其值非负, 又要保证积分的结果为1. 这一点可以通过要求核函数$K(\boldsymbol{x})$是分布密度得到保证, 即

$$K(x) \geqslant 0, \quad \int K(x)\ \mathrm{d}x = 1.$$

实际上有

$$\begin{aligned}
&\int \hat{p}(x)\mathrm{d}x \\
=&\int \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\ \mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h}K\left(\frac{x-x_i}{h}\right)\ \mathrm{d}x \\
=&\frac{1}{n}\sum_{i=1}^{n}\int K(u)\ \mathrm{d}u = \frac{1}{n}\cdot n = 1 \ \left(\text{其中 } u = \frac{x-x_i}{h}\right).
\end{aligned} \tag{7.3}$$

由$\int \widehat{p}(x)dx = 1$可知，上述定义的$\widehat{p}(x)$是一个合理的密度估计函数。

核密度估计中，一个重要的部分就是核函数。以一维为例, 常用的核函数如表7.1所示：

表7.1 常用核函数

核函数名称	核函数$K(u)$
Parzen窗(Uniform)	$\frac{1}{2}I(\|u\| \leqslant 1)$
三角(Triangle)	$(1-\|u\|)I(\|u\| \leqslant 1)$
Epanechikov	$\frac{3}{4}(1-u^2)^2 I(\|u\| \leqslant 1)$
四次(Quartic)	$\frac{15}{16}(1-u^2)^2 I(\|u\| \leqslant 1)$
三权(Triweight)	$\frac{35}{32}(1-u^2)^3 I(\|u\| \leqslant 1)$
高斯(Gauss)	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$
余弦(Cosinus)	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) I(\|u\| \leqslant 1)$
指数(Exponent)	$\exp\{-\|u\|\}$

表7.1中不同的核函数表达了根据距离分配各个样本点对密度贡献的不同情况.

例7.2(例7.1续) 图7.2给出了各种带宽之下根据正态核函数做出的密度估计曲线. 由图可知, 带宽$h=10$是最平滑的(右边), 相反带宽$h=1$噪声很多, 它在密度中引入了很多虚假的波形. 从图中比较, 带宽$h=5$是较为理想的, 它在不稳定和过于平滑之间作了较好的折中.

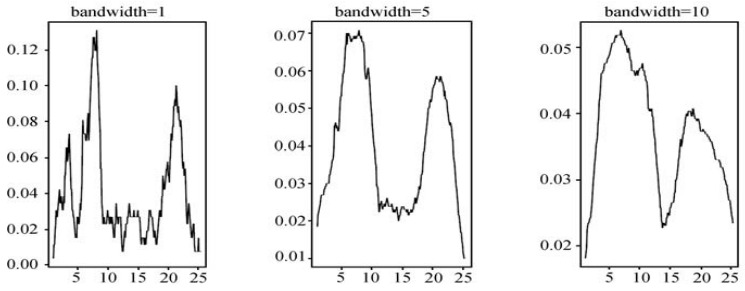

图 7.2 鲑鱼和鲈鱼身长的密度核估计

```
plot(density(length,kernel="gaussian",bw=1), main="Bandwidth=1")
plot(density(length,kernel="gaussian",bw=2), main="Bandwidth=2")
plot(density(length,kernel="gaussian",bw=8), main="Bandwidth=8")
```

§7.2.2 理论性质和带宽

核函数的形状通常不是密度估计中最关键的因素, 和直方图一样, 带宽对模型

光滑程度的影响作用较大. 因为如果h非常大, 将有更多的点对x处的密度产生影响. 由于分布是归一化的, 即

$$\int \omega_i(x-x_i)\mathrm{d}x = \int \frac{1}{h}K\left(\frac{x-x_i}{h}\right)\mathrm{d}x = \int K(u)\ \mathrm{d}u = 1,$$

因而距离x_i较远的点也分担了对x的部分权重, 从而较近的点的权重ω_i 减弱, 距离远和距离近的点的权重相差不大. 在这种情况下, $\hat{p}(x)$是n个变化幅度不大的函数的叠加, 因此$\hat{p}(x)$非常平滑; 反之, 如果h很小, 则各点之间的权重由于距离的影响而出现大的落差, 因而$\hat{p}(x)$是n个以样本点为中心的尖脉冲的叠加, 就好像是一个充满噪声的估计.

如何选择合适的带宽, 是核函数密度估计能够成功应用的关键. 类似于定性数据联合分布的误差平方和的分解, 理论上选择最优带宽也是从密度估计与真实密度之间的误差开始的.

对于每个固定的x，我们可以使用均方误(Mean Squared Error, MSE)。均方误可以分解为两个部分

$$\begin{aligned} &MSE(x;h) = E\left[\widehat{p}(x)-p(x)\right]^2 \\ &= \left[E\widehat{p}(x)-p(x)\right]^2 + E\left[\widehat{p}(x)-E\widehat{p}(x)\right]^2 \\ &= Bias(x)^2 + V(x) \end{aligned}$$

其中，$Bias(x) = E\widehat{p}(x)-p(x)$，$V(x) = E\left[\widehat{p}(x)-E\widehat{p}(x)\right]^2$

这里由于分布密度是连续的, 因而通常考虑估计的积分均方误(mean integral square error, MISE), 如下定义：

$$\mathrm{MISE} = E\left[\int (\widehat{p}(x)-p(x))^2\ \mathrm{d}x\right] = E(\widehat{p}(x)-p(x))^2$$

考虑大样本的渐近积分均方误(asymptotic integral mean square error, AMISE), 它可以分解为两部分：

$$\mathrm{AMISE} = \int [(\mathrm{Bias}(x))^2 + \mathrm{var}(x)]\ \mathrm{d}x.$$

等式右边分别为积分偏差平方(以下简称偏差)与方差.

与直方图类似, 也可以得到大样本情况下核估计的如下一些基本结论.

我们先来估计$Bias(\widehat{p})$, 首先，令$(x-x_i)/h = t$ and $x_i = x-ht$，计算可得

$$\begin{aligned} \int h^{-1}K\left(\frac{x-x_i}{h}\right)p(x_i)dx_i \quad &= \int h^{-1}K(u)p(x-ht)d(x-ht) \\ &= \int h^{-1}K(u)p(x-ht)|-h|dt \\ &= \int K(u)p(x-ht)dt \end{aligned}$$

使用泰勒展开$p(x-ht)-p(x)=-htp'(x)+\frac{1}{2}h^2t^2p''(x)+O(h^3)$ 因此，我们得到

$$\begin{aligned}&\int h^{-1}K\left(\frac{x-x_i}{h}\right)p(x_i)dx_i-p(x)\\&=\int K(u)\left[p(x-ht)-p(x)\right]dt\\&=-hp'(x)\int tK(t)dt+\frac{1}{2}h^2p^{(2)}(x)\int t^2K(t)dt+O(h^3)\\&=\frac{h^2}{2}\mu_2(K)p^{(2)}(x)+O(h^3)\end{aligned}$$

其中$\mu_2(K)=\int t^2K(t)dt.$

定理7.3 假设$\hat{p}(x)$定义如式(7.2), 是$p(x)$的核估计, 令$\text{supp}(p)=\{x:p(x)>0\}$是密度$p$的支撑. 设$x\in\text{supp}(p)\subset\mathbb{R}$为$\text{supp}(p)$的内点(非边界点), 当$n\to+\infty$时, $h\to0$, $nh\to+\infty$, 核估计有如下性质：

$$\begin{aligned}\text{Bias}(x)&=\frac{h^2}{2}\mu_2(K)p^{(2)}(x)+O(h^2);\\\text{V}(x)&=(nh)^{-1}p(x)R(K)+O((nh)^{-1})+O(n^{-1}).\end{aligned}$$

若$\sqrt{(nh)}\ h^2\longrightarrow0$,则

$$\sqrt{(nh)}(\hat{p}_n(x)-p(x))\longrightarrow N(0,p(x)R(K)).$$

其中$R(K)=\int K(x)^2\mathrm{d}x.$

从均方误的偏差和方差分解来看, 带宽h越小, 核估计的偏差越小, 但核估计的方差越大; 反之, 带宽h增大, 则核估计的方差变小, 但核估计偏差却增大. 所以, 带宽h的变化不可能一方面使核估计的偏差减小, 同时又使核估计的方差减小. 因而, 最佳带宽选择的标准必须在核估计的偏差和方差之间作一个权衡, 使积分均方误达最小. 实际上，由定理7.3，我们可以得到渐近积分均方误(AMISE) $\frac{h^4}{4}\mu_2^2\int p^{(2)}(x)^2dx+n^{-1}h^{-1}\int K(x)^2dx$ 由此可知，最优带宽为

$$h_{opt}=\mu_2(K)^{-4/5}\left[\int K(x)^2dx\right]^{1/5}\left[\int p^{(2)}(x)^2dx\right]^{-1/5}n^{-1/5}$$

对于上式中的最优带宽，核函数$K(u)$是已知的，但是密度函数$p(x)$是未知的。在实际操作中，我们经常把$p(x)$看成正态分布去求解，即$\int p^{(2)}(x)^2dx=\frac{3}{8}\pi^{-1/2}\sigma^{-5}$, 这样，对于不同的核函数，我们可以得到相应的最优带宽。例如当核函数是高斯

时，我们可以得到$\mu_2 = 1, \int K(u)^2 du = \int \frac{1}{2\pi}\exp(-u^2)du = \pi^{-1/2}$，这样，最优带宽就是$h_{opt} = 1.06\sigma n^{-1/5}$.

除了上述的方法，从实际计算的角度, 鲁德默(Rudemo,1982)和鲍曼(Bowman,1984)提出用交叉验证法确定最终带宽的递推方法. 具体来说, 考虑积分平方误

$$\mathrm{ISE}(h) = \int (\hat{p}(x) - p(x))^2 \mathrm{d}x = \int \hat{p}^2 \ \mathrm{d}x + \int p^2 \ \mathrm{d}x - 2\int \hat{p}p \ \mathrm{d}x \tag{7.4}$$

达到最小, 将右边展开, 因此这等价于最小化式：

$$\mathrm{ISE}(h)_{\mathrm{opt}} = \int \hat{p}^2 \mathrm{d}x - 2\int \hat{p}p \ \mathrm{d}x. \tag{7.5}$$

注意到等式的第二项为$\int \hat{p}p \, \mathrm{d}x = E(\hat{p})$, 因此, 可以用$\int \hat{p}p \, \mathrm{d}x$的一个无偏估计$n^{-1}\sum_{i=1}^{n} \hat{p}_{-i}(X_i)$来估计, 其中$\hat{p}_{-i}$是将第$i$个观测点剔除后的概率密度估计. 下面只要估计第一项即可. 将核估计定义式代入第一项, 不难验证：

$$\begin{aligned}\int \hat{p}^2 \ \mathrm{d}x &= n^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_x K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right) \ \mathrm{d}x \\ &= n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_t K\left(\frac{X_i - X_j}{h} - t\right) K(t) \ \mathrm{d}t,\end{aligned}$$

于是, $\int \hat{p}^2 \ \mathrm{d}x$可用$n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} K\cdot K\left(\frac{X_i - X_j}{h}\right)$ 估计, 其中$K\cdot K(u) = \int_t K(u-t)K(t)\mathrm{d}t$是卷积. 所以, 鲁德默(Rudemo)和鲍曼(Bowman)提出的交叉验证法(cross validation)实际上是选择h使下一步

$$\mathrm{ISE}(h)_1 = n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} K\cdot K\left(\frac{X_i - X_j}{h}\right) - 2n^{-1}\sum_{i=1}^{n}\hat{p}_{-i}(X_i) \tag{7.6}$$

达到最小. 当K是标准正态密度函数时, $K\cdot K$是$N(0,2)$密度函数, 有

$$\begin{aligned}\mathrm{ISE}(h)_1 = &\frac{1}{2\sqrt{\pi}n^2 h}\sum_i\sum_j \exp\left[-\frac{1}{4}\left(\frac{X_i - X_j}{h}\right)^2\right] \\ &-\frac{2}{\sqrt{2\pi}n(n-1)h}\sum_i\sum_{j\neq i}\exp\left[-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2\right].\end{aligned}$$

§7.2.3 置信带和中心极限定理

首先，对于单点x而言，令$s_n(x) = \sqrt{\mathrm{Var}(\hat{p}_h(x))}$. $p_h(x) = \mathrm{E}(\hat{p}_h(x))$.有中心极限定理：

$$Z_n(x) = \frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} \to h \to 0N(0, \tau^2(x)).$$

值得注意的是，上述的中心极限定理只能对$p_h(x)$产生一个近似的置信区间估计，不能对$p(x)$产生置信区间估计。注意到

$$\frac{\hat{p}_h(x)-p(x)}{s_n(x)}=\frac{\hat{p}_h(x)-p_h(x)}{s_n(x)}+\frac{p_h(x)-p(x)}{s_n(x)}.$$

上式的第一项是一个近似标准正态统计量。第二项是偏差和标准差之比，这一项有下面的收敛：

$$\frac{\hat{p}_h(x)-p(x)}{s_n(x)}\to N(c,\tau^2(x).)$$

其中c不为0.这表示置信区间$\hat{p}_h(x)\pm z_{\alpha/2}s(x)$不会以概率$1-\alpha$覆盖$p(x)$.

如果对于多个点而言，求置信区间的方法可以使用Bootstrap方法，它的计算算法如下：

核密度的Bootstrap算法

---

(1) 从经验分布$\hat{F}_n$中重抽样 $X_1^*,X_2^*,\cdots,X_n^*$, 经验分布在每个样本点上的概率密度为 $1/n$;

(2) 基于Bootstrap样本 $X_1^*,X_2^*,\cdots,X_n^*$抽样计算 $\hat{p}_h^*$;

(3) 计算$R=\sup_x\sqrt{nh}||\hat{p}_h^*-\hat{p}_h||_\infty$.

(4) 重复步骤(1)(2)(3)共 B 次, 得到 $R_1,R_2,\cdots,R_B$;

(5) 令a_α是$\{R_j,j=1,\cdots,B\}$的α分位数

$$\frac{1}{B}\sum_{j=1}^{B}I(R_j>z_\alpha)\approx\alpha.$$

(6) 令

$$l_n(x)=\hat{p}_h(x)-\frac{z_\alpha}{\sqrt{nh}},\quad u_n(x)=\hat{p}_h(x)+\frac{z_\alpha}{\sqrt{nh}}.$$

定理7.4 在比较弱的条件下，有下面定理

$$\lim_{n\to\infty}\inf_{\forall x}\mathrm{P}(l_n(x)\leqslant p_h(x)\leqslant u_n(x))\geqslant 1-\alpha.$$

如果要求p的置信带，需要降低偏差，一种较为简单的办法是用二次估计法(twicing).假设有两个核估计$\hat{p}_h$和$\hat{p}_{2h}$. 对于同一个$C(x)$,

$$\mathrm{E}(\hat{p}_h(x))=p(x)+C(x)h^2+o(h^2); \tag{7.7}$$

$$\mathrm{E}(\hat{p}_{2h}(x))=p(x)+C(x)4h^2+o(h^2); \tag{7.8}$$

其中偏差的决定项是$b(x)=C(x)h^2$.可以如下定义

$$\hat{b}(x)=\frac{\hat{p}_{2h}(x)-\hat{p}_h(x)}{3}$$

那么根据(7.7)和(7.8)有

$$\mathrm{E}(\hat{b}(x))=b(x).$$

定义偏差降低法密度估计量为：

$$\tilde{p}_h(x)=\hat{p}_h(x)-\hat{b}(x)=\frac{4}{3}(\hat{p}_h(x)-\frac{1}{4}\hat{p}_{2h}).$$

例7.3 数据见chap7 murder.txt.是英国威尔士18 年间的凶杀案数据，尝试Bootstrap方法，每次有放回选择9 个数据进行0.025 尾分位数估计，由此产生置信区间，比较偏差，尝试偏差降低法密度估计。

解7.3 选定$h=$ 26.23。每次重抽样n =9次，重复$B=5000$ 次,得到如下两个估计：

图7.3上图蓝线为置信上下带，图7.3下图红线为$\widetilde{p}_h(x)=\frac{4}{3}(\hat{p}_h(x)-\frac{1}{4}\hat{p}_{2h}(x))$ 改进结果，黑色虚线为$\hat{p}_{2h}(x)$ 结果。

§7.2.4 多维核密度估计

以上我们考虑的是一维情况下的核密度估计，下面我们考虑多维情况下的核密度估计。

定义7.2 假设数据$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 是d维向量，并取自一个连续分布$p(\boldsymbol{x})$, 在任意点$\boldsymbol{x}$处的一种核密度估计定义为

$$\hat{p}(\boldsymbol{x})=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right), \tag{7.9}$$

注意到这里$p(\boldsymbol{x})$是一个d维随机变量的密度函数。$K(\cdot)$ 是定义在d维空间上的核函数，即$K:\mathbb{R}^d\to\mathbb{R}$, 并满足如下条件：

$$K(\boldsymbol{x})\geqslant 0,\quad \int K(\boldsymbol{x})\ \mathrm{d}\boldsymbol{u}=1.$$

类似于一维情况，我们可以证明$\int_{\mathbb{R}^d}\widehat{p}(\boldsymbol{x})d\boldsymbol{x}=1$，进而可知，$\widehat{p}(\boldsymbol{x})$是一个密度估计。

对于核函数的选择，我们经常选取对称的多维密度函数来作为核函数。例如我们可以选取多维标准正态密度函数来作为核函数，$K_n(\boldsymbol{x})=(2\pi)^{-d/2}exp(-\boldsymbol{x}^T\boldsymbol{x}/2)$。

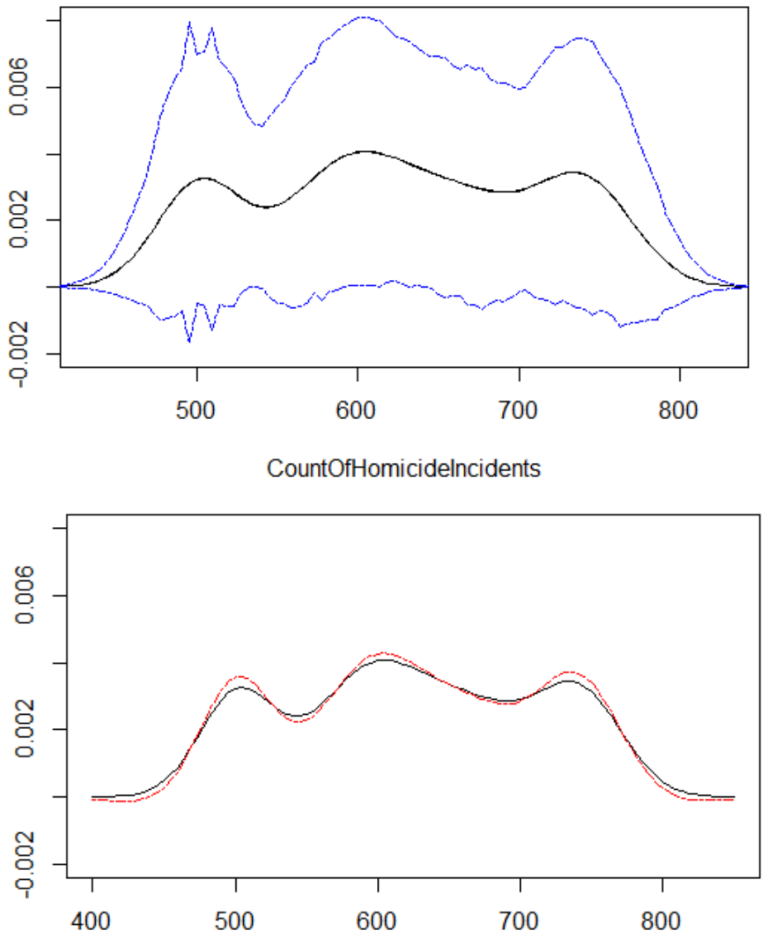

图 7.3 置信区间估计和偏差降低法有偏核密度估计结果

其他常用的核函数还有

$$K_2(\boldsymbol{x}) = 3\pi^{-1}(1-\boldsymbol{x}^T\boldsymbol{x})^2 I(\boldsymbol{x}^T\boldsymbol{x}<1)$$
$$K_3(\boldsymbol{x}) = 4\pi^{-1}(1-\boldsymbol{x}^T\boldsymbol{x})^3 I(\boldsymbol{x}^T\boldsymbol{x}<1)$$
$$K_e(\boldsymbol{x}) = \frac{1}{2}c_d^{-1}(d+2)(1-\boldsymbol{x}^T\boldsymbol{x})I(\boldsymbol{x}^T\boldsymbol{x}<1)$$

K_e被称为多维Epanechinikow核函数，其中c_d是一个和维度有关的常数，$c_1=2$, $c_2=\pi$, $c_3=4\pi/3$.

上述的多维核密度估计中，我们只使用了一个带宽参数h，这意味着在不同方向上，我们取的带宽是一样的。事实上，我们可以对不同方向取不同的带宽参数，即

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh_1\cdots h_d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{\boldsymbol{h}}\right).$$

其中，$\boldsymbol{h}=(h_1,\cdots,h_d)$是一个$d$维向量。在实际数据中，有时候一个维度上的数据比另外一个维度上的数分散得多，这个时候上述的核函数就有用了。比如说数据在一个维度上分布在$(0,100)$区间上，而在另一个维度上仅仅分布在区间$(0,1)$ 上，这时候采用不同带宽的多维核函数就比较合理了。

例7.4 下例是美国黄石国家公园的Old Faithful Geyser 数据, 它包含272对数据, 分别为喷发时间和喷发的间隔时间。我们以此数据估计喷发时间和喷发的间隔时间的联合密度函数.

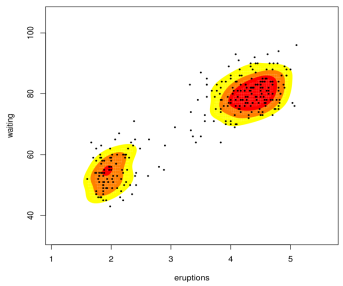

图 7.4 喷发时间和喷发的间隔时间的联合密度函数估计

```
library(ks)
data(faithful)
H <- Hpi(x=faithful)
fhat <- kde(x=faithful, H=H)
```

```
plot(fhat, display="filled.contour2")
points(faithful, cex=0.5, pch=16)
```

关于最优带宽的选择，也有类似一维情况下的结论。对于多维核密度估计，运用多维泰勒展开，有

$$Bias(\boldsymbol{x}) \approx \frac{1}{2}h^2\alpha\nabla^2 p(\boldsymbol{x}),$$
$$V(\widehat{p}(\boldsymbol{x}) \approx n^{-1}h^{-d}\beta p(\boldsymbol{x}).$$

其中，$\alpha = \int \boldsymbol{x}^2 K(\boldsymbol{x})d\boldsymbol{x}$, $\beta = \int K(\boldsymbol{x})^2 d\boldsymbol{x}$.

因此我们可以得到渐进积分均方误

$$AMISE = \frac{1}{4}h^4\alpha^2 \int \nabla^2 p(\boldsymbol{x})d\boldsymbol{x} + n^{-1}h^{-d}\beta.$$

由此可得最优带宽为

$$h_{opt} = \left[d\beta\alpha^{-2}(\int \nabla^2 p(\boldsymbol{x})d\boldsymbol{x})\right]^{1/(d+4)} n^{-1/(d+4)}$$

在上述的最优带宽中，真实密度$p(\boldsymbol{x})$是未知的，因此我们可以采用多维正态密度$\phi(\boldsymbol{x})$来代替，进而得到

$$h_{opt} = A(K)n^{-1/(d+4)},$$

其中$A(K) = \left[d\beta\alpha^{-2}(\int \nabla^2 \phi(\boldsymbol{x})d\boldsymbol{x})\right]^{1/(d+4)}$.

对于$A(K)$，在知道估计中的核函数类型后，可以计算出来，并进而得到最优带宽h_{opt}. 以下是不同核函数的$A(K)$，

表7.2 不同核函数下的$A(K)$值

ID	核函数	维度	$A(K)$
1	K_n	2	1
2	K_n	d	$\{4/(d+2)\}^{1/(d+4)}$
3	K_e	2	2.40
4	K_e	3	2.49
5	K_e	d	$\{8c_d^{-1}(d+4)(2\sqrt{\pi})\}^{1/(d+4)}$
6	K_2	2	2.78
7	K_3	2	3.12

§7.2.5 贝叶斯决策和非参数密度估计

分类决策是对一个概念的归属作决定的过程, 比如：生物物种的分类、手写文字的识别、西瓜是否成熟、疾病的诊断等. 如果一个概念的自然状态是相对确定的, 要对比不同决策的优劣是相对容易的. 比如：一个人国籍身份的归属, 根据我国国籍法规定“父母双方或一方为中国公民, 本人出生在中国, 具有中国国籍”. 即父母的身份和一个人的出生地可以作为公民国籍归属的基本识别属性. 一个不在中国出生的婴儿如果已有他国国籍, 则不具有中国国籍. 这是一个概念规则相对比较清晰的例子, 然而现实中更多问题的根本是需要形成较为清晰的、可操作性较强的分类规则, 比如：信用评价问题、垃圾邮件识别问题、欺诈侦测问题等. 在诸如此类的问题中, 我们可能收集到信用不良事件和信用良好事件的线索记录, 比如发生时间、发生地点、当事人历史记录等, 希望通过对收集到的信息进行分析比较, 从而找出可用于信用概念评价的一些识别属性, 完成分类规则建制的基本任务.

不仅如此, 由于决策过程常常面对的是一个信息不充分的环境, 这就是说决策不可避免地会犯错误, 于是决策研究中对分类决策的评价就成为不可或缺的核心内容. 综上所述, 一个分类框架一般由四项基本元素构成.

(1) 参数集：概念所有可能的不同自然状态. 在分类问题中, 自然参数是可数个, 用$\Theta = \{\theta_0, \theta_1, \cdots\}$表示.

(2) 决策集：所有可能的决策结果$\mathcal{A} = \{a\}$. 比如：买或卖、是否癌症、是否为垃圾邮件,在分类问题中, 决策结果就是决策类别的归属, 所以决策集与参数集往往是一致的.

(3) 决策函数集：$\Delta = \{\delta\}$,函数$\delta : \Theta \to \mathcal{A}$.

(4) 损失函数：联系于参数和决策之间的一个损失函数. 如果概念和参数都是有限可数的, 那么所有的概念和相应的决策所对应的损失就构成了一个矩阵.

例7.5 两类问题中, 真实的参数集为θ_1和θ_0(分别简记为1或0), 可能的决策集由四个可能的决策构成$\Delta = \{\delta_{1,1}, \delta_{0,0}, \delta_{0,1}, \delta_{1,0}\}$.其中, $\delta_{i,j}$表示把i判为$j, i, j = 0, 1$, 相应的损失矩阵可能为

$$\boldsymbol{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

这表示判对没有损失, 判错有损失. 真实的情况为1判为0, 或真实的情况为0判为1, 则发生损失1, 称为“0-1”损失.

从分布的角度来看, 分类问题本质上是概念属性分布的辨识问题, 于是可能通过密度估计回答概念归属的问题. 以两类问题为例：真实的参数集为θ_1和θ_0,在没有

观测之前, 对θ_1和θ_0的决策函数可以应用先验$p(\theta_1)$和$p(\theta_0)$确定, 即定义决策函数

$$\delta = \begin{cases} \theta_1, & p(\theta_1) > p(\theta_0), \\ \theta_0, & p(\theta_1) < p(\theta_0). \end{cases}$$

很多情况下, 我们对概念能够收集到更多的观测数据, 于是可以建立类条件概率密度$p(x|\theta_1), p(x|\theta_0)$.显然, 两个不同的概念在一些关键属性上一定存在差异, 这表现为两个类别在某些属性上面分布呈现差异. 综合先验信息, 可以对类别的归属通过贝叶斯公式重新组织. 即

$$p(\theta_1|x) = \frac{p(x|\theta_1)p(\theta_1)}{p(x)},$$
$$p(\theta_0|x) = \frac{p(x|\theta_0)p(\theta_0)}{p(x)}.$$

根据贝叶斯公式, 我们可以通过后验分布制定决策：

$$\delta = \begin{cases} \theta_1, & p(\theta_1|x) > p(\theta_0|x), \\ \theta_0, & p(\theta_1|x) < p(\theta_0|x). \end{cases}$$

注意到后验概率比较中, 本质的部分是分子, 所以上式等价于

$$\delta = \begin{cases} \theta_1, & p(x|\theta_1)p(\theta_1) > p(x|\theta_0)p(\theta_0), \\ \theta_0, & p(x|\theta_1)p(\theta_1) < p(x|\theta_0)p(\theta_0). \end{cases}$$

定理7.5 后验概率最大化分类决策是“0-1”损失下的最优风险.

证明 注意到条件风险

$$\begin{aligned} R(\theta_1|x) &= p(\theta_0|x)L(\theta_0, \theta_1) + p(\theta_1|x)L(\theta_1, \theta_1) \\ &= 1 - p(\theta_1|x). \end{aligned}$$

上述定理很容易扩展到$k, k \geqslant 3$个不同的分类(此处不再赘述, 留作练习). 后验概率最大相当于“0-1”损失下的最小风险.

于是给出如下的非参数核密度估计分类计算步骤：

后验分布构造贝叶斯分类

1. $\forall i = 1, 2, \cdots, k, \theta_i$下观测$x_{i1}, x_{i2}, \cdots, x_{in} \sim p(x|\theta_i)$;
2. 估计$p(\theta_i), i = 1, 2, \cdots, k$;
3. 估计$p(x|\theta_i), i = 1, 2, \cdots, k$;
4. 对新待分类点x,计算$p(x|\theta_i)p(\theta_i)$;
5. 计算$\theta^* = \text{argmax}\{p\ (x|\theta_i)p\ (\theta_i)\}$.

例7.6(例7.1续) 根据核密度估计贝叶斯分类对例7.1中的两类鱼进行分类.

解 假设θ_1表示鲑鱼, θ_0表示鲈鱼, 记两类鱼的先验分布为

$$\text{鲑鱼:}\hat{p}\ (\theta_1) \quad \leftrightarrow \quad \text{鲈鱼:}\hat{p}\ (\theta_0).$$

用两类分别占全部数据的频率估计先验概率. 在本例中, 由于鲑鱼为100条, 鲈鱼为130条, 两类先验概率分别估计为: $p\ (\theta_1) = 100/230 = 0.4348; p\ (\theta_0) = 130/230 = 0.5652$.

接着, 对每一类别独立估计概率密度, 两类鱼身长的核概率密度分别记为

$$\text{鲑鱼:}p\ (x|\theta_1) \quad \leftrightarrow \quad \text{鲈鱼:}p\ (x|\theta_0).$$

根据“最大后验概率”的原则进行分类制定如下判别原则: 对$\forall x$,

$$\delta_x \in \begin{cases} \theta_0, & \text{当 } p\ (\theta_0|x) > p\ (\theta_1|x), \\ \theta_1, & \text{当 } p\ (\theta_1|x) > p\ (\theta_0|x). \end{cases}$$

下面我们针对一组数据点, 得到表7.3所示的分类结果.

表7.3 用核密度估计对鲑鱼和鲈鱼的分类结果表

位置	数值	$p^*(\theta_1\|x)$	$p^*(\theta_0\|x)$	真实的类别	判断的类别
83	19.6	0.0506	0.0071	1	1
82	22.3	0.0593	0.0069	1	1
220	14.07	0.0076	0.0179	0	0
89	8.5	0.0046	0.0634	1	0
93	17.3	0.0135	0.0112	1	1
167	7.6	0.0044	0.0777	0	0
140	6.3	0.0051	0.0583	0	0
107	2	0.0001	0.0293	0	0

注: p^*表示没有归一化的分布密度.

核函数密度曲线如图7.5所示.

表中有下滑线的数据表示分类错误. 如上结果有8个数据, 7个分类正确, 1个分类错误, 在表7.2中用下划线标记.

上述的概率密度估计和分类的例子已经较好地说明了非参数密度估计的优点. 如果能采集足够多的训练样本, 无论实际采取哪一种核函数形式, 从理论上最终可以得到一个可靠的收敛于密度的估计结果. 概率密度估计和分类例子的主要缺点是为了获得满意的密度估计, 实际需要的样本量却是非常惊人的. 非参数估计要求的

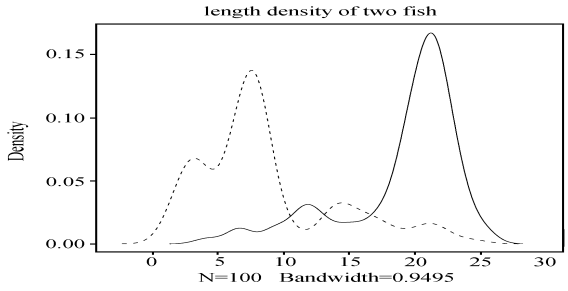

图 7.5 鲑鱼和鲈鱼核函数密度曲线图

样本量远超过在已知分布参数形式下估计所需要的样本量. 这种方法对时间和内存空间的消耗都是巨大的, 人们也正在努力寻找有效降低估计样本量的方法.

然而, 非参数密度估计最严重的问题是高维应用问题. 一般在高维空间上, 会考虑定义一个d维核函数为一维核函数的乘积, 每个核函数有自己的带宽, 记为$h_1, h_2, \cdots, h_d$, 参数数量与空间维数呈线性关系. 然而在高维空间中, 任何一个点的邻域里没有数据点是很正常的, 因而出现了体积很小的邻域中的任意两个点之间的距离却很远, 比如10维空间上位于一个体积为0.001的小邻域内的两个点的距离可以允许高到0.5, 这样基于体积概念定义的核函数没有样本点估计. 这种现象被称为“维数灾难”问题(curse of dimensionality). 为了使核估计能够应用, 则需要更多的样本作为代价. 因此这也严重限制了非参数密度估计在高维空间上的应用.

§7.3 k近邻估计

Parzen窗估计一个潜在的问题是每个点都选用固定的体积. 如果h_n定的过大, 则那些分布较密的点由于受到过多点的支持, 使得本应突出的尖峰变得扁平; 而对于另一些相对稀疏的位置或离群点, 则可能因为体积设定过小, 而没有样本点纳入邻域, 从而使密度估计为零. 虽然可能选择像正态等一些连续核函数, 能够在一定程度上弱化该问题, 但很多情况下并不具有实质性的突破, 仍然没有一个标准指明应该按照哪些数据的分布情况制定带宽. 一种可行的解决方法就是让体积成为样本的函数, 不硬性规定窗函数为全体样本个数的某个函数, 而是固定贡献的样本点数, 以点$\boldsymbol{x}$为中心, 令体积扩张, 直到包含进k_n个样本为止, 其中的k_n是关于n的某一个特定函数. 被吸收到邻域中的样本就称为点$\boldsymbol{x}$的k_n个最近邻. 用停止时的体积定义估计点的密度如下：

$$\tilde{p}_n(\boldsymbol{x}) = \frac{k_n/n}{V_n}. \tag{7.10}$$

如果在点$\boldsymbol{x}$附近有很多样本点, 那么这个体积就相对较小, 得到很大的概率密度; 而如果在点$\boldsymbol{x}$附近很稀疏, 那么这个体积就会变大, 直到进入某个概率密度很高的区域, 这个体积就会停止生长, 从而概率密度比较小.

如果样本点增多, 则k_n也相应增大, 以防止V_n快速增大导致密度趋于无穷. 从另一方面, 我们还希望k_n的增加能够足够慢, 使得为了包含进k_n个样本的体积能够逐渐地趋于零. 在选择k_n方面, 福永(Fukunaga)和霍斯特勒(Hosterler)(1975)给出了一个计算$k(n)$的公式, 对于正态分布而言：

$$k = k_0 n^{4/(d+4)}. \tag{7.11}$$

式中, k_0是常数,与样本量n和空间维数d无关.

如果取$k_n = \sqrt{n}$, 并且假设$\tilde{p}_n(\boldsymbol{x})$是$p(\boldsymbol{x})$的一个较准确的估计, 那么根据方程式(7.10), 有$V_n \approx 1/(\sqrt{n}p(\boldsymbol{x}))$. 这与核函数中的情况是一样的. 但是这里的初始体积是根据样本数据的具体情况确定的, 而不是事先选定的. 而且不连续梯度的点常常并不出现在样本点处, 见图7.6.

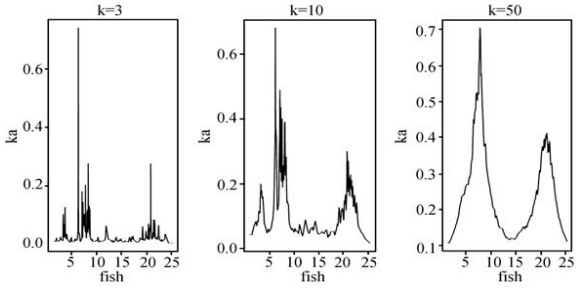

图 7.6　不同k的近邻密度估计图

与核函数一样, k_n近邻估计也同样存在维度问题. 除此之外, 虽然$\tilde{p}_n(\boldsymbol{x})$是连续的, 但$k$近邻密度估计的梯度却不一定连续. k_n近邻估计需要的计算量相当大, 同时还要防止k_n增加过慢导致密度估计扩散到无穷. 这些缺点使得用k_n近邻法产生密度并不多见, k_n近邻法更常用于分类问题.

习题

7.1　使用R里的library(MASS)中的案例数据geyser老忠实温泉数据,对间隔时间作核估计.

(1) 取h=0.3,选用标准正态密度函数、Parzen窗函数和三角函数分别作图, 分析不同窗函数对结果的影响.

(2) 固定核函数为标准正态密度, 取h为四个不同的值:$h = 0.3, 0.5, 1$和1.5, 从图上分析带宽对核密度估计的影响.

7.2 对鲑鱼和鲈鱼识别数据, 尝试用k_n方法估计两类的分布密度, 再尝试贝叶斯方法设计分类器.

(1) 选择所使用的k近邻数.

(2) 在不同的k之下计算训练误差率.

7.3 考虑一个正态分布$p(x) \sim N(\mu, \sigma^2)$和核函数$K(x) \sim N(0,1)$. 证明Parzen窗估计$p_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right)$有如下性质.

(1) $\hat{p}_n(x) \sim N(\mu, \sigma^2 + h_n^2)$.

(2) $\mathrm{var}[p_n(x)] \approx \frac{1}{2nh_n\sqrt{\pi}}p(x)$.

(3) 当h_n较小时, $p(x) - \bar{p}_n(x) \approx \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2\left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right]p(x)$. 注意, 如果$h_n = h_1/\sqrt{n}$, 那么这个结果表示由于偏差而导致的误差率以$1/n$的速度趋向于零.

7.4 令$p(x) \sim U(0,a)$为0到a之间的均匀分布, 而Parzen窗函数为当$x > 0$时, $\varphi(x) = \mathrm{e}^{-x}$, 当$x \leqslant 0$时则为零.

(1) 证明Parzen窗估计的均值为

$$\hat{p}_n(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{a}(1 - \mathrm{e}^{-x/h_n}), & 0 \leqslant x \leqslant a, \\ \frac{1}{a}(\mathrm{e}^{a/h_n} - 1)\mathrm{e}^{-x/h_n}, & a \leqslant x. \end{cases}$$

(2) 画出当$a = 1$, h_n分别等于1, 1/4, 1/16时的$\hat{p}_n(x)$关于x的函数图像.

(3) 在这种情况下, 即$a = 1$时, 求h_n的值. 并且画出区间$0 \leqslant x \leqslant 0.05$的$\hat{p}_n(x)$的函数图像.

7.5 假设x_1, x_2相互独立满足$(0,1)$间的均匀分布, 考虑指数核函数$K(u) = \exp\{-|u|\}/2$.

(1) 写出核函数密度估计的表达式$\hat{p}(x)$.

(2) 计算$Bias(x) = E\hat{p}(x) - p(x)$.

7.6 对于多维核密度函数$K_e(\boldsymbol{x}) = \frac{1}{2}c_d^{-1}(d+2)(1 - \boldsymbol{x}^T\boldsymbol{x})I(\boldsymbol{x}^T\boldsymbol{x} < 1)$, 其中$d$是多元核函数的维度

(1)当$d = 2$和3时,分别计算c_d的值并写出对应的多维核密度函数的表达式

(2)当$d = 2$时, 我们有数据$(1,1),(1,2),(2,1),(2,2)$,试计算核密度估计$\hat{p}(\boldsymbol{x})$在$\boldsymbol{x} = (1.5, 1.5)$的值.

(3)当$d = 3$时, 我们有数据$(1,1,1),(1,2,1),(2,1,2),(2,2,2)$, 试计算核密度估计$\hat{p}(\boldsymbol{x})$ 在$\boldsymbol{x} = (1,1,2)$的值.

7.7 信用卡信用被分为三级,试利用光盘上所给的Credit.txt数据根据核估计法和后验概率来构造分类器. 尝试R中的所有可能的核函数, 并比较不同的结果.

7.8 对于凶杀案数据，尝试R中包sm.envelope 之间的分析差异。

案例与讨论:景区游客时空分布密度与预测框架

案例背景

在我国，旅游业是保护生态资源环境可持续发展的绿色产业。中国旅游研究院的数据显示，2015年中国旅游接待总人数已经突破41亿人次。而伴随而来的则是在旅游旺季，知名景区的旅游线路超负荷承载，配套服务协调的失控等。游客数量暴增，特别是大散客时代的到来，让游览需求更加多样化：附近的停车场还有空车位吗？而2020 年受新冠疫情影响,全球旅游业进入寒冬,最近的洗手间在哪儿，安全卫生的餐厅距离当前位置有多远，排队状况如何，去往下一个景点的电瓶车、哪个泊位的游船人数较少几点能来… 这些细微琐碎的服务需求已经难以再靠传统方式满足。同时，对于景区管理者来而言，旅游管理中面对的种种问题亟需大数据的帮忙。如何快速向游客推送景区各类信息，如何获知人流热度以便及时指挥调度，如何管理景区的景点、道路、设施相关数据，这些都是国内更多传统景区转型中亟需攻克的难点。景区游客实时预报和对游客流动分布的监测与客流量合理疏导是“旅游产业管理”转型升级的必要之路。

其中景区热度分析即景区人群密度预测技术是其中的关键。具体而言，基于大数据热度信息，可以帮助游客和景区绘制景区内精准的基础地图数据，帮助游客和景区进行拥堵、排队等人流、车流大数据采集、分析基于位置（LBS）的大数据，帮助景区进行实时活动信息、地址信息变更等在线数据管理。作为“智慧景区”主要引擎的热力图，游客可以通过其显示的不同颜色，判断该处游客人数的多少，合理安排游览时间。

据报道，建设“智慧景区”已经成为我国旅游业发展的一个新趋势。2015年9月，国家旅游局发布了《“旅游+互联网”行动计划》，明确到2020年，推动全国所有4A 级景区实现免费WIFI、智能导游、电子讲解、在线预订、信息推送等功能全覆盖。据统计，截至2015年底，全国共有5A景区213家，4A景区617家，3A景区更是不计其数。

请看下面这张图，下图反映的是颐和园某时间段内景区热度密度估计结果，从中可以发现同一时段景区不同位置人口密度的分布存在异质性结构，通过传统的核密度估计只能得到这张图右下边的结果，这张图反映出许多潜在的弱密度区域有被低估而高密度结构被高估的现象,这样就产生了数据点在构成密度估计中的权重选择问题。

研讨问题

请阅读论文《数理统计与管理》2018年3期438-448页上的文章《基于权重时变的混合正态模型的游客分布预测模型》，讨论以下4个问题：

1.为什么传统的核密度估计会出现低密度区域被低估的现象，这类问题在单一的正态分布中会出现吗？

2.为什么传统的核密度估计会出现高密度区域被高估的现象，这类问题在单一的Gamma分布中会出现吗？

3.变权重的估计在不平衡的混合密度估计中有怎样的作用，结合图上左下的输出结果进行思考。这些估计技术在体现不同分支的机会平等要求和不同时间前驱后继结构的发现与解读方面有哪些独特的作用；

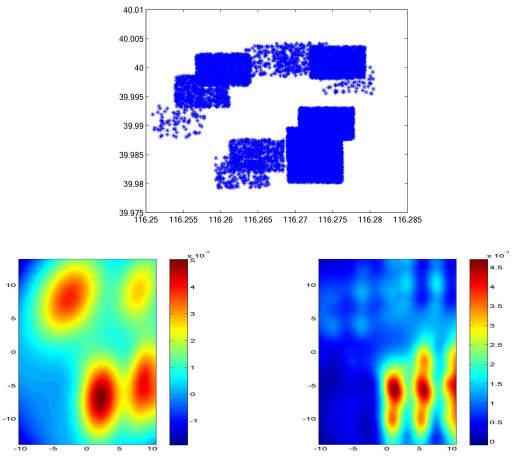

图 7.7 2017年某日10:00am景区游客分布空间分布图(上)，时变权重的密度估计（左下），传统的核密度估计（右下）

4.这类模型的建立需要导入怎样的数据，它对景区游客的互动信息服务平台的哪些决策会有帮助，会创新哪些新的业务模式，请收集文献给予分析和讨论。