模式识别的理论与方法

Pattern Recognition

裴继红

Chapter 4 (Part 1):
Nonparametric Techniques

(Sections 4.1-4.3)

本讲内容

- 引言
- 非参数概率密度估计的一般原理
- 概率密度估计的**Parzen**窗方法

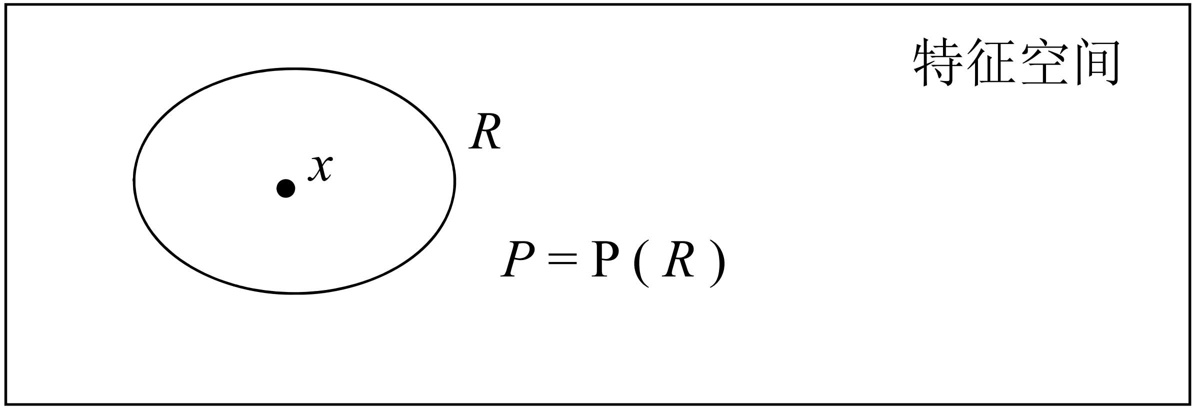
引言

- 概率密度的参数估计方法
  - 要求知道概率的总体分布形式。而许多的实际问题并不知道总体分布形式，或者不能写成通常的参数形式。
  - 所有密度的参数估计方法都是针对单模的，即只有一个局部最大值。而在许多的实际问题中往往涉及到多模的情况。

- 非参数估计方法
  - 可用于任意分布形式的估计，不必事先假设概率分布的形式。
  - 很多非参数估计方法的**核心思想都非常简单**

- **两种典型的非参数估计方法：**
  - 用样本估计**类条件概率密度** $P(x \mid \omega_j)$
  - 用样本直接估计**后验概率密度** $P(\omega_j \mid x)$

区域 R 中事件发生的概率



- 若已知区域 R的概率P，如何估计 $p(x)$?

例如 R 为区间 : $\{ x \mid 20 \text{ cm} < \text{length}(x) < 30 \text{ cm} \}$

问在 x = 25cm处的概率为多少?

概率密度 $p(x)$的估计

若 $p(x)$ 是连续函数，且 R 足够小， $p(x)$在 R 内近似不变，则

$$P = \int_R p(x')\,dx' \cong p(x)V$$

这里V是R包围的体积

可得 $p(x) = P / V$

现在，问题转化为如何估计 P

概率密度估计

- **基本思路:**

特征空间中的点（矢量）x 落在区域 R 中的概率为:

$$P = \int_R p(x')dx'$$

其中，P 是密度函数 $p(x)$ 的某种平均。

若已知样本数为 n，且每个样本都是独立同分布抽取的，则其中 k 个样本落在区域 R 中的概率服从二项式分布:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k}$$

由二项式分布可知，**k** 的期望值为:

$$\mathbf{k}_n = E(\mathbf{k}) = n\mathrm{P}$$

概率密度估计

$$\frac{k_n}{n} \cong P = \int_R p(x')dx' \cong p(x)V$$

$$\boxed{p_n(x) \cong \frac{k_n/n}{V_n}}$$

$p(x)$ 的估计

- 假设 n 个样本中的 k 个样本落在区域 R 中. 若样本是独立同分布的，则 P 近似为 k/n.
  - 若采用最大似然ML估计，令 $P=\theta$ 则同样

$$\underset{\theta}{Max}(P_k \mid \theta) \quad \text{达到最大时} \quad \hat{\theta}=\frac{k}{n}\cong P$$

因此，比值k/n 是概率 P 的一个好的估计

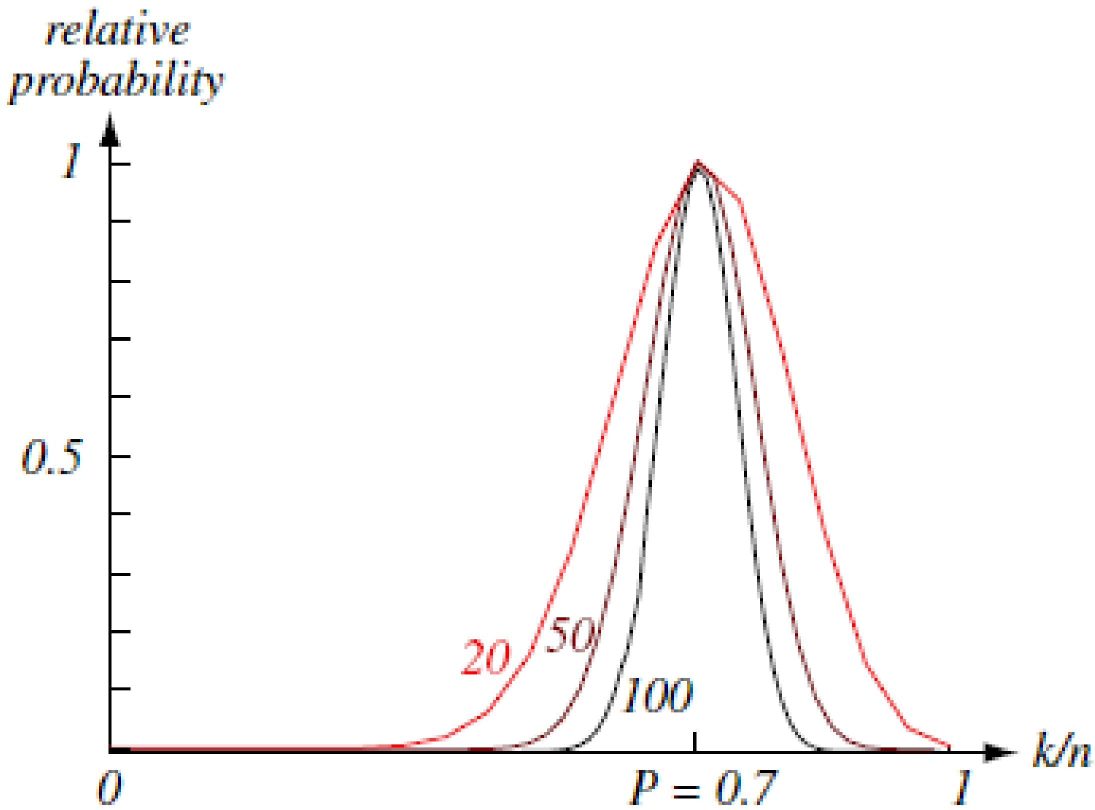
由于 P 可以近似为 k/n，因此可得

$$\boxed{p_n(x)\cong\frac{k_n/n}{V_n}}$$

相对概率的最大似然估计 Figure4.1

- 由最大似然估计ML估计的相对概率 k / n，
- 图中　P= 0.7是产生样本的真实的概率。

  每条曲线上的数字表示产生该曲线的样本n 的个数。当 n→ ∞时，曲线在P= 0.7处近似为一个 δ 函数。



对密度估计公式的简单证明

- 公式

$$\int_R p(x')\,dx' \cong p(x)V$$

由于假设 $p(x)$是连续的，且R足够小， p 在R中没有明显的变化，因此 $p(x)$ ≌常数，这样

$$\int_R p(x')dx' = p(x')\int_R dx' = p(x')\int_R 1_R(x)dx' = p(x')\mu(R)$$

这里， $\mu(R)$: 在二维欧氏空间R^2中代表面积，在三维欧氏空间R^3中代表体积，在 n 维欧氏空间 R^n 中代表超体积。因此

$$p(x) \cong \frac{k}{nV}$$

上述密度估计收敛的条件

- 分数 $k/(nV)$ 是 $p(x)$ 的空间平均值，
  - 只有在体积 *V* 趋近于 0 时，才可以得到真正的 $p(x)$。
  - 而若 n=常数，则这有可能产生 R 中没有任何样本的无意义的情况，这时

$$\lim_{V \to 0,\ k=0} p(x) = 0$$

  - 也可能是另一种相反的情况：R中恰好有一个样本，此时

$$\lim_{V \to 0,\ k \neq 0} p(x) = \infty$$

收敛条件讨论

- 在实际中，由于样本数量 n 总是有限的，因此体积 V 不可能任意小。
  - 为此必须允许比值 k/n 有一定的波动
- 理论上，若样本数量无限的话，可以使用下面的方法进行估计：
  - 构造并组成一个包含 x 的区域的序列 $R_1, R_2, \ldots,$
  - 第一个区域包含一个样本，
  - 第二个区域包含2个样本，…，以此类推。
- 令 V_n 是区域 R_n 的体积， k_n 是落在区域 R_n 中的样本数量，是对 $p(x)$ 的第 n 次估计，则

$$p_n(x) = (k_n/n)/V_n$$

收敛条件讨论

- 要使 $p_n(x)$ 收敛于 $p(x)$，需要满足下面的三个必要条件：

$$1)\ \lim_{n\to\infty} V_n = 0 \qquad 2)\ \lim_{n\to\infty} k_n = \infty \qquad 3)\ \lim_{n\to\infty} k_n / n = 0$$

为了找到真实的而非平均的 $p(x)$, 在逼近过程中需要随着样本数目的增大减小体积V。

本章介绍两种满足上面条件的构造区域序列的方法

1）**Parzen窗估计技术**

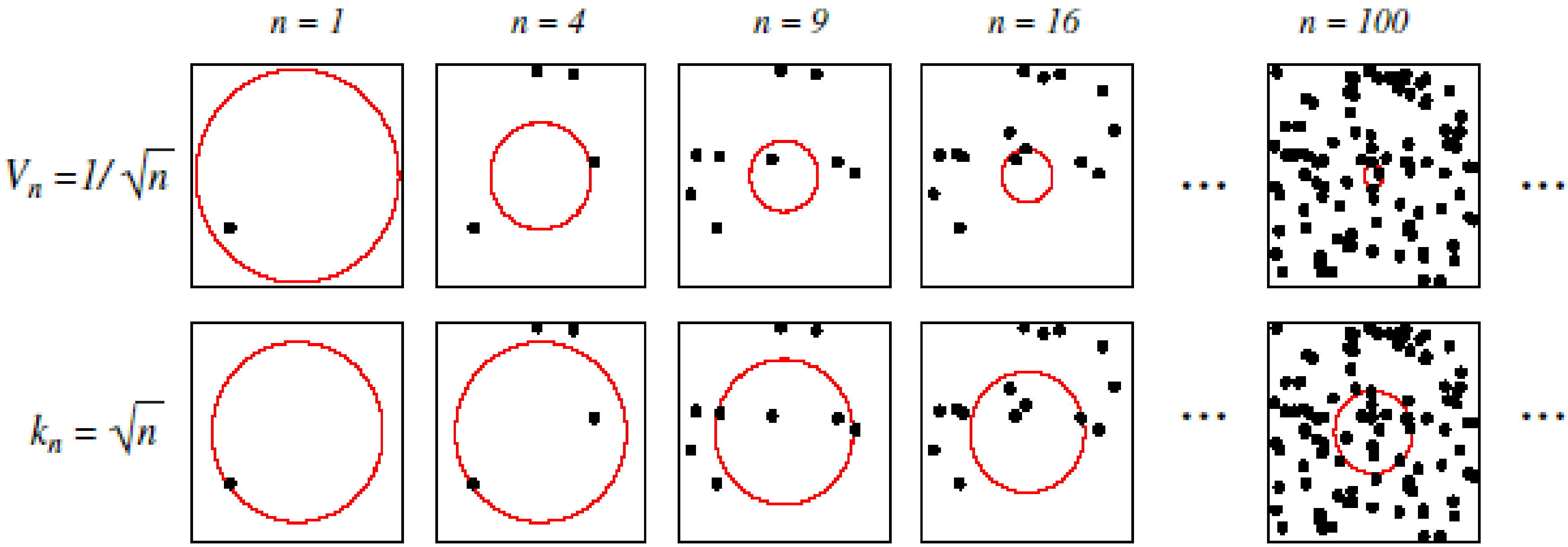
将体积作为n的函数进行收缩，如 $V_n \propto \frac{1}{\sqrt{n}}$

2）**最近邻估计技术**

将具体的 ***k*** 值作为***n***的函数，如 $k_n \propto \sqrt{n}$

两种主要的估计方法原理图示

上面一行为 Parzen窗方法，下面一行为 ***k*** 最近邻方法。



Parzen窗方法

- 在Parzen窗方法中，首先假设体积V是d-维空间的一个超立方体，若 h 是该超立方体的边长，则

$$V = h^d$$

- 概率密度估计公式为

$$p(x) = \frac{P}{V} = \frac{k/n}{V}$$

- 在Parzen窗方法中，样本数 n 确定以后，体积 ***V*** 就确定了, 但是落在超立方体中的样本数未知

窗函数

➢ 定义窗函数:

$$\varphi(u)=\begin{cases}1 & |u_j|\le 1/2,\quad j=1,2,\cdots,d\\ 0 & otherwise\end{cases}$$

这是一个中心在原点的超立方体。易知，若 $\boldsymbol{x}_{\mathrm{i}}$ 落入中心在 $\boldsymbol{x}$ 的超立方体内部，则

$$\varphi\left(\left(x-x_i\right)/h\right)=1$$

窗函数

- 因此，落入超立方体内部的样本数为

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h}\right)$$

进一步我们有

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{x - x_i}{h}\right)$$

即对 $p(x)$的估计$P_n(x)$ 是关于 $\mathbf{x}$ 和样本 $\mathbf{x}_i$ 的窗函数的平均，每一个样本根据其到 $\mathbf{x}$ 的距离对函数值有不同的贡献。

窗函数的形式

- 指数函数
- 高斯函数
- 。。。。。。

➢ 具有对称衰减特性的函数

窗口宽度 *h* 的影响

在估计 ***p*(*x*)** 时，窗口宽度 ***h*** 对估计的影响如何?

- 首先定义 $\delta(x)$ 函数：

$$\delta(x)=\frac{1}{V}\varphi\left(\frac{x}{h}\right)$$

- 这样可将 $p(x)$ 写为

$$p(x)=\frac{1}{n}\sum_{i=1}^{n}\delta(x-x_i)$$

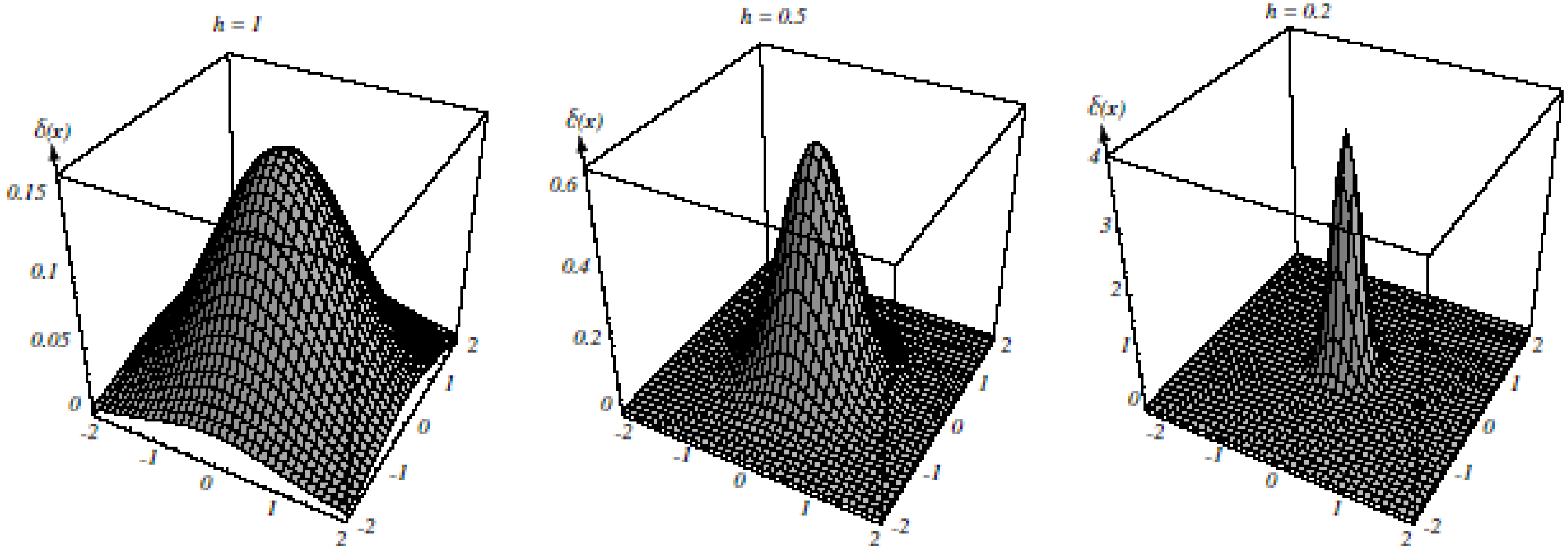
- 窗口宽度 ***h*** 影响着 $\delta(x)$ 函数的幅度和宽度

窗口宽度 *h* 的影响

- 若 ***h*** 大，则函数的幅度小，同时在距离 $\mathbf{x}_i$ 较远的地方才能观察到 $\delta(x\text{-}x_i)$ 的明显变化。
  - 在这种情况下， ***p***(**x**) 是一个非常平滑的（或称为散焦的）估计
- 若 ***h*** 非常小, 则$\delta(x\text{-}x_i)$ 在靠近 x_i 的 x 处就有非常大的值。
  - 在这种情况下，估计的 ***p***(**x**) 在样本点 x_i 为中心局部空间存在一个非常尖锐的脉冲，
  - ***p***(**x**) 是一个充满“噪声”的估计

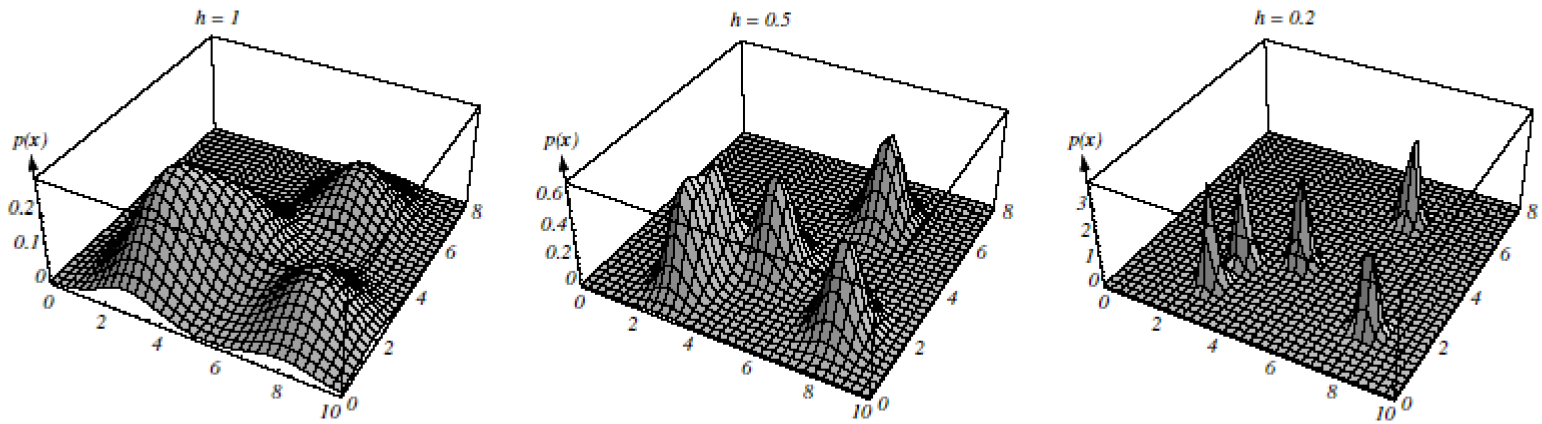
不同宽度的Parzen窗 示例

- 对应于三个不同的h 值的二维圆对称正态Parzen窗。
  - 注意，由于$\delta(\mathbf{x})$ 已经被归一化，每一个图的垂直尺度经过了调整



不同宽度Parzen窗的密度估计

- 分别使用图4.3 中的窗函数对具有5个样本点的相同的样本集合进行Parzen窗密度估计。
  - 与前面的图类似，每个图的垂直尺度已经过调整



Parzen窗估计的例子

- 令 ***p***(***x***) 是一个 0 均值、单位方差的正态密度，令窗函数具有下面的形式：

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$

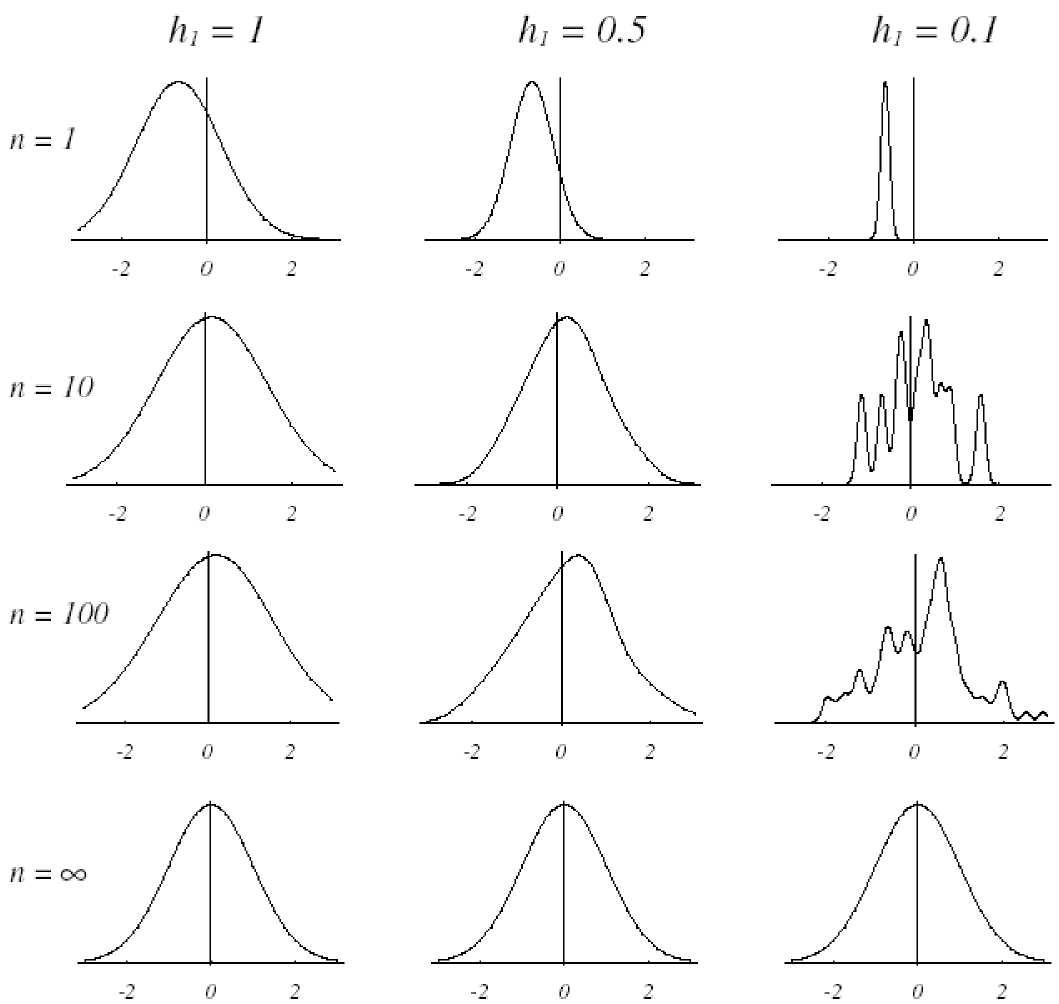
这样，得到的估计结果是一个以每一个样本点所在的位置为中心的正态密度函数的所有样本点窗函数的平均

$$p_n(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)$$

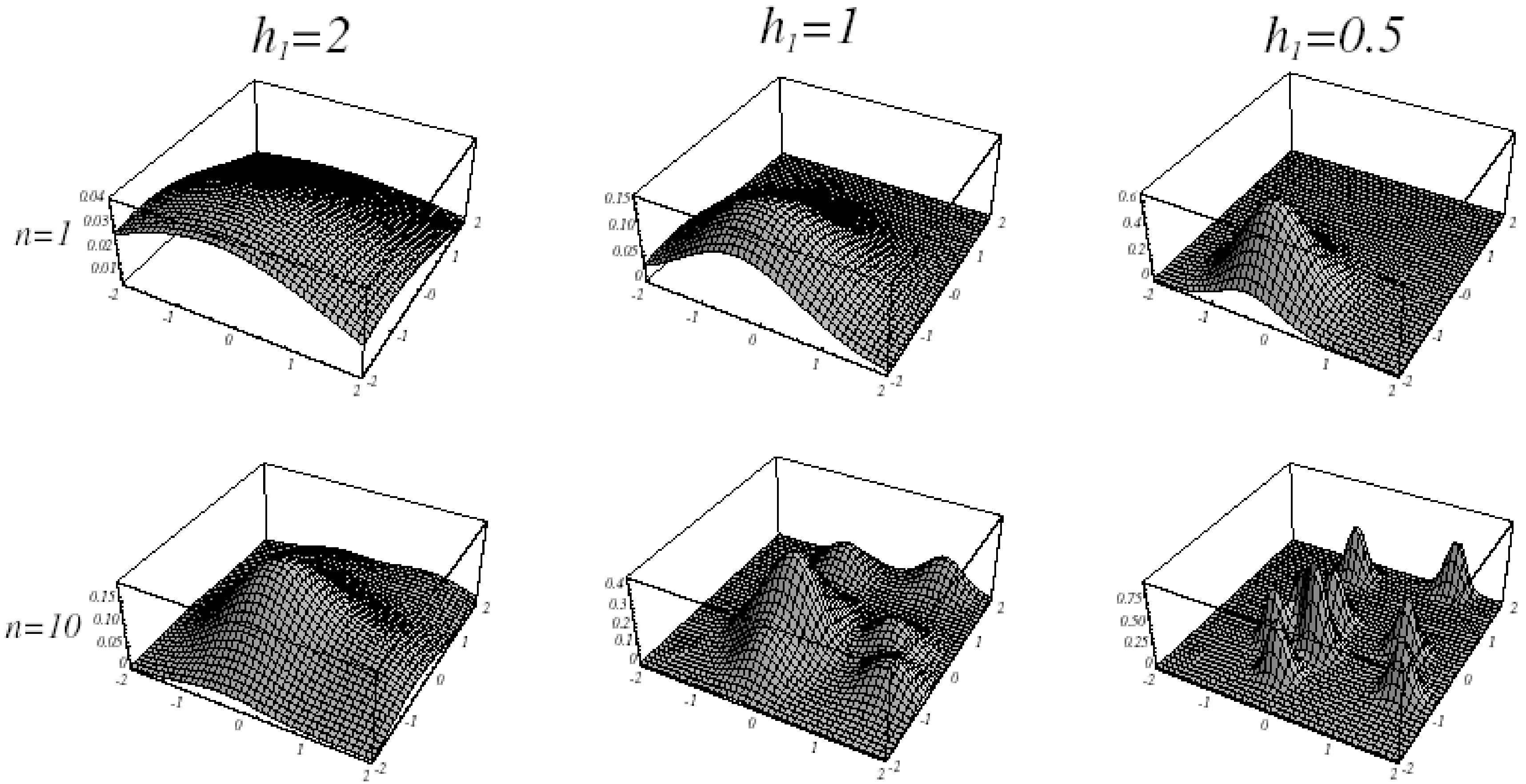
这里 $h_n=h_1/\sqrt{n}$ $\boldsymbol{h_1}$ 是一个我们可以控制调整的参数

一维正态窗的密度估计：Figure 4.5

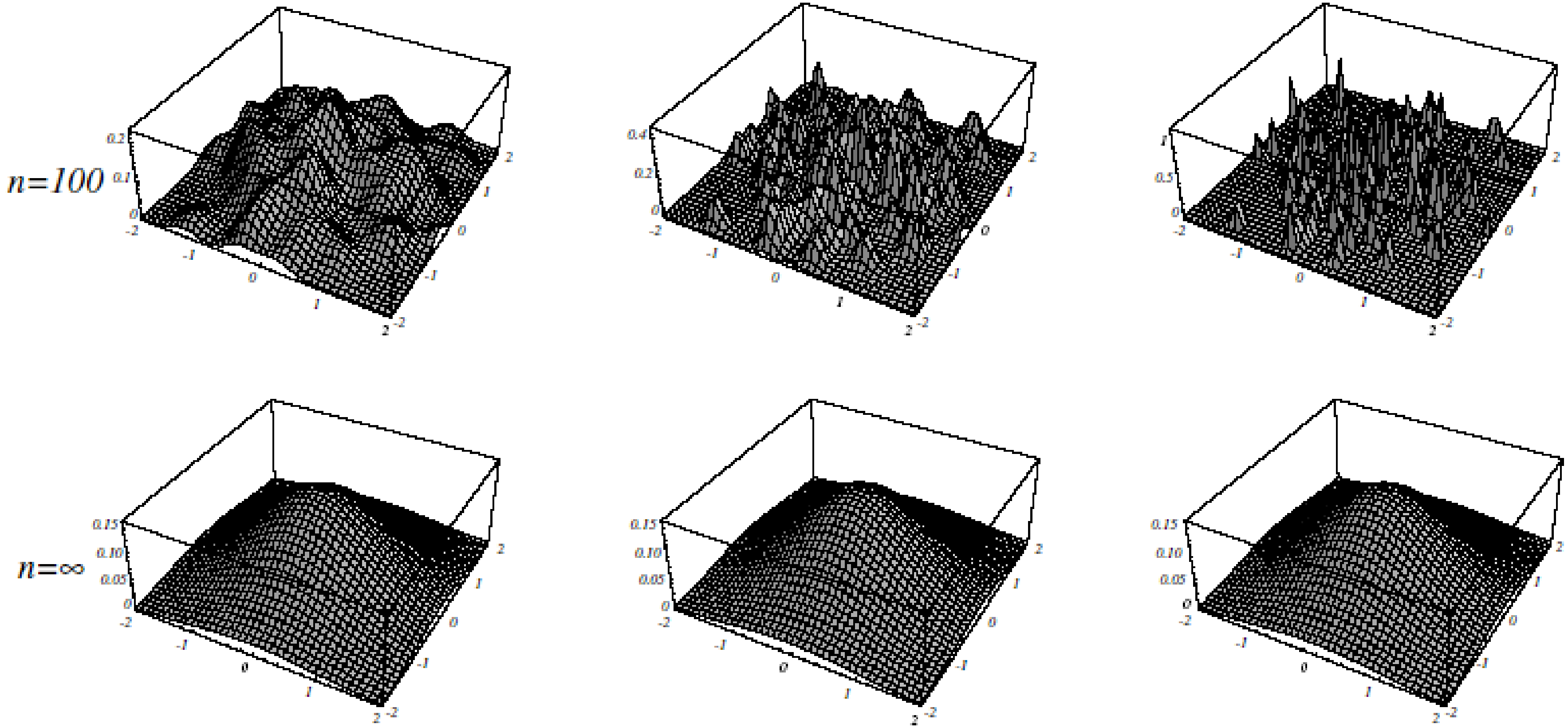
- 采用不同的窗口宽度、不同的样本数量对单变量正态密度进行Parzen估计的结果。各图中的垂直轴已经经过尺度调整。
- 注意，在 n= ∞时，不论窗口宽度如何，各个估计的结果相同(与真实的密度函数相匹配)



二维情况下不同窗口宽度、不同样本数的 Parzen估计：Figure 4.6(a)

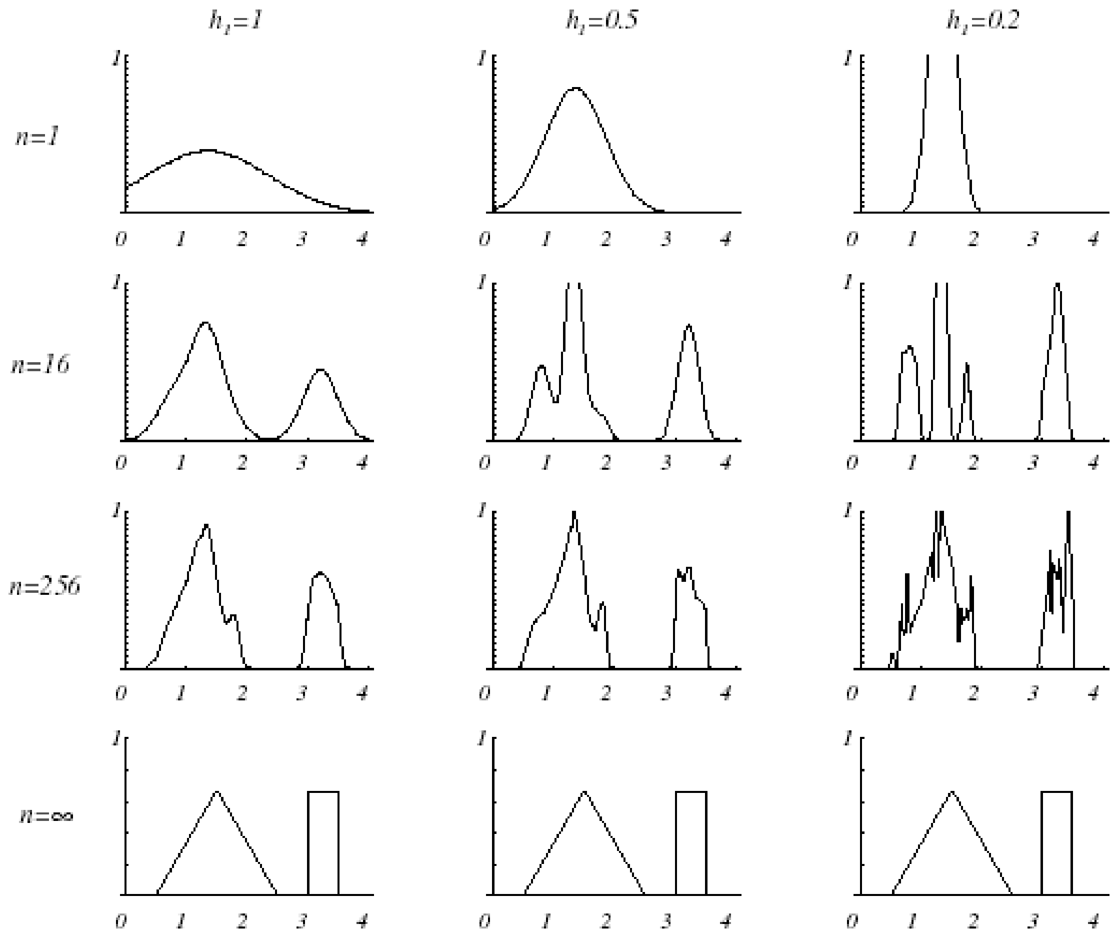


二维情况下不同窗口宽度、不同样本数的Parzen估计： Figure 4.6(b)



Parzen窗估计的另一个例子：Figure4.7

- 采用不同的窗口宽度、不同的样本数对双模混合分布（三角分布和均匀分布）进行Parzen估计的结果。每个图中的垂直轴已经经过尺度调整。
- 注意，在 n= ∞时，不论窗口宽度如何，各个估计的结果相同(与真实的密度函数相匹配)



Parzen窗估计的分类问题

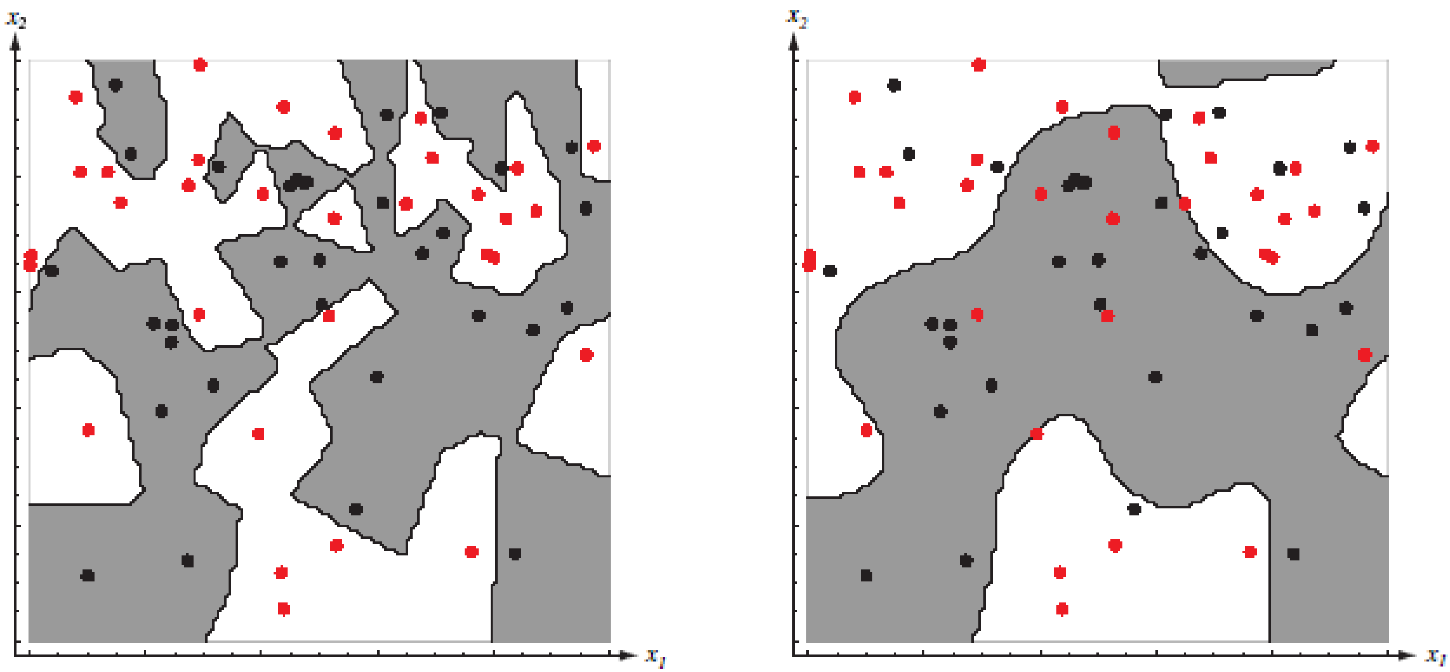
➢ 基于Parzen窗估计的分类器

1. 对每一个类，估计其密度函数，然后根据后验概率最大原则，将待测试样本标记为对应的类。

2. 对Parzen窗分类器的决策区域，决策边界依赖于窗函数的选择（窗函数的形状、大小）。

Parzen窗分类的例子：Figure4.8

- 在一个二维的Parzen窗二分类器中，其决策边界依赖于窗口的宽度 h.
  - 在左上图中的一个小的 h 形成了比右上图中一个大的h 更复杂的分类边界。
  - 对该数据集合，上半区域的数据适合于小的h，而下半区域的数据适合大的h，但对所有数据来说，不存在对所有数据均理想的单一窗口。



Parzen窗分类器的网络实现：概率神经网络 PNN

- 若令$\|\mathbf{x}\| = \mathbf{x}^t\mathbf{x} = 1$，$\|\mathbf{w}_k\| = \mathbf{w}_k^t\mathbf{w}_k = 1$，则

$$(\mathbf{x}-\mathbf{w}_k)^t(\mathbf{x}-\mathbf{w}_k) = \mathbf{x}^t\mathbf{x} + \mathbf{w}_k^t\mathbf{w}_k - 2\mathbf{x}^t\mathbf{w}_k = 2 - 2\mathbf{x}^t\mathbf{w}_k$$

令净输入（投影） $net_k = \mathbf{w}_k^t\mathbf{x}$，则

$$(\mathbf{x}-\mathbf{w}_k)^t(\mathbf{x}-\mathbf{w}_k) = -2(net_k - 1)$$

对于正态密度窗函数

$$e^{-\frac{(\mathbf{x}-\mathbf{w}_k)^t(\mathbf{x}-\mathbf{w}_k)}{2\sigma^2}} = e^{\frac{net_k - 1}{\sigma^2}}$$

概率神经网络PNN

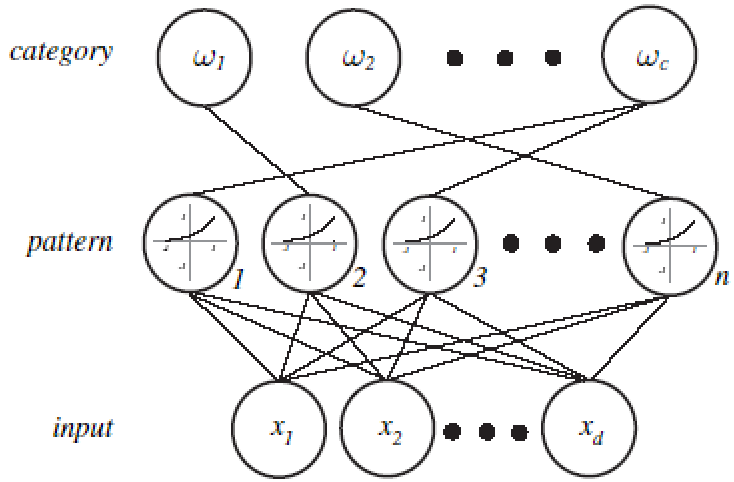
- 概率神经网络(probabilistic neural network, PNN)是由d-维的输入单元、n个模式单元，以及c个分类单元构成的三层网络结构。
- 每个模式单元的输入是其权值矢量和归一化的模式矢量 **x** 的内积 z = **w**^{t}**x** ， 其输出为 exp[(z− 1)/σ^2]。
- 每一个分类单元的功能是计算由连接的各个模式单元输出值的和。
- 这样，每个分类单元的激励代表了 Parzen 窗密度估计，其中窗函数为协方差为$\sigma^2 \boldsymbol{I}$ 的圆对称高斯窗， $\boldsymbol{I}$ 代表 d× d的单位矩阵

第 k 个分类单元的输出为：

$$p_{n_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{1}{\sqrt{2\sigma}} e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}}$$

k = 1,…c，n_k为第k类的训练样本数

$n_1+n_2+\ldots+n_c = n$



非参数技术的特点

- 优点:
  - 方法非常通用，不需要对数据作任何的分布假设。
  - 在样本足够多时，能够保证收敛
- 缺点:
  - 要求的样本数量非常大
  - 特别是当数据的维数增加时，对样本数量的要求呈指数增长，出现所谓的“维数灾难”。