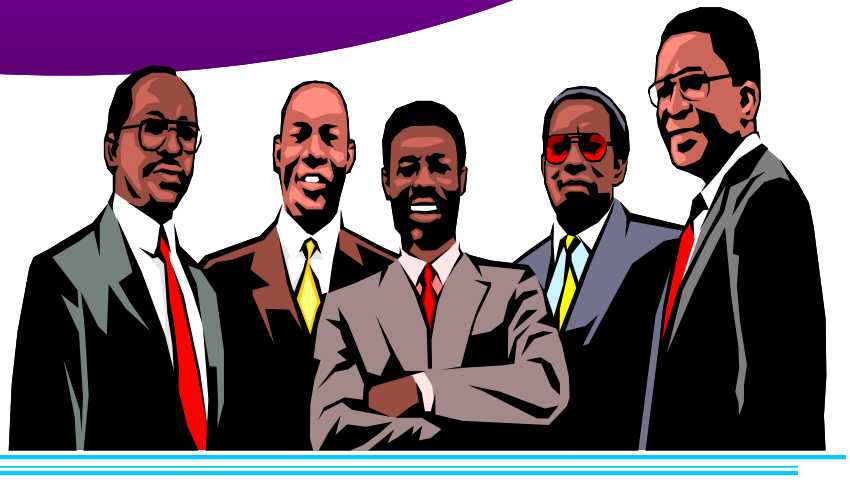
第3章 概率密度函数估计



Outline:

第三章 概率密度函数估计

- 引言
- 最大似然估计
  - 最大似然估计基本原理
  - 最大似然估计的求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

引言

- 贝叶斯决策：已知 $P(\omega_i)$和 $p(\boldsymbol{x}\,|\,\omega_i)$，对未知样本分类（设计分类器）
- 实际问题：已知一定数目的样本，对未知样本分类（设计分类器）

- 怎么办？一种很自然的想法：
  - 首先根据样本估计 $P(\omega_i)$和 $p(\boldsymbol{x}\,|\,\omega_i)$，记 $\hat{\boldsymbol{P}}(\omega_i)$和 $\hat{\boldsymbol{p}}(\boldsymbol{x}\,|\,\omega_i)$
  - 然后用估计的概率密度设计贝叶斯分类器。

-----（基于样本的）两步贝叶斯决策

引言

- 希望：当样本数 $N \to \infty$ 时，如此得到的分类器收敛于理论上的最优解。
- 为此， 需 $\hat{p}(\boldsymbol{x} \mid \omega_i) \xrightarrow{N \to \infty} p(\boldsymbol{x} \mid \omega_i)$

  $\hat{P}(\omega_i) \xrightarrow{N \to \infty} P(\omega_i)$
- 重要前提：
  - ✓ 训练样本的分布能代表样本的真实分布，所谓**i.i.d**条件(独立同分布)
  - ✓ 有充分的训练样本
- 如何利用样本集估计概率密度函数？
- 估计概率密度的两种基本方法：
  - ✓ 参数方法 **(parametric methods)**
  - ✓ 非参数方法 **(nonparametric methods)**

引言

基本概念

- 参数估计(parametric estimation)：
  - ✓ 已知概率密度函数的形式，只是其中几个参数未知，目标是根据样本估计这些参数的值。

- 几个名词：
  - ✓ 统计量(statistics)：样本的某种函数，用来作为对某参数的估计
  - ✓ 参数空间(parametric space)：待估计参数的取值空间 $\theta \in \Theta$
  - ✓ 估计量(estimation)：$\hat{\theta}(x_1, x_2, \cdots, x_N)$

引言

➢参数估计

✓在数理统计学中，似然函数是一种关于统计模型中的参数的函数，表示模型参数中的似然性.

✓“似然性”和“概率”意思相近，但在统计学中，两者又有明确的区分。“概率”用于在已知一些参数的情况下，预测接下来的观测所得到的结果；而“似然性”在是用于在已知某些观测所得到的结果时，对有关事物性质的参数进行估计.

- 概率（probability)和似然（likelihood)，都是指可能性，都可以被称为概率，但在统计应用中有所区别。 概率是给定某一参数值，求某一结果的可能性的函数。

  例如，抛一枚匀质硬币，抛10次，6次正面向上的可能性多大？

  解读：“匀质硬币”，表明参数值是0.5，“抛10次，六次正面向上”这是一个结果，概率（probability)是求这一结果的可能性。

  似然是给定某一结果，求某一参数值的可能性的函数。

  例如，抛一枚硬币，抛10次，结果是6次正面向上，其是匀质的可能性多大？

  解读：“抛10次，结果是6次正面向上”，这是一个给定的结果，问“匀质”的可能性，即求参数值=0.5的可能性。

Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

最大似然估计(Maximum Likelihood Estimation)

假设条件：

- ✓① 参数θ是确定的未知量，（不是随机量）
- ✓② 各类样本集$\chi_i, i=1,\cdots,c$中的样本都是从密度为$p(\boldsymbol{x}\mid\omega_i)$的总体中独立抽取出来的，(独立同分布，**i.i.d.**）
- ✓③ $p(\boldsymbol{x}\mid\omega_i)$具有某种确定的函数形式，只是其中参数$\theta$未知
- ✓④ 各类样本只包含本类分布的信息

其中，参数 θ 通常是向量，比如一维正态分布 $N(\mu_i,\sigma_1^2)$，未知参数可能是 $\theta_i=\begin{bmatrix}\mu_i\\ \sigma_i^2\end{bmatrix}$，此时 $p(\boldsymbol{x}\mid\omega_i)$可写成$p(\boldsymbol{x}\mid\omega_i,\theta_i)$ 或 $p(\boldsymbol{x}\mid\theta_i)$ 。

最大似然估计

某一类样本集 $\mathrm{K} = \{x_1, x_2, \ldots, x_N\}$

似然函数

$$l(\theta) = p(K \mid \theta) = p(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{k=1}^{N} p(x_k \mid \theta)$$

对数似然函数 $$H(\theta) = \ln l(\theta) = \sum_{k=1}^{N} \ln\, p(x_k \mid \theta)$$

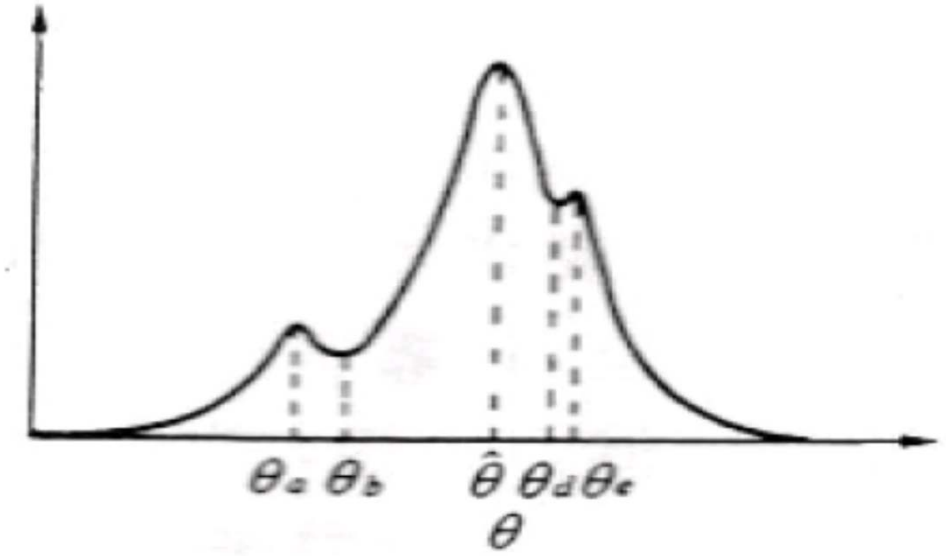
$$\hat{\theta} = \arg\max l(\theta)$$

$$= \arg\max \sum_{k=1}^{N} \ln\, p(x_k \mid \theta)$$

$$\nabla_\theta H(\theta) = 0$$

最大似然估计

！ 求得的满足方程的参数估计值有可能有多个，有的是局部最优解，需要寻找到全局最优解！



具有局部最优解的最大似然估计

最大似然估计

- 讨论：
  - ✓ 如果连续、可微，存在最大值，且上述必要条件方程组有唯一解，则其解就是最大似然估计量。（比如多元正态分布）；
  - ✓ 如果必要条件有多解，则需从中求似然函数最大者
  - ✓ 若不满足条件，则无一般性方法，用何方法？（如均匀分布）

Outline:

- 引言
- 最大似然估计
  - ✓ 最大似然估计基本原理及求解
  - ✓ 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - ✓ Bayes估计
  - ✓ Bayes学习
  - ✓ 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - ✓ 非参数估计基本原理及直方图方法
  - ✓ Parzen窗法
  - ✓ k_N近邻估计方法

最大似然估计示例

多元正态分布 情况一：**Σ**已知, 均值向量**μ**未知

$$p(x_k \mid \theta) = \frac{1}{(2\pi)^{d/2}\left|\sum\right|^{1/2}} \exp\left[-\frac{1}{2}(x_k-\mu)^T \sum^{-1}(x_k-\mu)\right]$$

对数似然函数为：

$$H(\theta) = \sum_{k=1}^{N} \ln p(x_k \mid \theta^i) = \sum_{k=1}^{N} -\frac{1}{2}\ln(2\pi)^d \cdot |\Sigma| - \frac{1}{2}(x_k-\mu)^T \sum^{-1}(x_k-\mu)$$

$$\left.\frac{dH(\theta)}{d\mu}\right|_{\mu=\hat{\mu}} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

即均值向量的最优估计值是训练样本集中所有样本的均值

最大似然估计示例

- 情况二：Σ、μ均未知，一维情形

$$\theta = [\theta_1, \theta_2]^T, \quad \theta_1 = \mu, \quad \theta_2 = \sigma^2$$

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

样本集 $\mathcal{X} = \{x_1, x_2, \cdots, x_N\}$

似然函数 $l(x) = p(\mathcal{X} \mid \theta) = \prod_{k=1}^{N} p(x_k \mid \theta)$

对数似然函数 $H(\theta) = \ln l(x) = \sum_{k=1}^{N} \ln P(x_k \mid \theta)$

最大似然估计示例

最大似然估计量 $\hat{\theta}$ 满足方程：

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \nabla_\theta \ln p(x_k \mid \theta) = 0$$

而

$$\ln p(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta \ln p(x_k \mid \theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2 \end{bmatrix}$$

得方程组

$$\begin{cases} \sum_{k=1}^{N} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \\ -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\theta_2^2} = 0 \end{cases}$$

最大似然估计示例

解得：

$$\hat{\mu} = \hat{\theta}_1 = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \hat{\mu})^2$$

均匀分布示例

- 已知样本集 $K=\{x_1, x_2, \ldots, x_N\}$，$x$ 在区间$[\theta_1, \theta_2]$服从均匀分布，请用最大似然估计求概率分布。

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

$$l(\theta)=\begin{cases}\dfrac{1}{(\theta_2-\theta_1)^N} & \theta_1<x<\theta_2 \\ 0 & \text{otherwise}\end{cases}$$

$$H(\theta)=-N.\ln(\theta_2-\theta_1)$$

均匀分布示例

$$\frac{\partial \boldsymbol{H}}{\partial \theta_1} = \boldsymbol{N}.\frac{1}{\theta_2 - \theta_1} \qquad \frac{\partial \boldsymbol{H}}{\partial \theta_2} = \boldsymbol{N}.\frac{-1}{\theta_2 - \theta_1}$$

$$\boldsymbol{x}' = \min\left\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\right\}$$

$$\boldsymbol{x}'' = \max\left\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\right\}$$

$$\hat{\theta}_1 = \boldsymbol{x}' \qquad \hat{\theta}_2 = \boldsymbol{x}''$$

Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - k_N近邻估计方法
  - Parzen窗法

贝叶斯估计

最大似然估计：

把待估计的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；

贝叶斯估计：

把待估计的参数本身也看作是随机变量，要做的是根据观测数据对参数的分布进行估计。

贝叶斯估计

某一类样本集 $K=\{x_1,x_2,\ldots,x_N\}$

贝叶斯估计：

把未知参数Θ作为具有某种先验分布密度P(Θ)的随机变量，通过对样本的观察，使先验分布转化为后验分布p(Θ|K)，再修正原先对参数的估计。

从决策的角度研究估计问题：

贝叶斯决策：

未知参数类别ω，先验分布P(ω)，通过对样本的观察，使先验分布转化为后验分布p(ω|x)

贝叶斯估计

损失函数：把θ 估计为$\hat{\theta}$ 所造成的损失，记为$\lambda(\hat{\theta},\theta)$

期望风险：
$$R=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\boldsymbol{x},\theta)d\theta d\boldsymbol{x}$$
$$=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\mid\boldsymbol{x})p(\boldsymbol{x})d\theta d\boldsymbol{x}$$
$$=\int_{E^d}R(\hat{\theta}\mid\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

其中，$\boldsymbol{x}\in\boldsymbol{E}^{\boldsymbol{d}}$，$\theta\in\Theta$

条件风险：
$$R(\hat{\theta}\mid\boldsymbol{x})=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\mid\boldsymbol{x})d\theta \qquad \boldsymbol{x}\in E^d$$

给定条件下，估计量的期望损失

贝叶斯估计

最小化期望风险 $\Rightarrow$ 最小化条件风险 （对所有可能的 $\boldsymbol{x}$）

有限样本集下，最小化经验风险：

$$R(\hat{\theta} \mid \mathcal{X}) = \int_{\Theta} \lambda(\hat{\theta}, \theta)\, p(\theta \mid \mathcal{X}) d\theta$$

贝叶斯估计量：（在样本集 $\mathcal{X}$ 下）使条件风险（经验风险）最小的估计量 $\hat{\theta}$

损失： 离散情况：损失函数表（决策表）； 连续情况：损失函数

常用的损失函数： $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ （平方误差损失函数）

贝叶斯估计

定理3.1

如果采用平方误差损失函数，则 θ 的贝叶斯估计量 $\hat{\theta}$ 是在给定 $\boldsymbol{x}$ 时 θ 的条件期望，即 $\hat{\theta} = E[\theta \mid \boldsymbol{x}] = \int_{\Theta} \theta p(\theta \mid \boldsymbol{x}) d\theta$

同理可得，在给定样本集 $\mathcal{X}$ 下，θ 的贝叶斯估计是：

$$\hat{\theta} = E[\theta \mid \mathcal{X}] = \int_{\Theta} \theta p(\theta \mid \mathcal{X}) d\theta$$

贝叶斯估计

求贝叶斯估计的方法：（平方误差损失下）

（1）确定 θ 的先验分布 $p(\theta)$

（2）求样本集的联合分布

$$p(\mathcal{X}\mid\theta)=\prod_{i=1}^{N} p(\boldsymbol{x}_i\mid\theta)$$

（3）求 θ 的后验概率分布

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

（4）求 θ 的贝叶斯估计量

$$\hat{\theta}=E[\theta\mid\mathcal{X}]=\int_{\Theta}\theta p(\theta\mid\mathcal{X})d\theta$$

贝叶斯估计

也可直接推断总体分布

$$p(\boldsymbol{x}\,|\,\mathcal{X}) = \int_{\Theta} p(\boldsymbol{x}\,|\,\theta)\,p(\theta\,|\,\mathcal{X})\,d\theta$$

其中，
$$p(\theta\,|\,\mathcal{X}) = \frac{p(\mathcal{X}\,|\,\theta)\,p(\theta)}{\int_{\Theta} p(\mathcal{X}\,|\,\theta)\,p(\theta)\,d\theta}$$

理解： 参数θ是随机变量，它有一定的分布，而要估计的概率密度$p(\boldsymbol{x}\,|\,\mathcal{X})$就是所有可能的参数取值下的样本概率密度的加权平均，而这个加权就是在观测样本下估计出的参数θ的后验概率。

贝叶斯估计

$$p(\boldsymbol{x}\,|\,\mathcal{X}) = \int_{\Theta} p(\boldsymbol{x}\,|\,\theta)\,p(\theta\,|\,\mathcal{X})\,d\theta$$

$$p(\theta\,|\,\mathcal{X}) \sim p(\mathcal{X}\,|\,\theta)\,p(\theta)$$

设 θ 的最大似然估计为 $\hat{\theta}_l$，则在 $\theta = \hat{\theta}_l$ 处 $p(\theta\,|\,\mathcal{X})$ 很可能有一尖峰，若如此，且先验概率 $p(\theta)$ 在 $\hat{\theta}_l$ 处非零且在附近变化不大，则

$$p(\boldsymbol{x}\,|\,\mathcal{X}) \doteq p(\boldsymbol{x}\,|\,\hat{\theta}_l)\,,$$

即贝叶斯估计结果与最大似然估计结果近似相等。

如 $p(\theta\,|\,\mathcal{X})$ 的峰值不尖锐，则不能用最大似然估计来代替贝叶斯估计。

Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

Bayes学习

考虑估计的收敛性：记学习样本个数 N，样本集 $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$

$N>1$时有 $$p(\mathcal{X}^N\mid\theta)=p(\boldsymbol{x}_N\mid\theta)p(\mathcal{X}^{N-1}\mid\theta)$$

因此有递推后验概率公式：

$$p(\theta\mid\mathcal{X}^N)=\frac{p(\boldsymbol{x}_N\mid\theta)p(\theta\mid\mathcal{X}^{N-1})}{\int p(\boldsymbol{x}_N\mid\theta)p(\theta\mid\mathcal{X}^{N-1})d\theta}$$

设 $p(\theta\mid\mathcal{X}^0)=p(\theta)$，

则随着样本数增多，可得后验概率密度函数序列：

$$p(\theta),\ p(\theta\mid\boldsymbol{x}_1),\ p(\theta\mid\boldsymbol{x}_1,\boldsymbol{x}_2),\cdots$$

—— 参数估计的递推贝叶斯方法

如果此序列收敛于以真实参数值为中心的 δ 函数，则把这一性质称作贝叶斯学习。

此时 $$p(\boldsymbol{x}\mid\mathcal{X}^{N\to\infty})=p(\boldsymbol{x}\mid\hat{\theta}=\theta)=p(\boldsymbol{x})\ 。$$

Bayes学习

- 后验分布理解：
  - ✓ 后验分布的意义在于综合了关于 θ 的先验信息（反映在先验分布p(θ)中）和样本X关于 θ 的信息（反映在样本分布p(X|θ)中）。先验分布概括了在试验前对 θ 的认识，而得到样本观测值X之后，对 θ 的认识有了深化，这集中反映在后验分布中。Bayes公式反映了先验分布到后验分布的转化，即Bayes自己所说的“归纳推理”的统计方法。

Bayes统计推断的原则

- 样本X的作用
  - ✓ 对Bayes统计而言，样本X的唯一作用在于对θ的认识由先验分布转化成后验分布。
- Bayes统计推断的原则
  - ✓ 对参数θ所作的任何推断(估计、检验等)必须基于且只能基于θ的后验分布。
- 对原则的理解
  - ✓ 一经由样本X算出了θ的后验分布，就设想我们除了这一后验分布外，其余的东西(样本值、样本分布、先验分布)全忘记了。这时，对θ的推断的唯一凭证就是这一后验分布。

Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

正态分布的Bayes估计

一维， $p(x|\mu)\sim N(\mu,\sigma^2)$， σ^2 已知，估计 μ 。假设先验分布 $p(\mu)\sim N(\mu_0,\sigma_0^2)$

结论： $$\hat{\mu}=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\boldsymbol{m}_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$ 其中 $$m_N=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

----- 样本信息与先验知识的线性组合

讨论：

$N=0$ 时， $\hat{\mu}=\mu_0$； $N\to\infty$ 时， $\hat{\mu}\to m_N$

若 $\sigma_0^2=0$， 则 $\hat{\mu}\equiv\mu_0$ （先验知识可靠，样本不起作用）

若 $\sigma_0>>\sigma$，则 $\hat{\mu}=m_N$ （先验知识十分不确定，完全依靠样本信息）

正态分布的Bayes估计

μ 的密度：

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

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

当 $N \to \infty$ 时，$\sigma_N^2 \to 0$，$p(\mu \mid X) \to \delta$ 函数。

正态分布的Bayes估计

$$p(\boldsymbol{x}\,|\,\mathcal{X}^N)=\int p(\mu\,|\,\mathcal{X}^N)p(\boldsymbol{x}\,|\,\mu)d\mu$$

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

均值 μ_N，方差由 σ^2 增为 $\sigma^2+\sigma_N^2$ ----- 由于用了 μ 的估计值而不确定性增加

Maximal Likelihood vs. Bayesian

- ML and Bayesian estimations are asymptotically equivalent and "consistent". They yield the same class-conditional densities when the size of the training data grows to infinity.
- ML is typically computationally easier: in ML we need to do (multidimensional) differentiation and in Bayesian (multidimensional) integration.
- ML is often easier to interpret: it returns the single best model (parameter) whereas Bayesian gives a weighted average of models.
- But for a finite training data (and given a reliable prior) Bayesian is more accurate (uses more of the information).

Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

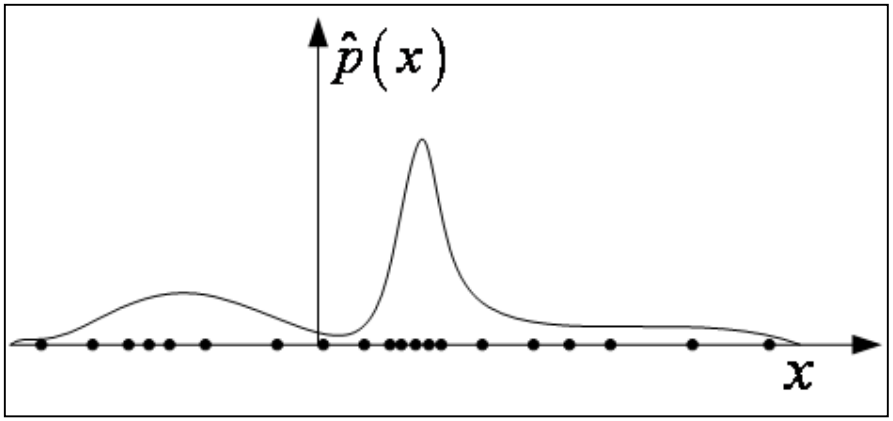
概率密度估计的非参数方法

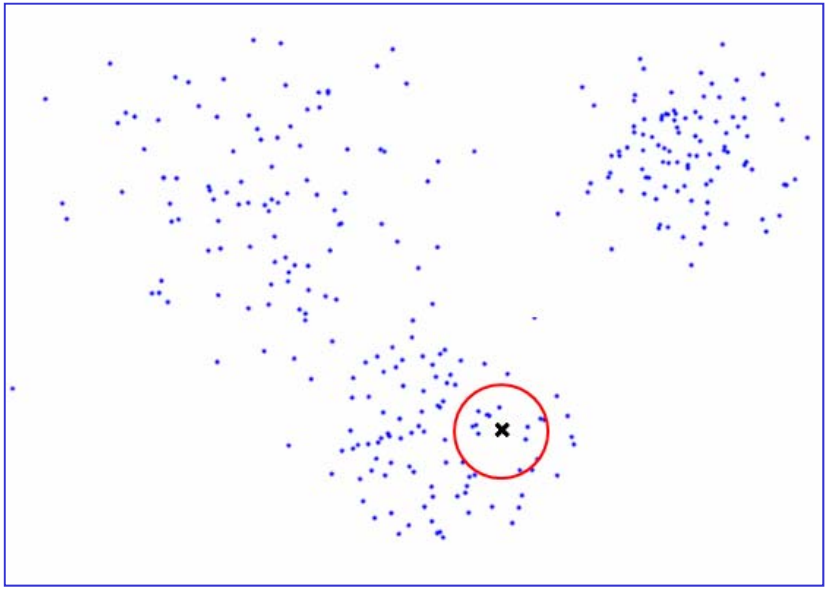
参数估计 **parametric estimation**

非参数估计 **nonparametric estimation**

给定i.i.d.样本集： $X=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_l\}$

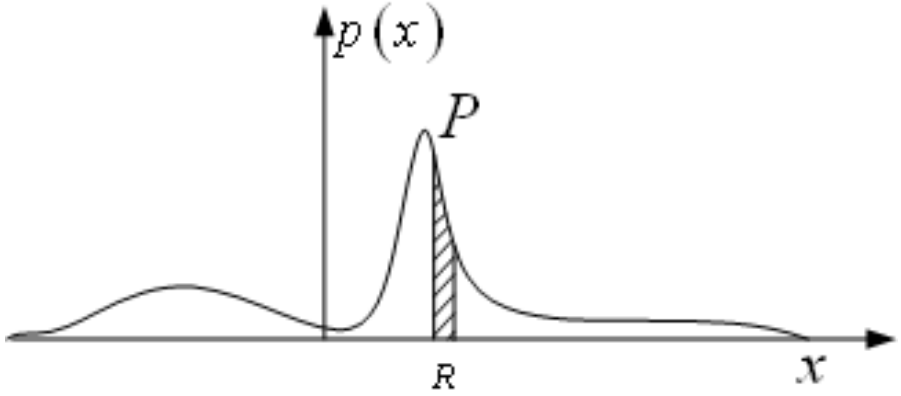
估计概率分布： $p(\mathbf{x})$





概率密度估计

- 非参数概率密度估计的核心思路：



一个向量x落在区域R中的概率P为：

$$P = \int_R p(\mathbf{x})d\mathbf{x}$$

因此，可以通过统计概率P来估计概率密度函数p(x)

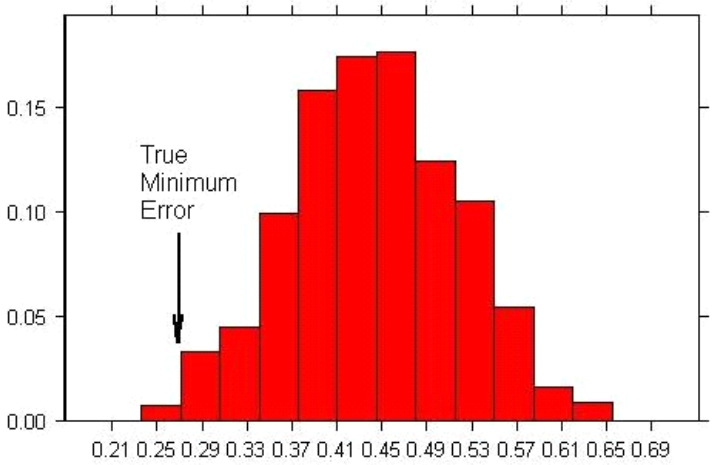
概率密度估计的非参数方法

直方图方法

非参数概率密度估计的最简单方法

（1）把 x 的每个分量分成 k 个等间隔小窗，（若 $x \in E^d$，则形成 $\boldsymbol{k}^d$ 个小舱）

（2）统计落入各个小舱内的样本数 q_i



（3）相应小舱内的概率密度为 $q_i/(NV)$（N:样本总数，V: 小舱体积）

非参数估计基本原理

问题：已知样本集 $\mathcal{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$，其中样本均从服从 $p(\boldsymbol{x})$ 的总体中独立抽取，求估计 $\hat{p}(\boldsymbol{x})$，近似 $p(\boldsymbol{x})$。

考虑随机向量 $\boldsymbol{x}$ 落入区域 $\Re$ 的概率 $P_R=\int_{\Re}p(x)dx$

$\mathcal{X}$ 中有 k 个样本落入区域 $\Re$ 的概率 $P_k=C_N^kP_{\Re}^k(1-P_R)^{N-k}$

k 的期望值 $E[k]=NP_R$

k 的众数（概率最大的取值）为 $m=[(N+1)P_R]$

P_R 的估计 $\hat{P}_R=\frac{k}{N}$（k：实际落到 $\Re$ 中的样本数）

非参数估计基本原理

设 $p(x)$ 连续，且 $\Re$ 足够小，$\Re$ 的体积为 V，则有

$$P_R = \int_R p(x)dx = p(x)V \qquad x \in \Re$$

因此

$$\hat{p}(x) = \frac{k}{NV}$$

其中，

N：样本总数，

V：包含 x 的一个小区域的体积

k：落在此区域中的样本数

$\hat{p}(x)$ 为对 $p(x)$ 在小区域内的平均值的估计。

非参数估计基本原理

- 当样本数量N固定时，体积V的大小对估计的效果影响很大。
  - ✓ 过大则平滑过多，不够精确；
  - ✓ 过小则可能导致在此区域内无样本点，k=0。

- 此方法的有效性取决于样本数量的多少，以及区域体积选择的合适。

非参数估计基本原理

- **收敛性问题：样本数量N无穷大时，估计的概率函数是否收敛到真实值？**

$$\lim_{N\to\infty}\hat{p}_N(\mathbf{x})=p(\mathbf{x})$$

实际中，$\hat{p}(\mathbf{x})$ 越精确，要求：$R\to 0$

实际中，N是有限的：

当 $R\to 0$ 时，绝大部分区间没有样本：$\hat{p}(\mathbf{x})=0$

如果侥幸存在一个样本，则： $\hat{p}(\mathbf{x})=\infty$

非参数估计基本原理

➢理论结果：

设有一系列包含x 的区域$\mathbf{R_1}$，$\mathbf{R_2}$，…,$\mathbf{R_n}$,…，对$\mathbf{R_1}$采用1个样本进行估计，对$\mathbf{R_2}$用2 个，…， $\mathbf{R_n}$包含$\mathbf{k_n}$个样本。$\mathbf{V_n}$为$\mathbf{R_n}$的体积。

$$p_n(\mathbf{x}) = \frac{k_n / N}{V_n}$$

为p(x)的第n次估计

非参数估计基本原理

- 如果要求 $p_n(\mathbf{x})$ 能够收敛到p(x)，那么必须满足：

$$\lim_{n\to\infty} V_n = 0 \qquad \lim_{n\to\infty} k_n = \infty \qquad \lim_{n\to\infty} k_n / n = 0$$

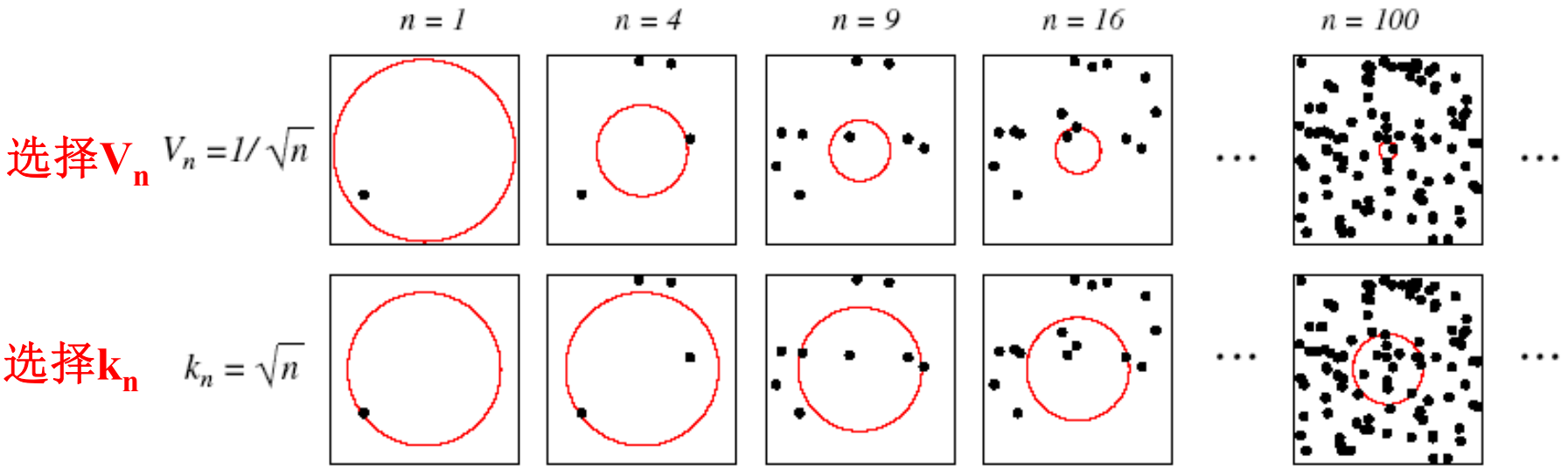
随着样本数的增加：

小舱体积应尽可能小；

保证小舱内有充分多的样本；

每个小舱内的样本数又必须是总样本数中很少的一部分。

非参数估计基本原理



估计某一点处的概率密度函数有两种最基本的方法。这里，我们假设这个点位于图中所示的正方形的中心。第一行表示的方法是从一个以目标样本点为中心的较大的区域开始，根据某个函数，例如 $V_n = 1/\sqrt{n}$，逐渐的缩小区域面积。第二种方法如第二行所示。这一方法缩小区域面积的方式是依赖于样本点的。例如，令区域必须包括 $k_n = \sqrt{n}$ 个样本点。这两种情况中的序列都是随机变量，它们一般会收敛，这样就能估计出测试样本点处的真正的概率密度函数

非参数估计基本原理

两种选择策略：

1. 选择V_n，（比如$V_n = \frac{1}{\sqrt{n}}$），同时对k_n和$\frac{k_n}{n}$加限制以保证收敛

—— Parzen 窗法

使区域序列V_n以n的某个函数的关系不断缩小

2. 选择k_n，（比如$k_n = \sqrt{n}$），V_n为正好包含x的k_n个近邻

—— k_N近邻估计

让k_n为n的某个函数

Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

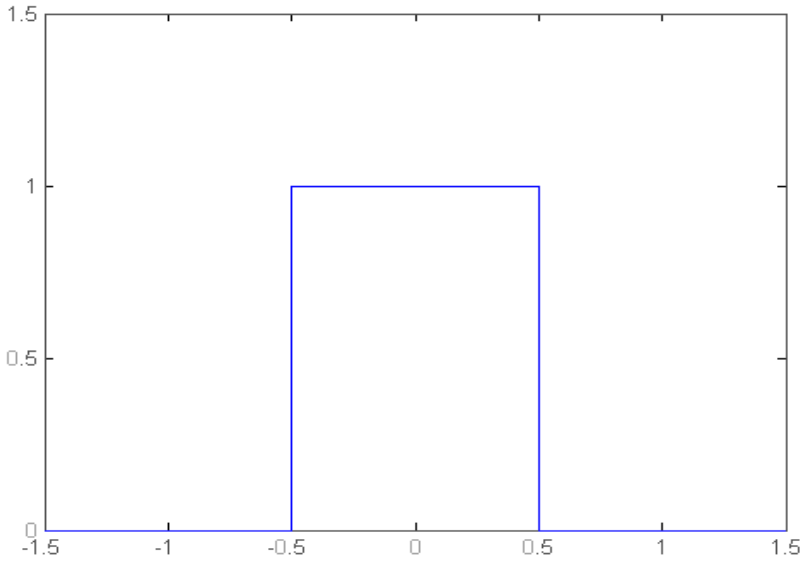
Parzen窗估计

- 定义窗函数：假设$\mathbf{R_n}$是一个d维的超立方体。令$\mathbf{h_n}$为超立方体一条边的长度，则体积：

$$V_n = h_n^d$$

立方体窗函数为：

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



中心在原点的单位超立方体：

Parzen窗估计

落入以**x**为中心的立方体区域的样本数为：

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

x处的密度估计为：

$$\hat{p}_n(\mathbf{x}) = \frac{k_n / n}{Vn} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

可以验证：　$\hat{p}_n(\mathbf{x}) \geq 0$　$\int \hat{p}_n(\mathbf{x})d\mathbf{x} = 1$

Parzen 窗法

- 估计量 $\hat{p}_n(x)$ 为密度函数的条件
  - ✓非负性显然

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

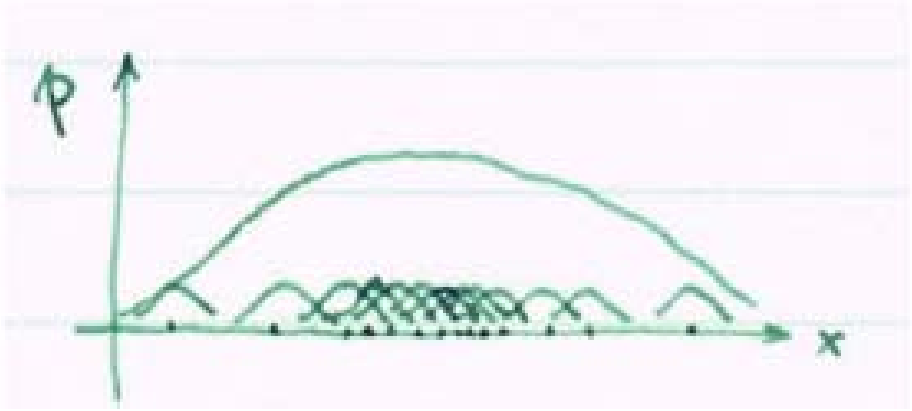
$$\int \hat{p}_N(x)dx=\int\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{x-x_i}{h_N}\right)dx$$

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

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

窗函数的要求

- **Parzen**窗估计过程是一个内插过程，样本x_i距离x越近，对概率密度估计的贡献越大，越远贡献越小。

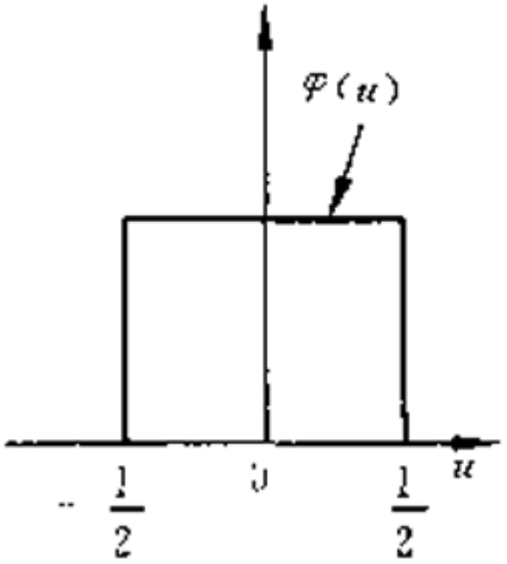


- 只要满足如下条件，就可以作为窗函数：

$$\varphi(\mathbf{u}) \geq 0$$

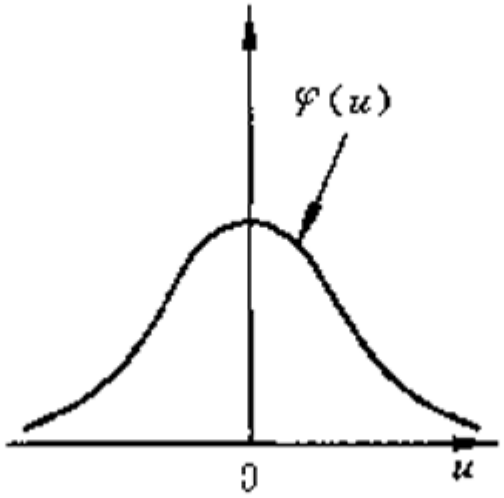
$$\int \varphi(\mathbf{u}) d\mathbf{u} = 1$$

窗函数的形式



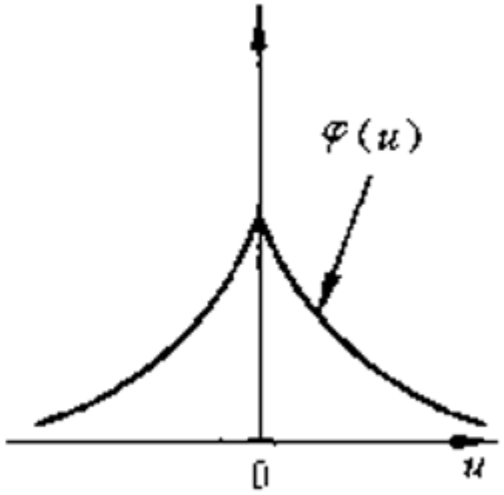
方窗函数

$$\varphi(u)=\begin{cases}1, |u| \le \dfrac{1}{2} \\ 0.\text{其他}\end{cases}$$



正态窗函数

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



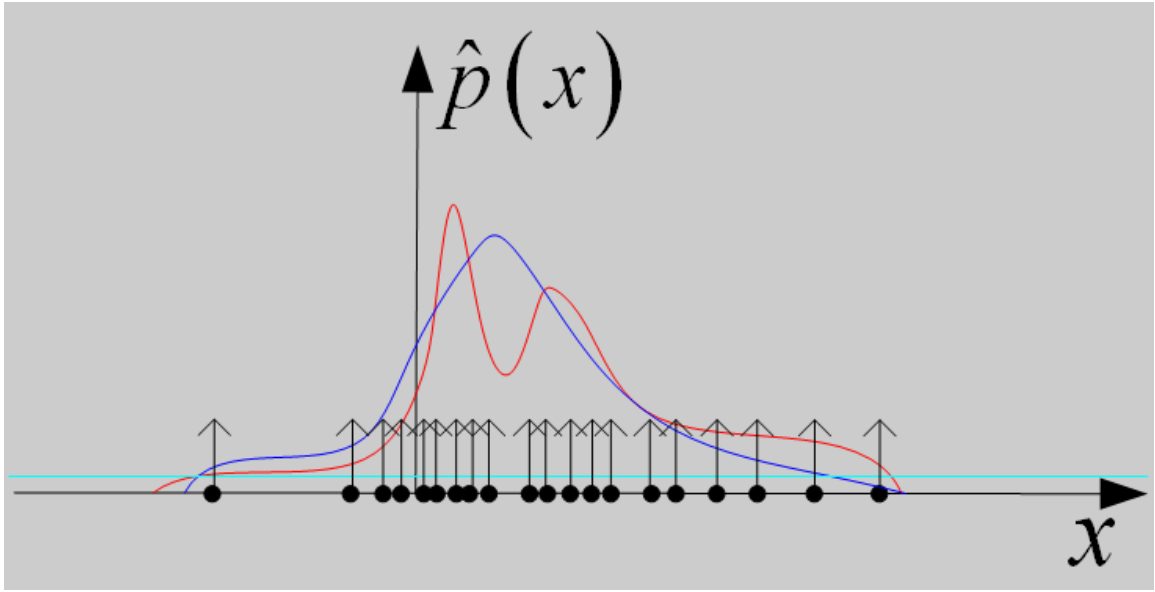
指数窗函数

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

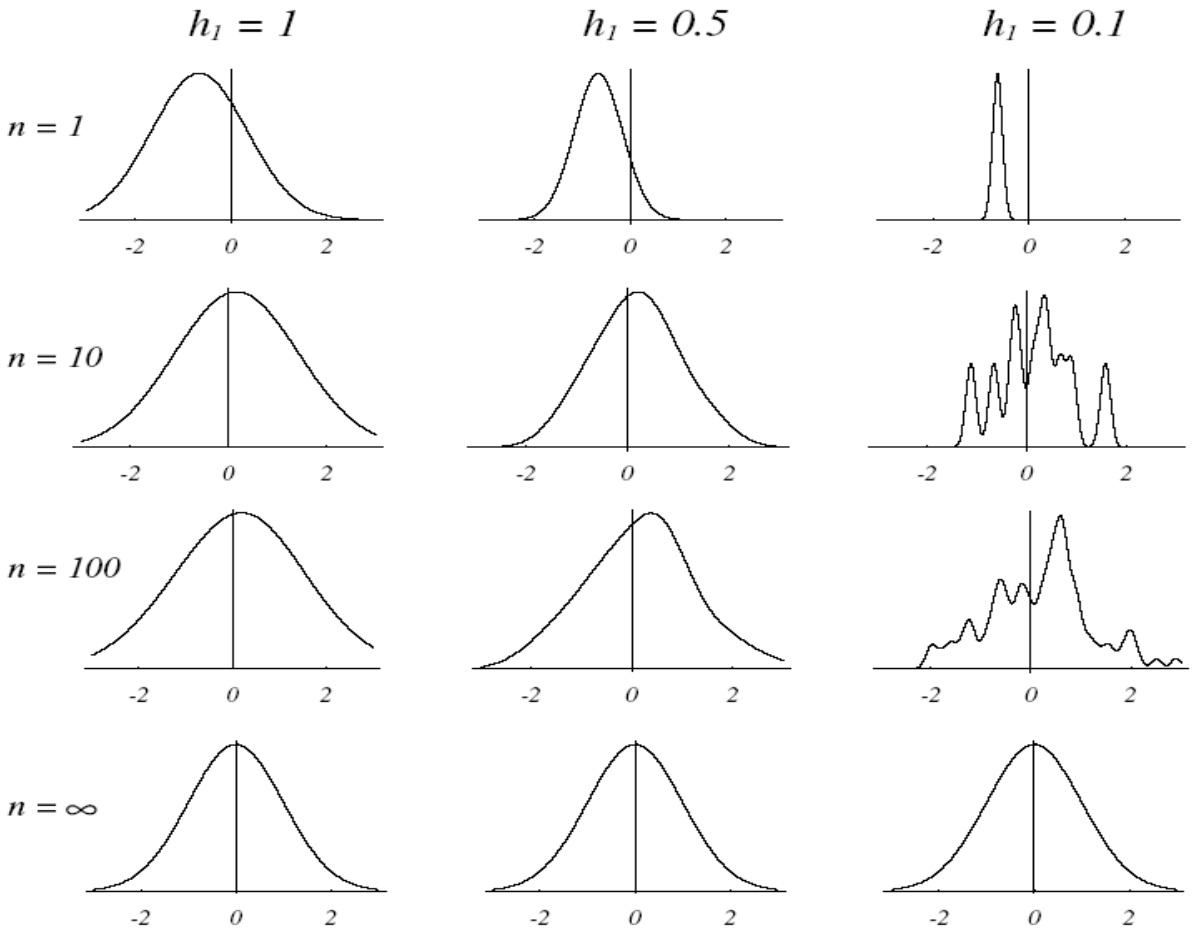
其中：$u=\dfrac{x-x_i}{h_n}$

窗口宽度的影响

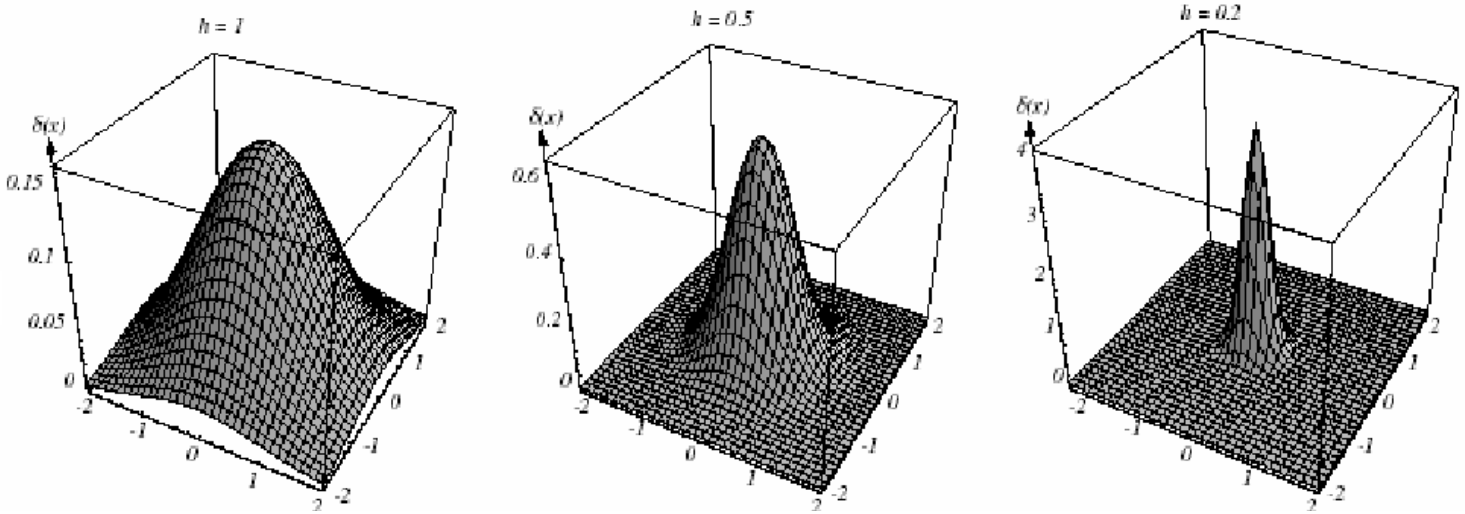
- **Parzen估计的性能与窗宽参数h_n紧密相关**
  - ✓ 当h_n较大时，x和中心x_i距离大小的影响程度变弱，估计的p(x)较为平滑，分辨率较差。
  - ✓ 当h_n较小时，x和中心x_i距离大小的影响程度变强，估计的p(x)较为尖锐，分辨率较好。



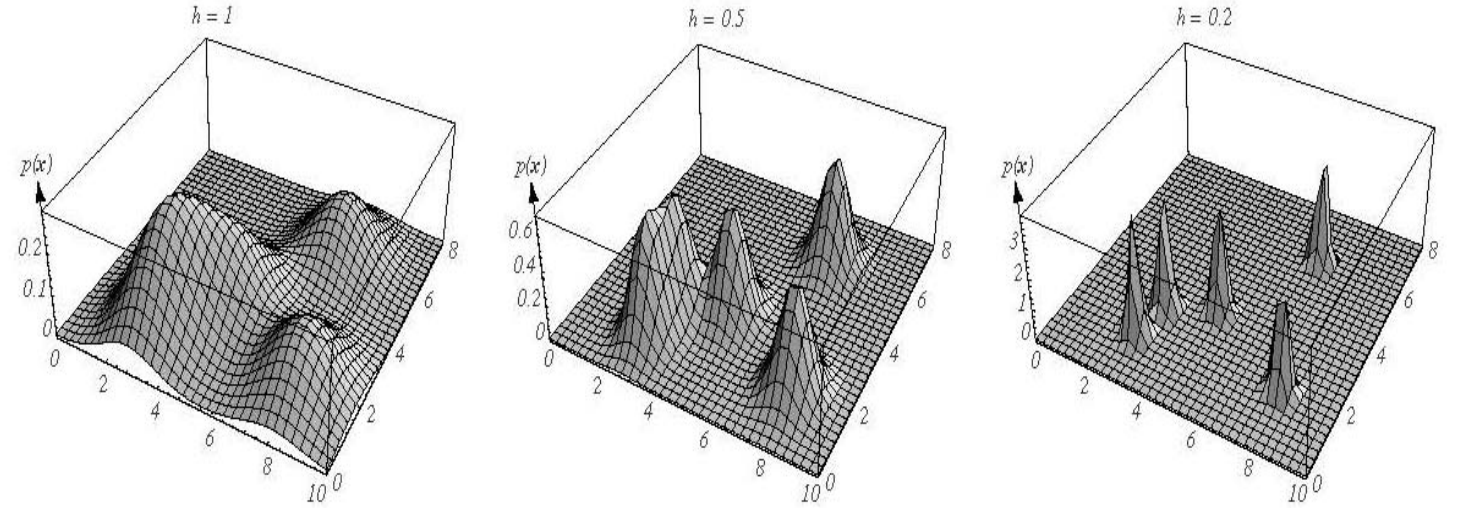
窗口宽度的影响



5个样本的Parzen窗估计：



窗函数



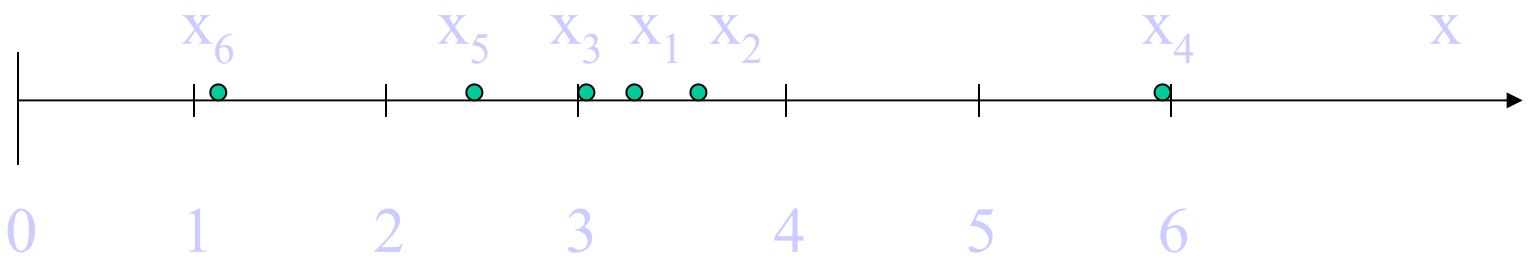
密度估计值

例：对于一个二类（ ω_1 ， ω_2 ）识别问题，随机抽取ω_1类的6个样本X=(x_1，x_2，….，x_6)

ω_1=(x_1，x_2，….，x_6)

=(x_1=3.2，x_2=3.6，x_3=3，x_4=6，x_5=2.5，x_6=1.1)

估计P(x|ω_1)即$P_N(x)$



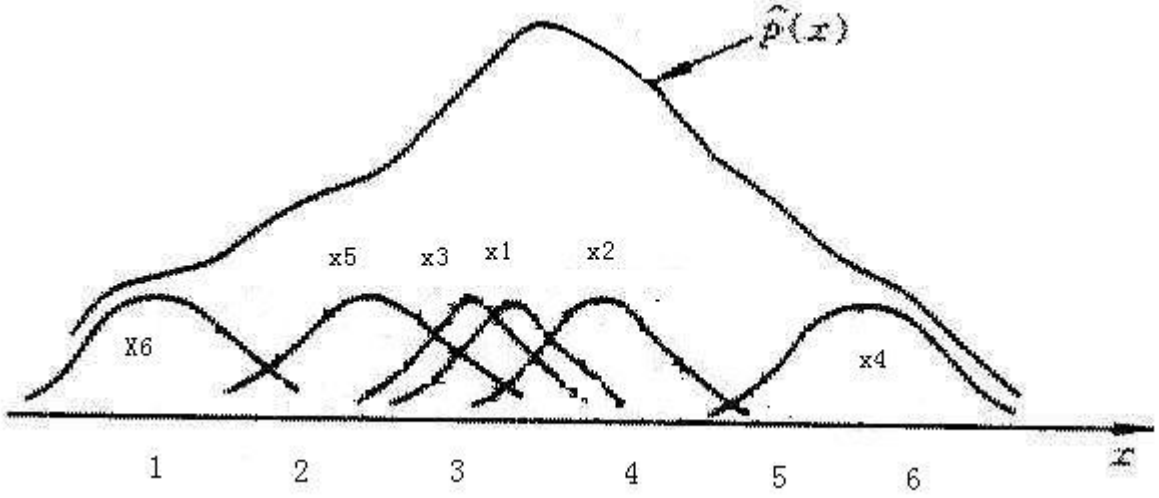
解：选正态窗函数 $\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$

$$\therefore \phi(u)=\phi(\frac{|x-x_i|}{h_N})=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{|x-x_i|}{h_N}\right)^2\right]$$

$\because$ x是一维的　　$\therefore V_N = h_N = \frac{h_1}{\sqrt{N}}$，其中选 $h_1 = 0.5\sqrt{6}$，$N = 6$

$\therefore V_N = h_N = \frac{0.5\sqrt{6}}{\sqrt{6}} = 0.5$　　代入：$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right)$

上式用图形表示是6个分别以3.2，3.6，3，6，2.5，1.1为中心的正态曲线，而$P_N(x)$则是这些曲线之和。



由图看出，每个样本对估计的贡献与样本间的距离有关，样本越多，$P_N(x)$越准确。

例：设待估计的p(x)是个均值为0，方差为1的正态密度函数。若随机地抽取X样本中的1个、 16个、 256个作为学习样本x_i,试用窗口法估计$P_N(x)$。

解：设窗口函数为正态的， σ＝1， μ ＝0

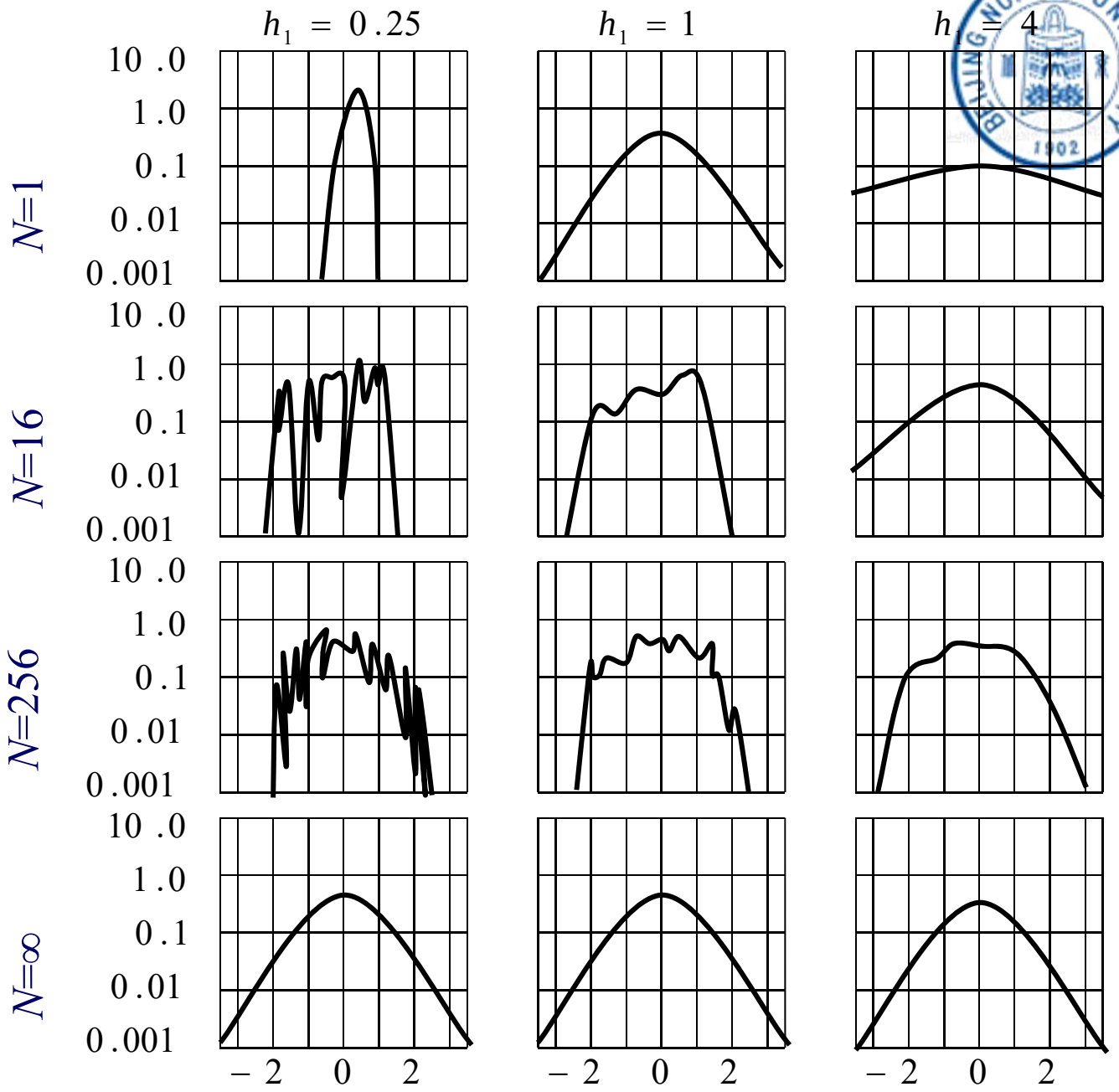
$$\varphi(\frac{|x-x_i|}{h_N})=\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{|x-x_i|}{h_N})^2]$$

$$h_N=\frac{h_1}{\sqrt{N}} \qquad V_N=h_N$$

h_N:窗长度，N为样本数，h1为选定可调节的参数。

$$P_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi(\frac{|x-x_i|}{h_N})=\frac{1}{h_1\sqrt{N}}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}\left(\frac{|x-x_i|\sqrt{N}}{h_1}\right)^2]$$

用 *Parzen* 窗法估计单一正态分布的实验



由图看出, $P_N(x)$随N, h_1的变化情况

①当N＝1时， $P_N(x)$是一个以第一个样本为中心的正态曲线，与窗函数差不多。

②当N＝16及N=256时

h_1＝0.25　曲线起伏很大，噪声大

h_1＝1　起伏减小

h_1＝4　曲线平坦

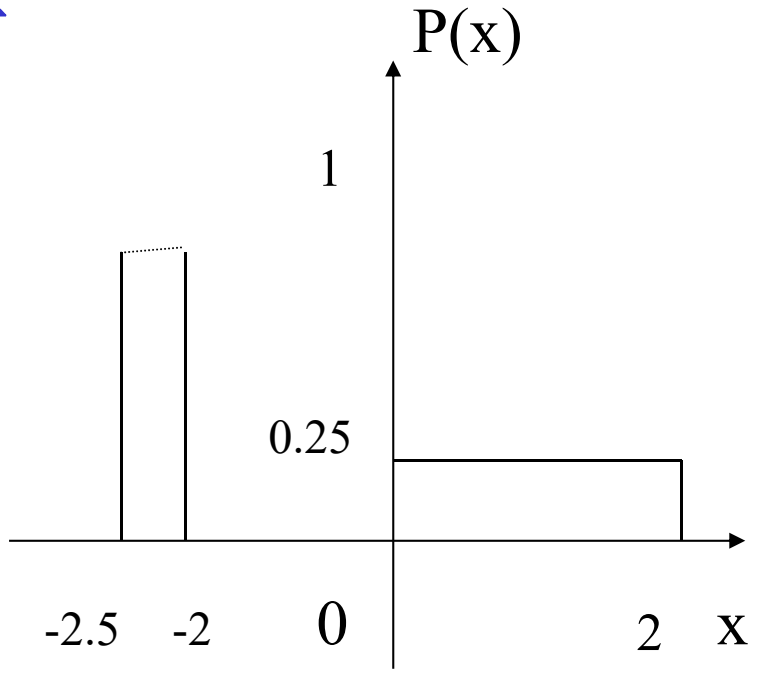
③当N→∞时， $P_N(x)$收敛于一平滑的正态曲线，估计曲线较好。

例：待估的密度函数为二项分布

解：此为多峰情况的估计

设窗函数为正态

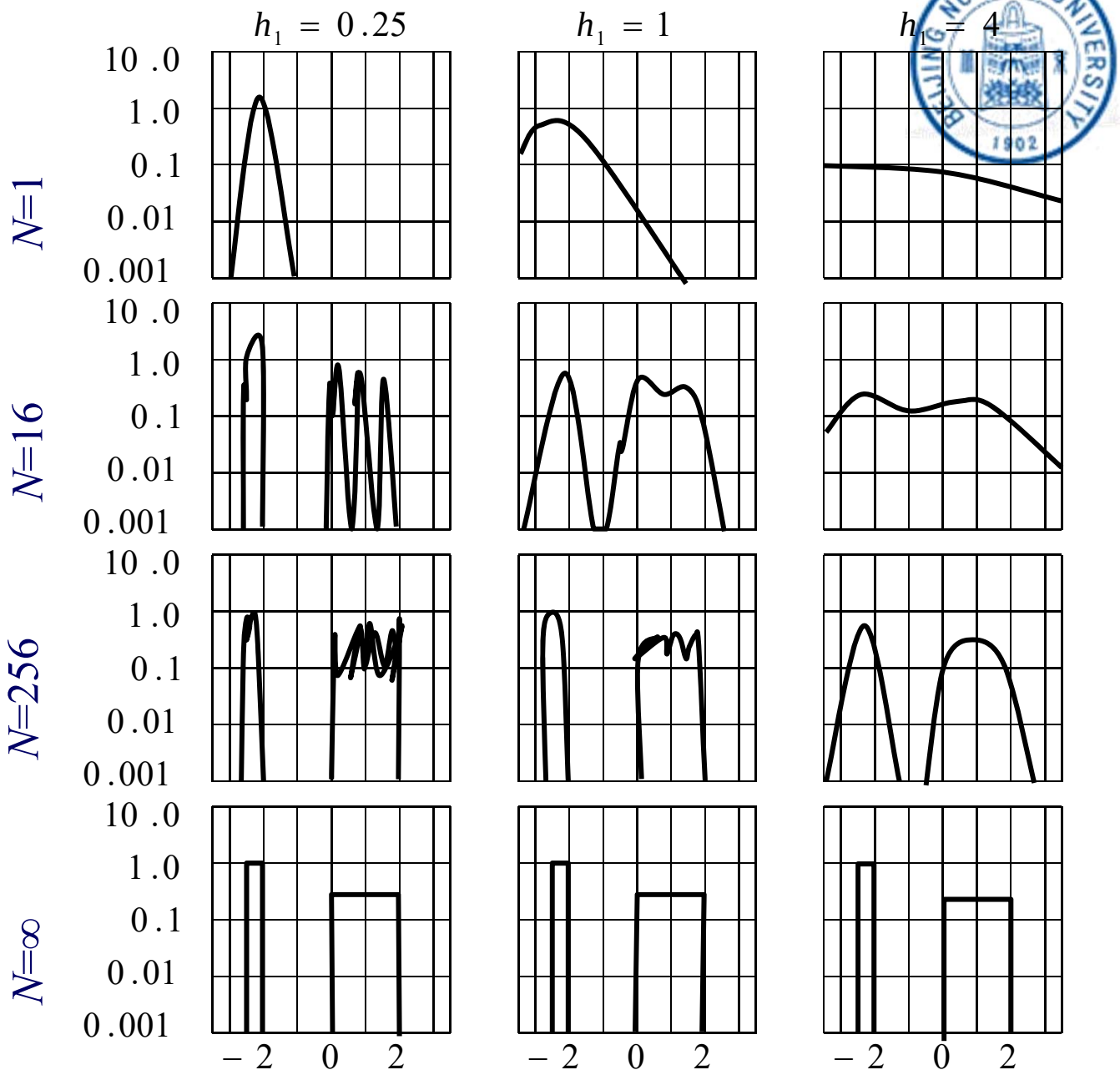
$$P(x)=\begin{cases}1 & -2.5<x<-2\\ 0.25 & 0<x<2\\ 0 & x\text{为其它}\end{cases}$$



解：此为多峰情况的估计

设窗函数为正态 $\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}u^2], h_N=\frac{h_1}{\sqrt{N}}$

- 用*Parzen*窗法估计两个均匀分布的实验



当N=1、16、256、 ∞时的$P_N(x)$估计如图所示

①当N＝1时， $P_N(x)$ 实际是窗函数。

②当N＝16及N=256时

h_1＝0.25 曲线起伏大

h_1＝1 曲线起伏减小

h_1＝4 曲线平坦

③当N→∞时，曲线较好。

Parzen窗估计

- 优点
  - ✓ 由前面的例子可以看出， **Parzen窗估计的优点是应用的普遍性。**对规则分布，非规则分布，单锋或多峰分布都可用此法进行密度估计。
  - ✓ 可以获得较为光滑且分辨率较高的密度估计，实现了光滑性和分辨率之间的一个较好平衡。
- 缺点
  - ✓ 要求样本足够多，才能有较好的估计。因此使计算量，存储量增大。
  - ✓ 窗宽在整个样本空间固定不变，难以获得区域自适应的密度估计。

识别方法

1. 保存每个类别所有的训练样本；
2. 选择窗函数的形式，根据训练样本数n选择窗函数的h宽度；
3. 识别时，利用每个类别的训练样本计算待识别样本x的类条件概率密度：

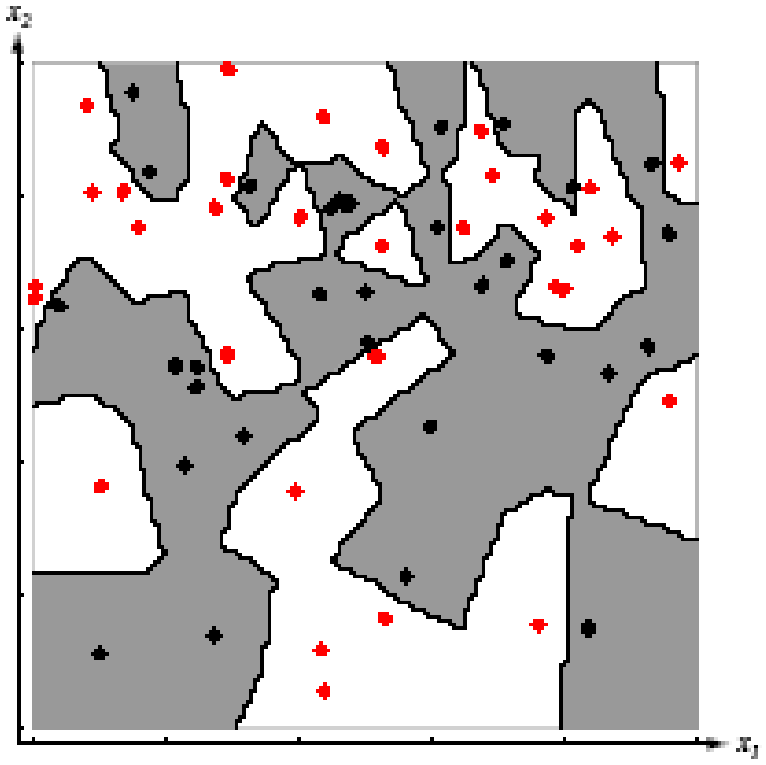
$$p_n\left(\mathbf{x}\middle|\omega_i\right)=\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_j^i}{h}\right)$$

4. 采用Bayes判别准则进行分类。

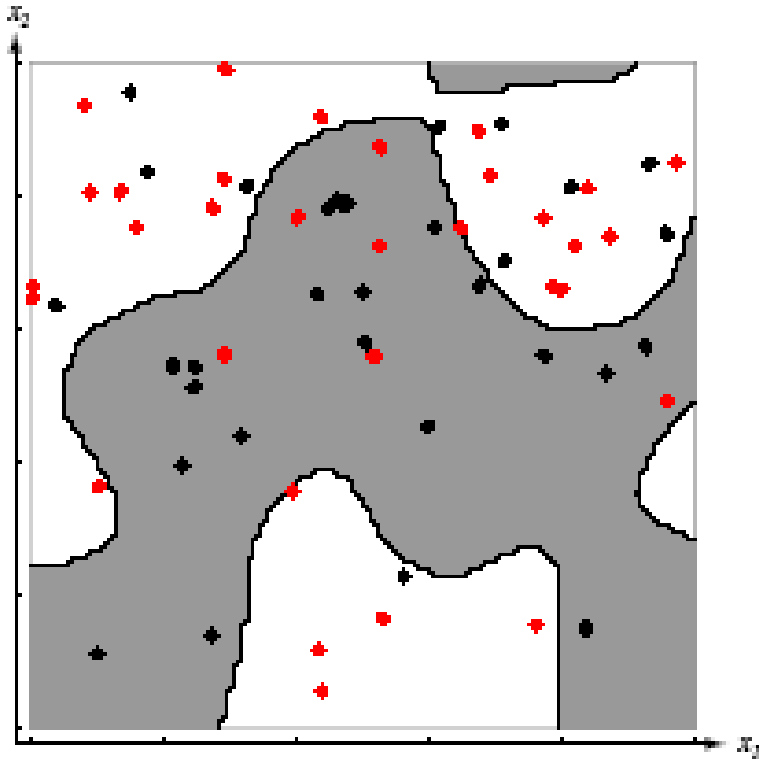
➢ 例子：

基于Parzen估计的Bayesian分类器

较小 h_n



较大 h_n



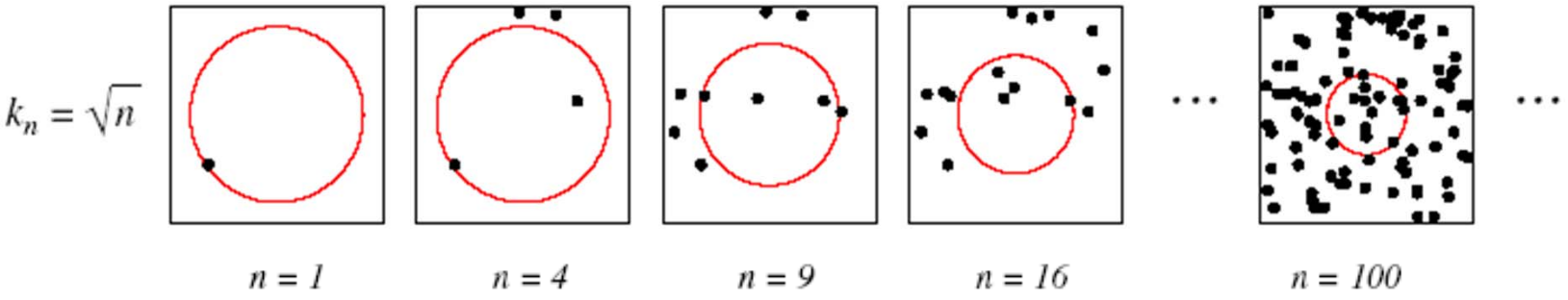
Outline:

- 引言
- 最大似然估计
  - 最大似然估计基本原理及求解
  - 正态分布的最大似然估计
- Bayes估计与Bayes学习
  - Bayes估计
  - Bayes学习
  - 正态分布下的Bayes估计
- 概率密度估计的非参数方法
  - 非参数估计基本原理及直方图方法
  - Parzen窗法
  - k_N近邻估计方法

K_n近邻估计

➢ K_n近邻密度估计：

固定样本数为 k_n ，在 **x** 附近选取与之最近的 k_n 个样本，计算该 k_n 个样本分布的最小体积 V_n



在**X**处的概率密度估计值为：

$$\hat{p}_n(\mathbf{x}) = \frac{k_n / n}{V_n}$$

渐近收敛的条件

$\hat{p}_n(\mathbf{x})$ 渐近收敛的充要条件为：

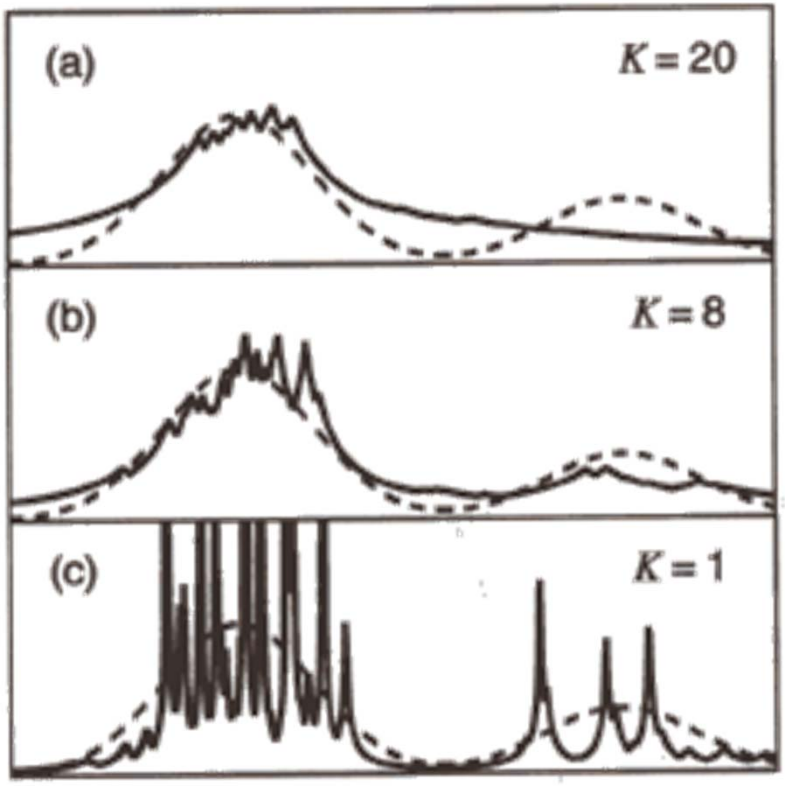
$$\lim_{n\to\infty} k_n = \infty$$

$$\lim_{n\to\infty} k_n / n = 0$$

通常选择： $k_n = \sqrt{n}$

K_n近邻估计

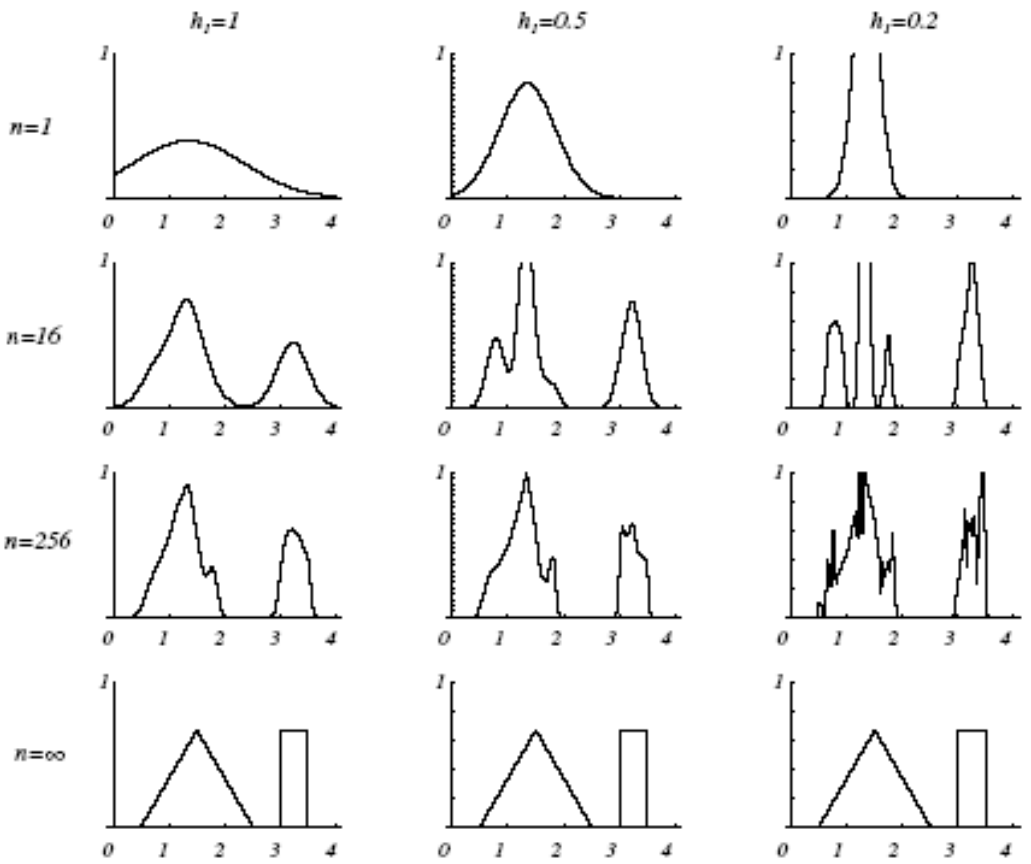
- 例子：



➢例子：

Parzen windows

k_n-nearest-neighbor



$$k_n = \sqrt{n}$$

k_N-近邻法


```
function [ f ] = Kn( N )
%KN Kn 估计法

x=linspace(-2,2,500);          % 产生-2,2 之间的 N 点行矢量
x2=randn(1,N);                 % 产生 N 个标准正态随机数
p=zeros(1,500);
for i=1:500
      Kn=sqrt(N);
      x3=sort(abs(x(i)-x2));   % 排序
      Vn=x3(Kn);               % 让体积扩张,直到包含 KN 个样本
      p(i)=(Kn/N)/Vn;
end
plot(x,p);
end
```

Discussion

- **Nearest Neighbor Estimation**

  构造 k_N 序列

- **Parzen Windows**

  构造 V_N 序列

- 非参数估计，对于样本量、计算量和存储量的要求

Discussion

- 密度函数估计
  - ✓有限样本下，密度函数的估计问题是一个很难的问题（不适定），比分类器设计问题甚至更难，也是一个更一般的问题。因此，通过首先估计密度函数来解决**PR** 问题似乎不是个好主意（除非有充分的先验知识）。
- 非参数分类判别方法