第四章　概率密度函数的估计

引言

贝叶斯决策：　已知$P(\omega_i)$和$p(\boldsymbol{x}\mid\omega_i)$，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$p(\boldsymbol{x}\mid\omega_i)$和$P(\omega_i)$，记$\hat{p}(\boldsymbol{x}\mid\omega_i)$和$\hat{P}(\omega_i)$
- 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

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
- 有充分的训练样本

本章讨论内容： 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)

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：样本的某种函数，用来作为对某参数的估计
- 参数空间：待估计参数的取值空间$\theta \in \Theta$
- 点估计：统计量$\hat{\theta}(x)$的估计值（根据样本得到的具体值）
- 区间估计

4.1 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D_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，只其参数$\theta$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$p(\boldsymbol{x}\mid\theta_i)$。

鉴于上述假设，我们可以只考虑一类样本，记已知样本为

$$D=\{x_1,x_2,\cdots,x_N\}$$

似然函数（likelihood function）

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

—— 在参数θ下观测到样本集 *D* 的概率（联合分布）密度

基本思想：

如果在参数$\theta=\hat{\theta}$下$L(\theta)$最大，则$\hat{\theta}$应是“最可能”的参数值，它是样本集的函数，记作$\hat{\theta}=\arg\max_{\theta}p(D\,|\,\theta)$。称作最大似然估计量。

为了便于分析，还可以定义对数似然函数$l(\theta)=\ln L(\theta)$。

求解：

若似然函数满足连续、可微的条件，则最大似然估计量就是方程

$$dL(\theta)/d\theta = 0 \text{ 或 } \quad dl(\theta)/d\theta = 0$$

的解（必要条件）。

若未知参数不止一个，即$\theta = [\theta_1, \theta_2, \cdots, \theta_s]^T$，记梯度算子

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_s}\right]^T$$

则最大似然估计量的必要条件由 S 个方程组成：

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

例子

- 正态分布下的最大似然估计

  以单变量正态分布为例 $\theta = [\theta_1, \theta_2]^T$, $\theta_1 = \mu$, $\theta_2 = \sigma^2$

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

- 均匀分布的最大似然估计

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

讨论：

- 如果$l(\theta)$或$L(\theta)$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，用其它方法求最大（例如均匀分布）
- 对分布的前提假设要对

思考：如果由均匀分布产生的数据用高斯分布做估计会如何？

4.2 贝叶斯估计和贝叶斯学习

贝叶斯估计

思路与贝叶斯决策类似，只是离散的决策状态变成了连续的估计。

基本思想：

把待估计参数 θ 看作具有先验分布 $p(\theta)$ 的**随机变量**，其取值与样本集 D 有关，根据样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 估计 θ。

$$p(\theta\,|\,D)=\frac{p(D\,|\,\theta)p(\theta)}{\int_{\Theta}p(D\,|\,\theta)p(\theta)d\theta}=\frac{p(D\,|\,\theta)p(\theta)}{p(D)}$$

思考：与贝叶斯决策的比较?

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

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

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 D 下，最小化经验风险：$R(\hat{\theta}\mid D)=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\mid D)d\theta$

贝叶斯决策与贝叶斯估计的比较

贝叶斯决策	贝叶斯估计
样本 $\boldsymbol{x}$	样本集合 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$
类的先验概率 $p(\omega_i)$	参数先验分布 $p(\theta_i)$
真实状态 ω_i	真实参数 θ_i
决策 α_i	估计 $\hat{\theta}_i$
类别状态：离散	分布参数：连续
损失函数表（决策表）	损失函数

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

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

可以证明（课后练习），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 D 下，θ 的贝叶斯估计是：

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

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 $p(D|\theta)=\prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta)$

（3）求θ的后验概率分布 $p(\theta|D)=\dfrac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)d\theta}$

（4）求θ的贝叶斯估计量 $\hat{\theta}=\int_{\Theta}\theta p(\theta|D)d\theta$

贝叶斯学习

- 我们的目标是希望通过对已有观测数据的学习，估计样本的真实分布 $p(\boldsymbol{x})$：

$$p(\boldsymbol{x}\mid D)=\int_{\Theta}p(\boldsymbol{x}\mid\theta,D)p(\theta\mid D)d\theta$$
$$=\int_{\Theta}p(\boldsymbol{x}\mid\theta)p(\theta\mid D)d\theta$$

其中，
$$p(\theta\mid D)=\frac{p(D\mid\theta)p(\theta)}{\int_{\Theta}p(D\mid\theta)p(\theta)d\theta}$$

$p(\boldsymbol{x}\mid\theta)$容易算，关键是求得 $p(\theta\mid D)$

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

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

因此有递推后验概率公式：

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

设 $p(\theta\mid D^0)=p(\theta)$，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。

正态分布下的贝叶斯估计

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

结论：$\hat{\mu}=\dfrac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\dfrac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$ ，　$\sigma_N^2=\dfrac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$，

其中　$m_N=\dfrac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$　　-----　样本信息与先验知识的线性组合

讨论：

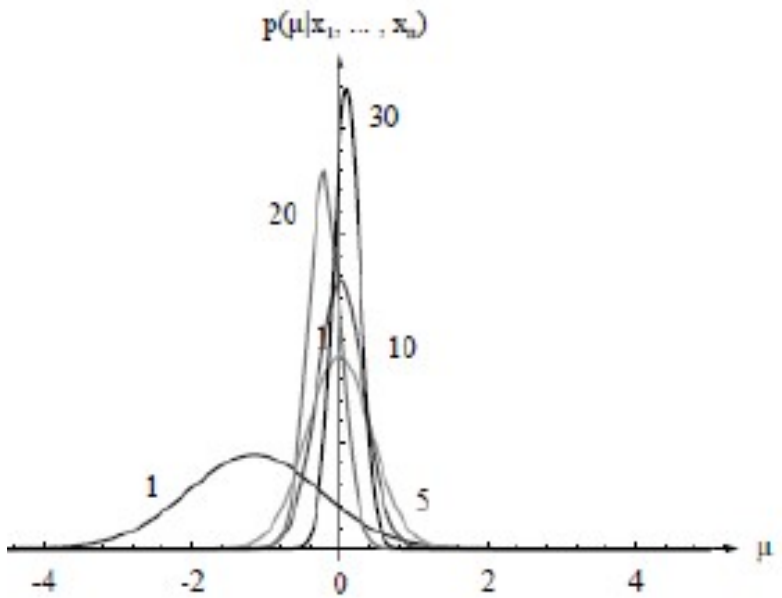
$N=0$ 时，$\hat{\mu}=\mu_0$；$N\to\infty$ 时，$\hat{\mu}\to m_N$

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）

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

对一维正态分布均值的贝叶斯学习过程



Richard Duda, Pattern Classification, second edition, figure 3.2

关于先验分布（Prior）

- 参数先验分布的选取取决于先验知识和对问题的理解

- 极端情况
  1. 很强的先验：脉冲函数
  2. 没有任何先验，Non-informative prior

最大似然估计 v.s. 贝叶斯估计

- 最大似然估计简单直观
- 当训练样本数无穷多的时候，最大似然估计和贝叶斯估计的结果是一样的
- 贝叶斯估计由于使用了先验概率，利用了更多的信息
- 如果这些信息是可靠的，那么有理由认为贝叶斯估计比最大似然估计的结果更准确

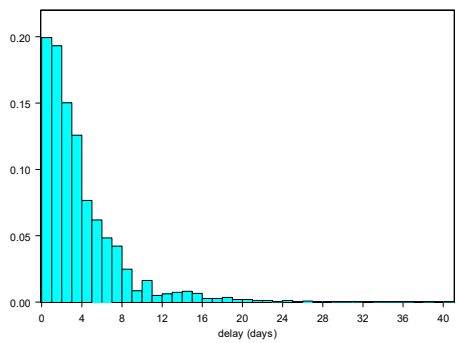
有时候先验概率很难设计

在没有特别先验知识的时候，取先验概率是这个区域中的均匀分布（无信息先验）

这种情况下最大似然估计结果和贝叶斯估计结果相似

非参数估计

直方图方法（Histogram）



- 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 k^d 个小舱）

（2）统计落入各个小舱内的样本数 q_i

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

非参数估计的基本原理

问题：已知样本集 $D=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$，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 $P_R=\int_{\Re}p(x)dx$

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

k 的期望值 $E[k]=NP_R$

k 的众数（出现频率最高的取值）为 $m=[(N+1)P_R]\approx NP_R$（[*]表示取整）

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

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。

关于V的选择：过大，估计粗糙；过小，可能某些区域中无样本

理论结果：

设有一系列包含x的区域$\Re_1,\Re_2,\cdots,\Re_n,\cdots$，对$\Re_1$采用$k_1$个样本进行估计，对$R_2$用$k_2$个，…，$(k_1<k_2<\cdots)$。设$\Re_n$包含$k_n$个样本，$V_n$为$\Re_n$的体积，$\hat{p}_n(x)=\dfrac{k_n}{nV_n}$为$p(x)$的第$n$次估计，有下面的结论：

如果：　（1）$\lim\limits_{n\to\infty}V_n=0$；　（2）$\lim\limits_{n\to\infty}k_n=\infty$；　（3）$\lim\limits_{n\to\infty}\dfrac{k_n}{n}=0$

则$\hat{p}_n(x)$收敛于$p(x)$。

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

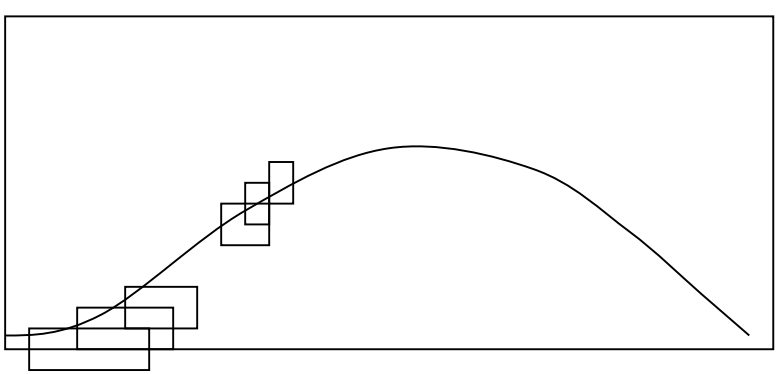
2. 选择k_n，（比如$k_n = \sqrt{n}$），V_n为正好包含x的k_n个近邻

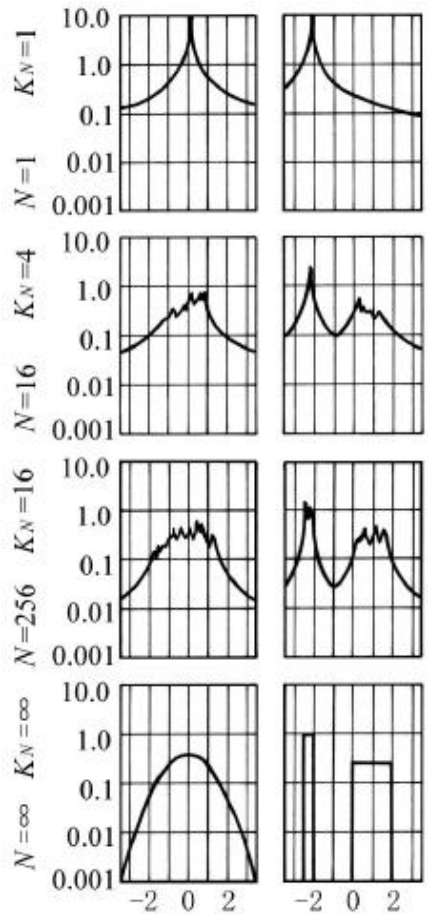
—— k_N近邻估计

k_N -近邻估计

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

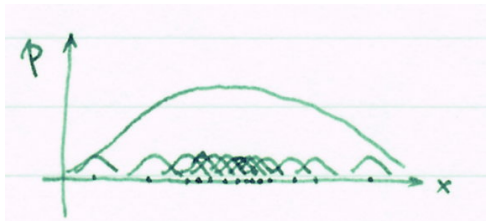
通过控制小区域内的样本数 k_n 来确定小区域大小。

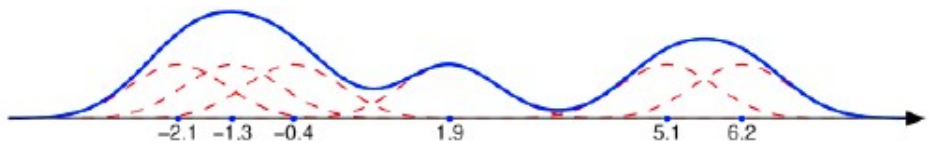




Parzen 窗法

$$\hat{p}(x)=\frac{1}{N}\sum_{i=1}^{N}k(\boldsymbol{x},\boldsymbol{x}_i)$$





窗函数（核函数）$k(x,x_i)$，反映 x_i 对 $p(x)$ 的贡献，实现小区域选择。

条件：$k(x,x_i)\geq 0$

$$\int k(x,x_i)dx=1$$

常用窗函数：

（1）超立方体窗（方窗）

$$k(x,x_i)=\begin{cases}\dfrac{1}{h^d} & \text{if} \quad \left|x^i-x_i^j\right|\le h/2,\ j=1,2,\cdots,d\\ 0 & \text{otherwise}\end{cases}$$

h 为超立方体棱长，$V=h^d$

（2）正态窗（高斯窗）

$$k(x,x_i)=\frac{1}{\sqrt{(2\pi)^d\,\rho^{2d}\,|Q|}}\exp\left\{-\frac{1}{2}\frac{(x-x_i)^T Q^{-1}(x-x_i)}{\rho^2}\right\},\quad (\Sigma=\rho^2 Q)$$

一维标准正态：$k(x,x_i)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-x_i)^2\right\}$

（3）超球窗

$$k(x,x_i)=\begin{cases}V^{-1} & \text{if} \quad \|x-x_i\|\le\rho \\ 0 & \text{otherwise}\end{cases}$$

（V：超球体积，半径ρ）

窗宽的选择：

- 样本数少则选大些，样本数多则选小些，

Parzen 窗估计的性质：

在满足一定的条件下，估计量 $\hat{p}_N(x)$ 是渐近无偏和平方误差一致的。条件是：

1. 总体密度 $p(x)$ 在 x 点连续；
2. 窗函数满足以下条件：

$$\varphi(u) \geq 0, \quad \int \varphi(u)du = 1$$ ：窗函数具有密度函数的性质

$$\sup_u \varphi(u) < \infty$$ ：窗函数有界

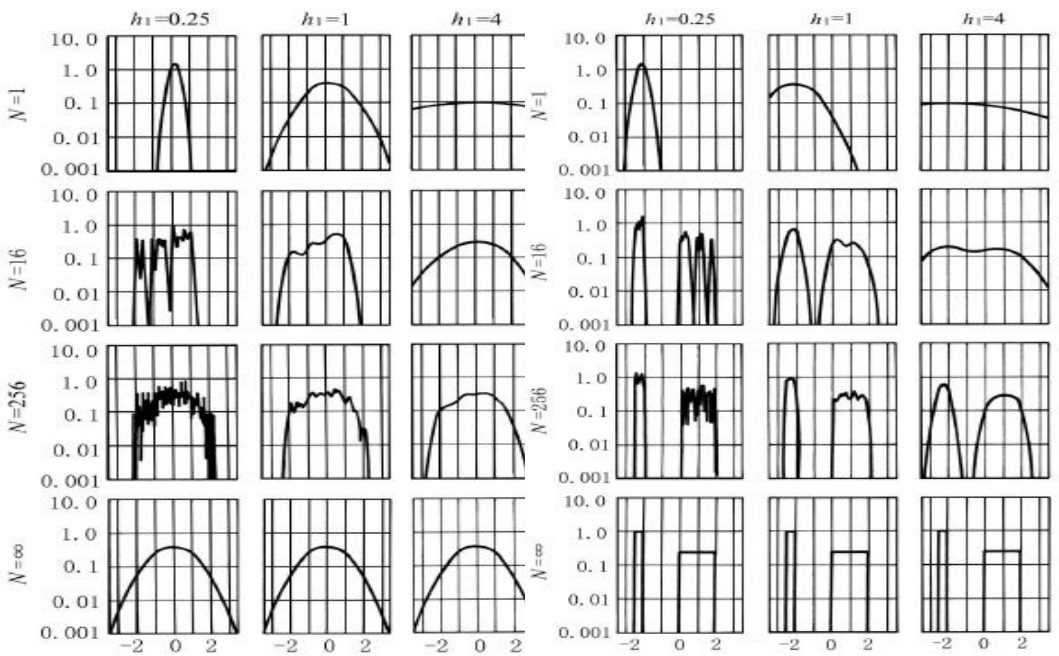
$$\lim_{\|u\| \to \infty} \varphi(u) \prod_{i=1}^{d} u_i = 0$$ ：窗函数随着距离的增大很快趋于零

3. 窗宽受以下条件约束：

$$\lim_{N\to\infty} V_N = 0$$ ：窗体积随着 N 的增大而趋于零

$$\lim_{N\to\infty} NV_N = \infty$$ ：但体积减小的速度要低于 1/N

举例：用已知的密度函数产生一系列样本，根据这些样本用 Parzen 窗法估计概率密度函数，与真实密度函数比较，分析样本数，窗宽等对估计结果的影响。右图是两种高斯窗进行估计的结果。



讨论

- 贝叶斯分决策理论上是最优的，前提条件是需要知道样本分布。
- 但在有限样本下，密度函数的估计问题是一个很难的问题，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模式识别问题不一定是个好主意（除非有充分的先验知识）。

小结：概率密度函数估计

- 参数估计：概率密度函数形式已知，只未知几个参数θ
  - **最大似然估计**

似然函数 $L(\theta)=p(D\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)$

对数似然函数 $l(\theta)=\ln L(\theta)$

最大似然估计量 $L(\hat{\theta})=\max_{\theta}L(\theta)$ 或记 $\hat{\theta}=\arg\max_{\theta}l(\theta)$

求解：连续可微条件下 $\nabla_{\theta}l(\theta)\big|_{\theta=\hat{\theta}}=0$

正态分布例： $\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i$ $\hat{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$

■ 贝叶斯估计

把$\hat{\theta}$看作随机变量，先验分布$p(\theta)$

最小化风险　$R = \int R(\hat{\theta} \mid x)\, p(x)dx$

对样本集　$R(\hat{\theta} \mid x) = \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta \mid D)d\theta$

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

贝叶斯估计　$\hat{\theta} = E[\theta \mid D] = \int_{\Theta} \theta p(\theta \mid D)d\theta$

求法：　$p(D \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$

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

贝叶斯学习

$$p(x \mid D) = \int_{\Theta} p(x \mid ,\theta) p(\theta \mid D) d\theta$$

递推

$$p(\theta \mid D^N) = \frac{p(x_N \mid \theta) p(\theta \mid D^{N-1})}{\int p(x_N \mid \theta) p(\theta \mid (D^{N-1}) d\theta}$$

- 非参数估计：直接估计密度函数（数值解），不对函数形式作假设

基本思想：将取值空间分为多个小区间，假定小区间内密度值不变，用小区间内的样本估计此值。 $\hat{p}(x) = \frac{k}{NV}$

- k_N 近邻估计 $\hat{p}_n(x) = k_N / NV_n$
- Parzen 窗法 $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} k(x, x_i)$