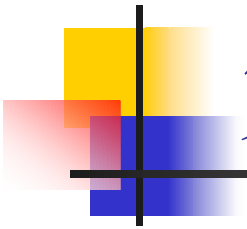
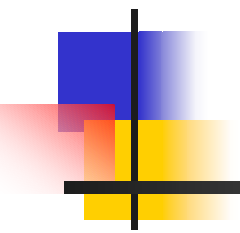


第三章 概率密度估计

- 3.0 引言
- 3.1 参数估计
- 3.2 非参数估计
- 3.3 说明



3.0 引言

3.0 引言

- 进行Bayes决策需要事先知道两种知识：
  - 各类的先验概率；
  - 观测向量的类条件概率密度。

- 知识的获取（估计）：
  - 一些训练数据；
  - 对问题的一般性的认识。

3.0 引言

- 类的先验概率的估计（较容易）：
  - 依靠经验；
  - 用训练数据中各类出现的频率估计。
  - 用频率估计概率的优点：
    - 无偏性；
    - 相合性；
    - 收敛速度快。

3.0 引言

- 类条件概率密度的估计（非常难）：
  - 概率密度函数包含了一个随机变量的全部信息；
  - 概率密度函数可以是满足下面条件的任何函数：

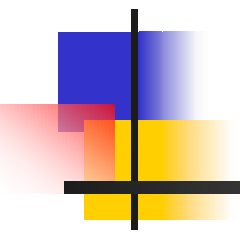
$$p(x) \geq 0, \qquad \int p(x)dx = 1.$$

3.0 引言

- 概率密度估计的两种主要思路：
  - 参数估计：

    根据对问题的一般性的认识，假设随机变量服从某种分布，分布函数的参数通过训练数据来估计。
  - 非参数估计：

    不用模型，而只利用训练数据本身对概率密度做估计。



3.1 参数估计

3.1 参数估计

- 估计随机变量 X的概率密度：
  - 给定某类训练数据 — 样本

$$x_1, x_2, \cdots, x_n, \qquad x_i \in \Re^d,$$

  - 假设已知 X 所服从的分布形式，待估计的参数为θ。例如，假定 X 服从正态分布$N(\mu, \Sigma)$，待估参数是 $\theta = (\mu, \Sigma)$.

3.1 参数估计

- 最大似然估计(Maximum Likelihood):
  - 把待估计的参数θ看作“数”，与后面将要讲到的Bayes估计中把 θ看作随机变量有区别。
  - 为了描述概率密度函数$p(x)$与参数 θ 的依赖关系，用$p(x;\theta)$表示。

3.1 参数估计

- 似然函数：$\prod_{i=1}^{n} p(x_i;\theta)$

等价

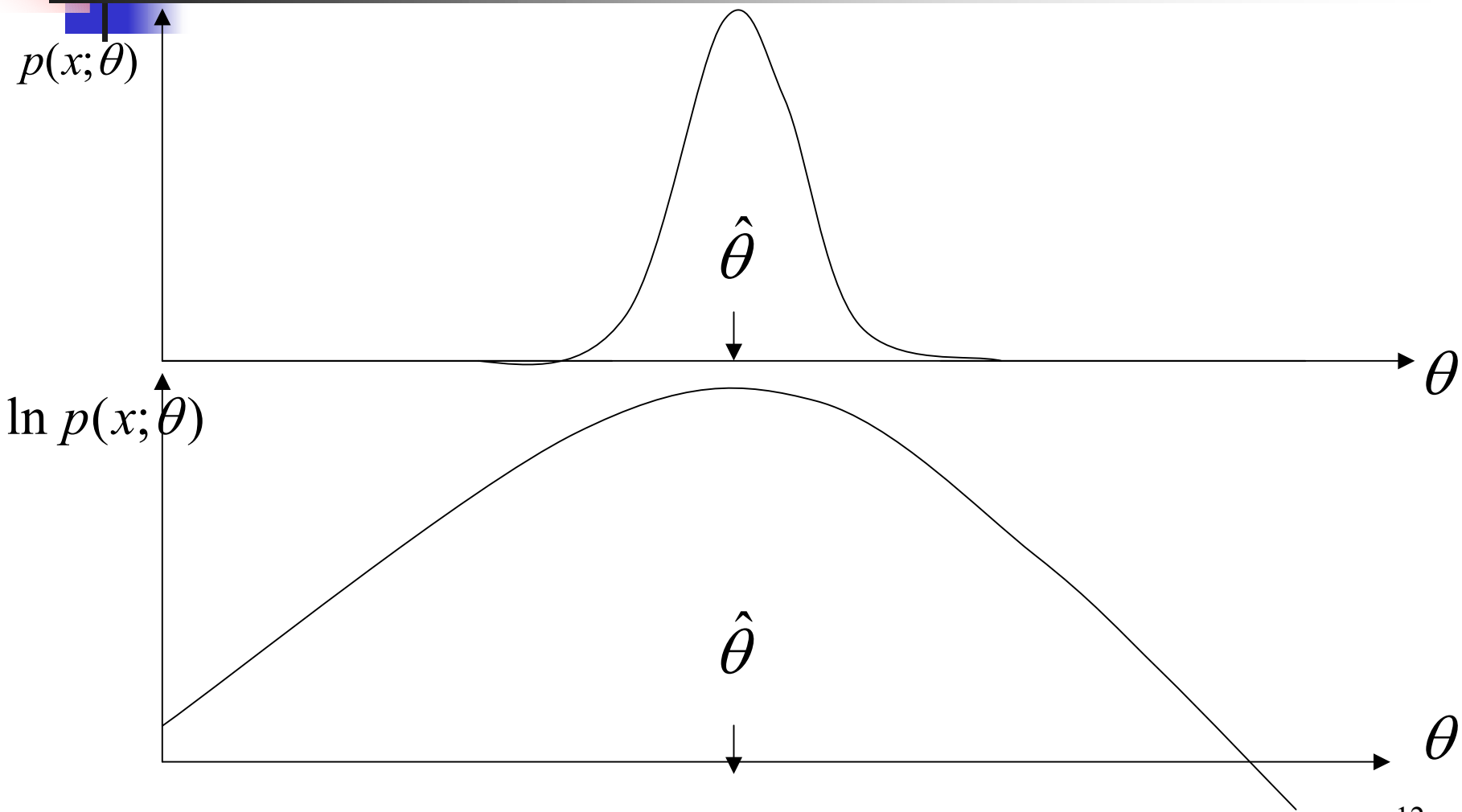
- 对数似然函数：$\sum_{i=1}^{n} \ln p(x_i;\theta)$
- 最大似然估计量$\hat{\theta}_{ML}$：

$$\hat{\theta}_{ML} = \arg\max_{\theta} \sum_{i=1}^{n} \ln p(x_i;\theta)$$

3.1 参数估计

- 似然函数给出了从总体样本中抽出n个样本 $x_1, x_2, \cdots, x_n$ 的概率。
- 假设样本是独立抽取的，并且不同类别的参数是相互独立的。
- 最大似然估计就是根据已经抽取的n个样本 $x_1, x_2, \cdots, x_n$ ，来估计这组样本“最可能”来自哪个密度函数。

最大似然估计示意图



3.1 参数估计

- 参数求解：必要条件是梯度为0。

$$\nabla_\theta(\sum_{i=1}^{n}\ln p(x_i;\theta))=0$$

3.1 参数估计

- 正态分布假设下的最大似然参数估计：

$$p(x_i;\mu,\Sigma)$$

$$=\frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{d/2}}\exp(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))$$

3.1 参数估计

- 似然函数：

$$\sum_{i=1}^{n} \ln p(x_i; \mu, \Sigma)$$

$$= -\frac{n}{2}\ln((2\pi)^d |\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu).$$

3.1 参数估计

- 均值 μ估计：

$$\nabla_{\mu}\left(\sum_{i=1}^{n}\ln p(x_i;\mu,\Sigma)\right)=\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu).$$

$$\hat{\Sigma}_{ML}^{-1}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})=0.$$

$$\hat{\mu}_{ML}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}.$$

均值的最大似然估计与 $\sum$ 无关！

3.1 参数估计

- 协方差矩阵 Σ的估计：
  - 对矩阵求导：

    $A=(a_{ij})_{m\times n}$是矩阵，f是A的函数。

$$\frac{\partial f(A)}{\partial A}=(\frac{\partial f}{\partial a_{ij}})_{m\times n}.$$

3.1 参数估计

- 几个矩阵求导的公式，A, B是矩阵：

$$\frac{\partial tr(A^{-1}B)}{\partial A} = -(A^{-1}(B+B^T)A^{-1} - diag(A^{-1}BA^{-1})).$$

$$\frac{\partial \ln|A|}{\partial A} = 2A^{-1} - diag(A^{-1}).$$

3.1 参数估计

令 $\mathrm{T} = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T .$

$$\nabla_{\Sigma} \left(\sum_{i=1}^{n} \ln p(x_i; \mu, \Sigma) \right)$$

$$= -\frac{n}{2} (2\Sigma^{-1} - diag\,\Sigma^{-1})$$

$$+ \frac{1}{2} (2\Sigma^{-1}\mathrm{T}\Sigma^{-1} - diag(\Sigma^{-1}\mathrm{T}\Sigma^{-1})).$$

3.1 参数估计

$$\nabla_{\Sigma}\left(\sum_{i=1}^{n}\ln p(x_i;\theta)\right)=0$$

$$\hat{\Sigma}_{ML}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})(x_i-\hat{\mu}_{ML})^T.$$

协方差矩阵的最大似然估计与 $\hat{\mu}_{ML}$ 有关！

3.1 参数估计

- 协方差矩阵的最大似然估计是有偏估计：

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T\right)$$

$$=\frac{n-1}{n}\Sigma$$

$$\neq \Sigma.$$

- 无偏估计：$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T.$

3.1 参数估计

- 最大后验概率估计（MAP）—maximum *a posteriori*.
  - 把待估计的参数θ看作随机变量，希望使后验概率最大。

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid X)$$

X 表示训练数据 $x_1, x_2, \cdots, x_n$.

3.1 参数估计

- 利用Bayes公式：

$$p(\theta \mid X) = \frac{p(X \mid \theta) p(\theta)}{p(X)}.$$

分母中的$p(X)$与参数估计无关，MAP估计化为：

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(X \mid \theta) p(\theta).$$

3.1 参数估计

- $p(\theta)$是参数作为随机变量的先验概率密度函数，一般根据经验确定。
- MAP估计的求解：

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(X \mid \theta) p(\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \ln p(x_i \mid \theta) + \ln p(\theta).$$

$$\nabla_{\theta}\left(\sum_{i=1}^{n} \ln p(x_i \mid \theta) + \ln p(\theta)\right) = 0$$

3.1 参数估计

- 正态分布假设下的最大后验概率估计：假设协方差矩阵已知，估计均值向量。
- 假设均值向量 μ 服从正态分布：

$$p(\mu)=\frac{1}{(2\pi)^{d/2}\sigma_{\mu}^{d}}\exp(-\frac{\|\mu-\mu_0\|^2}{2\sigma_{\mu}^{2}})$$

- MAP估计：

$$\nabla_{\mu}(\sum_{i=1}^{n}\ln p(x_i\mid\mu)+\ln p(\mu))=0.$$

3.1 参数估计

- 假设： $\Sigma = \sigma^2 I$

$$\sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i - \mu) - \frac{1}{\sigma_\mu^2}(\mu - \mu_0) = 0$$

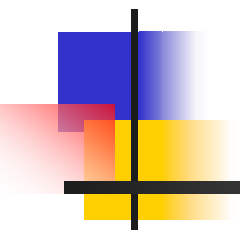
$$\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2}\sum_{i=1}^{n} x_i}{1 + n\frac{\sigma_\mu^2}{\sigma^2}}$$

3.1 参数估计

- Bayes(MAP)估计与ML估计的关系：
  - 当样本数趋于无穷时，MAP估计一般趋向于ML估计。（见习题1）
  - ML估计也可以看作参数的先验概率密度函数服从均匀分布（相当于没有先验知识）的MAP估计。
  - 当参数的先验概率密度函数比较准确时，MAP估计的小样本性质大大优于ML估计。

3.1 参数估计

- 参数估计中的模型选择问题：
  - 实际工作中处理的大都是高维数据：$d \geq 10$。
  - 统计学中经典的多元（高维）分布很少，研究的最详尽的是多元正态分布。
  - 近几十年的研究发现，实际所处理的高维数据几乎都不服从正态分布。
  - 通过增加模型的复杂程度（参数的个数），如正态模型的线性组合 — 高斯混合模型，试图“逼近”真实的分布，出现了过拟合问题。



3.2 非参数估计

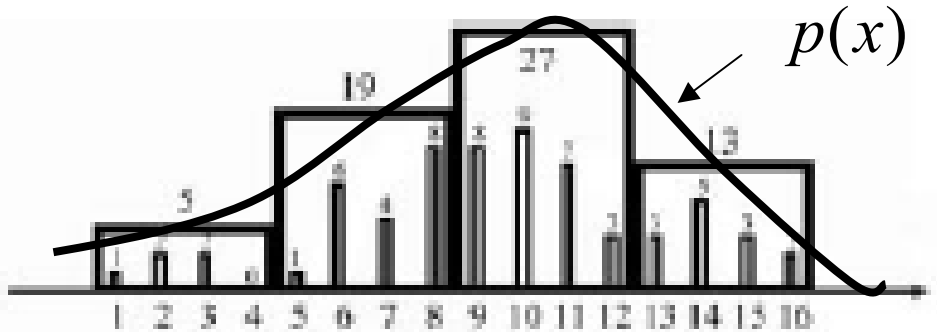
3.2 非参数估计

- 与参数估计需要事先假定一种分布函数不同，非参数估计不做任何模型假设。

- 两种主要方法：
  - 直方图法；
  - 核方法。

3.2 非参数估计

- 直方图法：
  - 用直方图逼近概率密度函数。



  - 高维空间由于数据稀疏，很难应用。

3.2 非参数估计

- 核方法：
  - 用某种核函数的线性组合估计概率密度。
  - 模式识别中常用的两种方法：

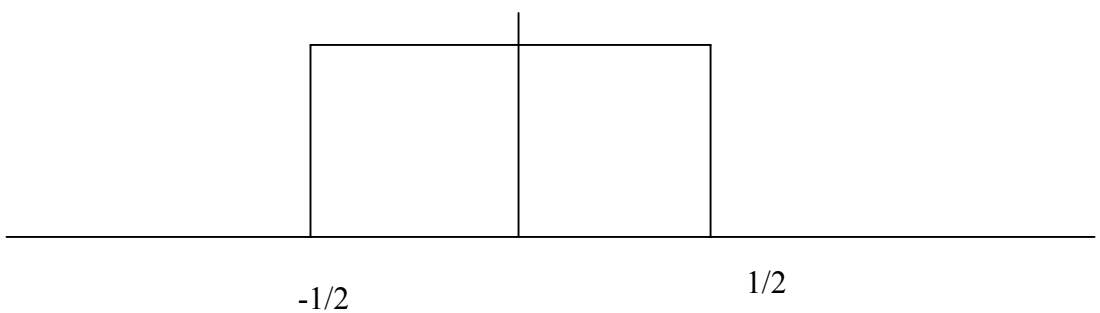
    Parzen窗法;

    k_n-近邻法。

3.2 非参数估计

- 两种常用的核函数：
  - 均匀核：$x=(x^1,x^2,\cdots,x^d)\in\Re^d$

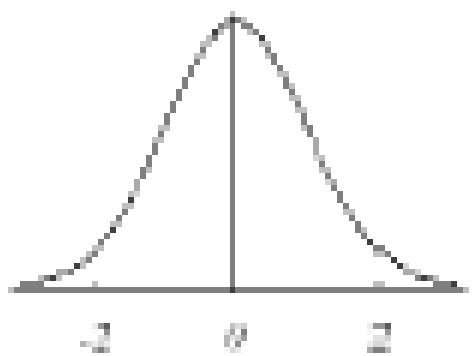
$$k(x)=\begin{cases}1 & |x^i|\le 1/2 \quad i=1,2,\cdots,d.\\ 0 & \text{其它}\end{cases}$$



3.2 非参数估计

- 正态（高斯）核：

$$k(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{\|x\|^2}{2}).$$



3.2 非参数估计

- 核函数要满足概率密度函数的条件。

$$\int_{\Re^d} k(x)dx = 1.$$

3.2 非参数估计

- Parzen窗法：

  把核函数看作“窗”，根据样本 $x_1, x_2, \cdots, x_n$，

$$\hat{p}_n(x) = \frac{1}{nh_n{}^d} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h_n}\right).$$

  h_n 是控制“窗”宽度的参数，根据样本的数量选择。

3.2 非参数估计

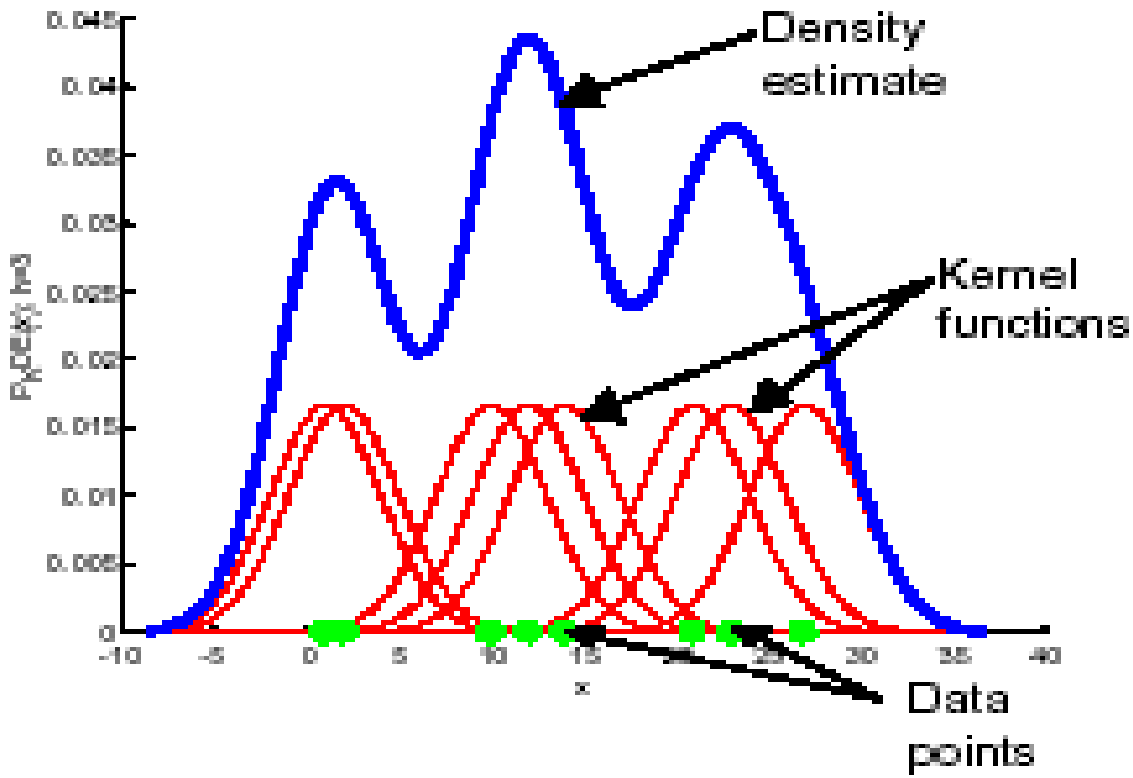
$\frac{1}{h_n^{\,d}} k(\frac{x-x_i}{h_n})$ 是以 x_i 为中心，宽 h_n 的窗。

满足归一化条件：

$$\int_{\Re^d} \frac{1}{h_n^{\,d}} k(\frac{x-x_i}{h_n}) dx = 1$$

3.2 非参数估计

- Parzen窗的例子：



3.2 非参数估计

- 窗宽h_n 的选择：
  - 保证依概率渐进收敛到真实的概率密度：

$$\hat{p}_n(x) \xrightarrow{P} p(x). \qquad \text{任何 } p(x).$$

收敛的充要条件：

$$\lim_{n\to\infty} h_n = 0.$$

$$\lim_{n\to\infty} n h_n^{\,d} = \infty.$$

3.2 非参数估计

- 样本有限—均方误差最小(MSE)准则，但

$$MSE(E(\hat{p}_n(x)-p(x))^2)\sim O(n^{-\frac{4}{d+4}}).$$

- 维数灾难(Curse of Dimensionality):

  当维数较高时，样本数量无法达到精确估计的要求。

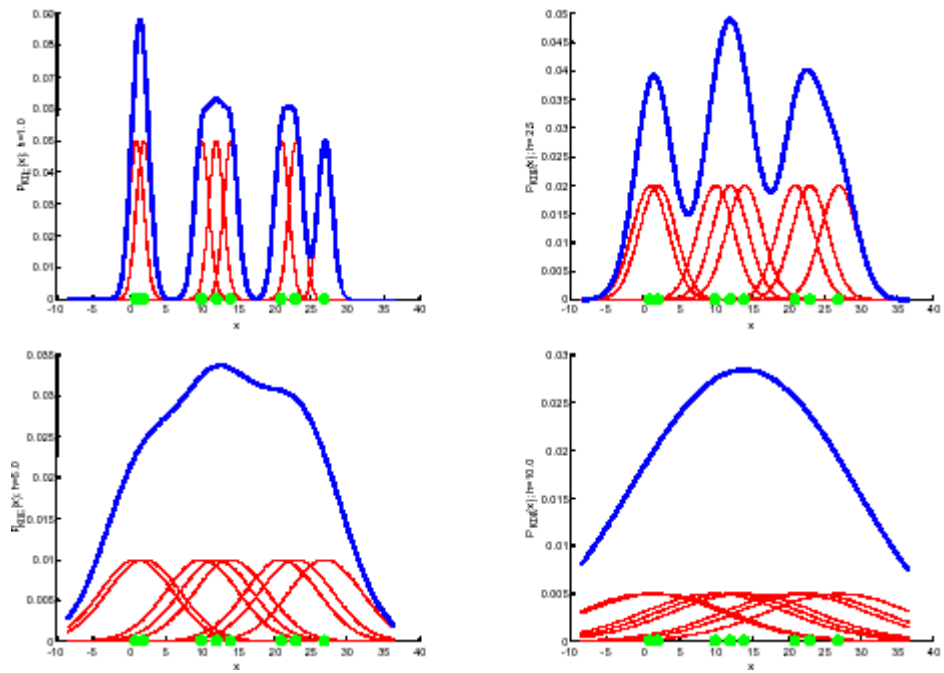
3.2 非参数估计

- 不同维数达到相同估计精度所需的样本数：

n	d	$n^{-\frac{4}{d+4}}$
16	1	0.1
32	2	0.1
178	5	0.1
3162	10	0.1
3×10^{13}	50	0.1

3.2 非参数估计

- 不同窗宽的估计效果：



3.2 非参数估计

- 均匀核函数Parzen窗估计的几何意义：

点 x 处概率密度=

$$\frac{\text{以}x\text{为中心,}2h_n\text{为边长的超立方体内的样本数}}{\text{总样本数}}$$

3.2 非参数估计

- k_n—近邻估计：
  - 均匀核函数Parzen估计，窗宽固定，不同位置落在窗内的样本点的数目是变化的。
  - k_n— 近邻估计：把窗扩大到刚好覆盖k_n个点。落在窗内的样本点的数目固定，窗宽是变化的。
  - k_n根据样本总数 n选择。

3.2 非参数估计

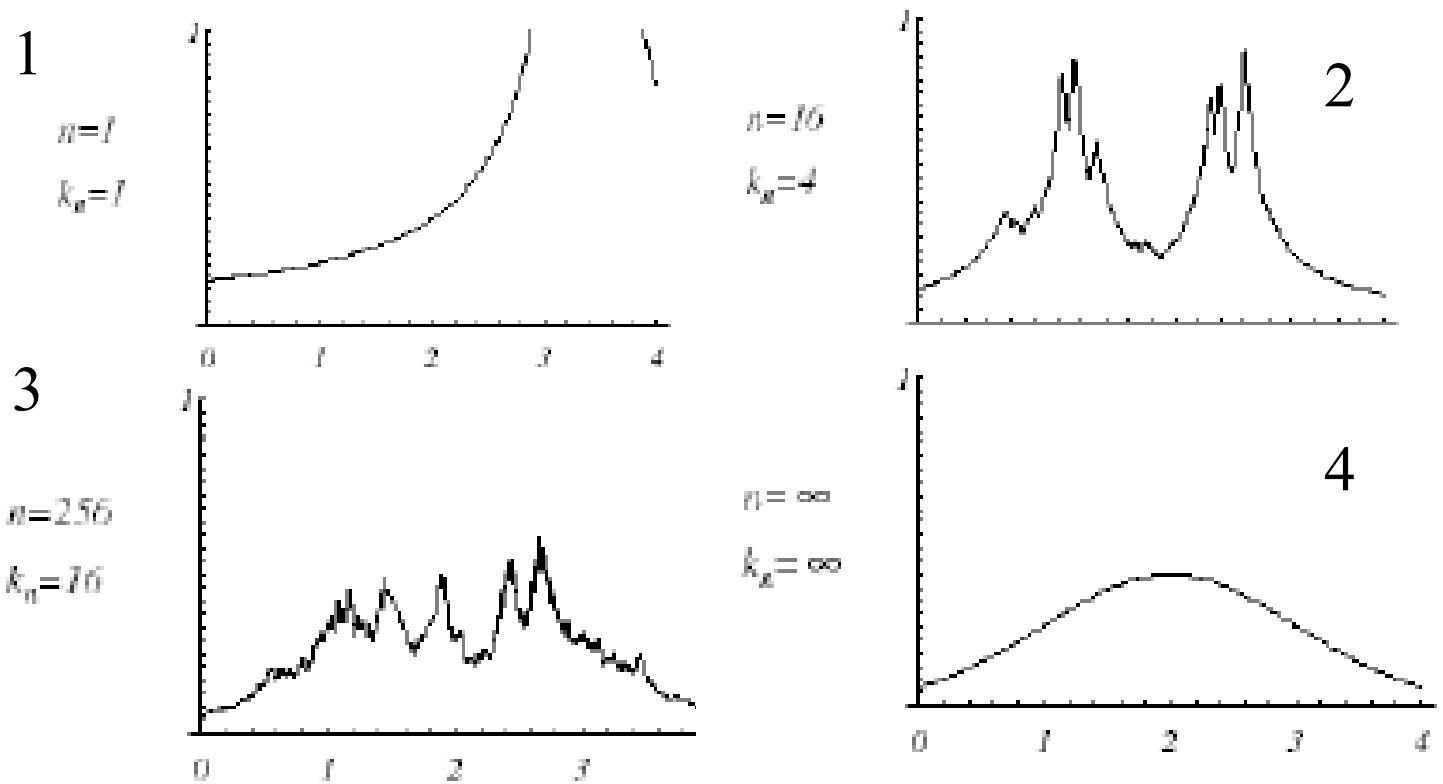
- 概率密度估计表达式：

点 x 处窗的“体积”是 V_n：

$$\hat{p}_n(x) = \frac{1}{V_n}\frac{k_n}{n}.$$

3.2 非参数估计

- 一个 k_n — 近邻估计例子

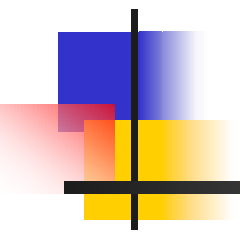


3.2 非参数估计

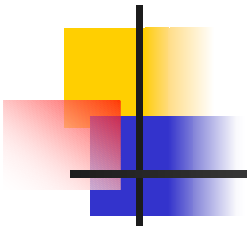
- k_n的选择：

  渐进收敛容易保证；

  有限样本性质、最小平方误差与Parzen窗几乎相同。

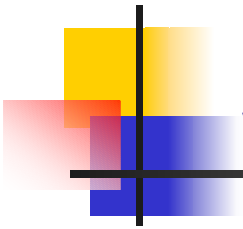


3.3 说明



3.3 说明

- 高维概率分布的估计无论在理论上还是实际操作中都是一个十分困难的问题。
- 概率密度函数包含了随机变量的全部信息，是导致估计困难的重要原因。
- 进行模式识别并不需要利用概率密度的所有信息，只需要求出分类面。
- 先估计概率密度，再进行分类，可能走了“弯路”。



参考文献

[1] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, 2nd Edition, John Wiley & Sons, Inc. 2001.

关于模式识别中矩阵微分的常用公式可参看[2]的附录：

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition. 2nd Edition, Academic Press, 1990.

非参数估计的理论性较强的的教材是[3]：

[3] L. Devroye, Nonparametric density estimation : the L1 view, New York : Wiley, c1985.