第三章
概率密度函数的估计

2009-10-20

贝叶斯估计 — 最小风险

- **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**：

$$\begin{aligned} R &= \int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\mathbf{x},\theta)d\theta d\mathbf{x} \\ &= \int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})p(\mathbf{x})d\theta d\mathbf{x} \\ &= \int_{E^d}p(\mathbf{x})\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta d\mathbf{x} \\ &= \int_{E^d}R(\hat{\theta}\,|\,\mathbf{x})p(\mathbf{x})d\mathbf{x}; \end{aligned}$$

- **条件风险**：

$$R(\hat{\theta}\,|\,\mathbf{x}) = \int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta.$$

贝叶斯估计 — 最小风险

- 最小化期望风险 ➔ 最小化条件风险。
- 在有限样本集下，最小化经验风险

$$R(\hat{\theta}\,|\,\boldsymbol{K}) = \int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\boldsymbol{K})d\theta.$$

- **贝叶斯估计量**：（在样本集 $\boldsymbol{K}$ 下）是条件风险（经验风险）最小的估计量 $\hat{\theta}_{\mathrm{BE}}$，即

$$\hat{\theta}_{\mathrm{BE}} = \arg\min_{\hat{\theta}} R(\hat{\theta}\,|\,\boldsymbol{K}).$$

贝叶斯估计 — 最小风险

- 把损失函数定义为平方误差：$\lambda(\hat{\theta},\theta) = (\theta - \hat{\theta})^2$.

$$\begin{aligned} R(\hat{\theta}\,|\,\mathbf{x}) &= \int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta \\ &= \int_{\Theta}[\theta - E(\theta\,|\,\mathbf{x})]^2 p(\theta\,|\,\mathbf{x})d\theta + \int_{\Theta}[E(\theta\,|\,\mathbf{x}) - \hat{\theta}]^2 p(\theta\,|\,\mathbf{x})d\theta; \end{aligned}$$

- **定理**：如果采用平方损失函数，则有

$$\hat{\theta}_{\mathrm{BE}} = E[\theta\,|\,\mathbf{x}] = \int_{\Theta}\theta\, p(\theta\,|\,\mathbf{x})d\theta;$$

- 同理，在给定样本集$\boldsymbol{K}$下，θ的贝叶斯估计是

$$\hat{\theta}_{\mathrm{BE}} = E[\theta\,|\,\boldsymbol{K}] = \int_{\Theta}\theta\, p(\theta\,|\,\boldsymbol{K})d\theta;$$

贝叶斯估计 — 最小风险

- 平方误差损失下，求解贝叶斯估计的步骤：
  - 确定 θ 的先验分布 $p(\theta)$;
  - 由样本集 $\boldsymbol{K}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 求出样本集的联合分布：

$$p(\boldsymbol{K}\,|\,\theta) = \prod_{k=1}^{N} p(\mathbf{x}_k\,|\,\theta);$$

  - 计算 θ 的后验分布

$$p(\theta\,|\,\boldsymbol{K}) = \frac{p(\boldsymbol{K}\,|\,\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{K}\,|\,\theta)p(\theta)d\theta};$$

  - 计算贝叶斯估计

$$\hat{\theta}_{\mathrm{BE}} = \int_{\Theta}\theta\, p(\theta\,|\,\boldsymbol{K})d\theta.$$

一元正态分布的贝叶斯估计

- 总体分布密度为：

$$p(x\,|\,\mu) \sim N(\mu,\sigma^2);$$

- 均值μ未知，μ的先验分布为：

$$p(\mu) \sim N(\mu_0,\sigma_0^2);$$

- 样本集：$\boldsymbol{K}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$
- 用贝叶斯估计方法求μ的估计量

一元正态分布的贝叶斯估计

- 计算 μ 的后验分布

$$p(\mu \mid \boldsymbol{K}) = \frac{p(\boldsymbol{K} \mid \mu)\, p(\mu)}{p(\boldsymbol{K})}$$

$$= \alpha \prod_{k=1}^{N} p(x_k \mid \mu)\, p(\mu) \sim N(\mu_N, \sigma_N^2)$$

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

其中 $m_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$ 为样本均值;

一元正态分布的贝叶斯估计

- 计算 μ 的贝叶斯估计

$$\hat{\mu} = \int \mu\, p(\mu \mid \boldsymbol{K})\, d\mu = \mu_N;$$

  - 是样本信息和先验知识的线性组合。
    - 当 $N=0$ 时，$\hat{\mu}_{BE} = \mu_0$;
    - 当 $N \to \infty$ 时，$\hat{\mu}_{BE} \to m_N$;
  - 特例:
    - 如 $\sigma_0^2 = 0$，则 $\hat{\mu}_{BE} \equiv \mu_0$; 即先验知识可靠，样本不起作用。
    - 如 $\sigma_n >> \sigma$，则 $\hat{\mu}_{BE} = m_N$; 即先验知识十分不确定，完全依靠样本信息。

贝叶斯学习

- 由局部推导总体：利用 θ 的先验分布 $p(\theta)$ 及训练样本提供的信息（似然函数）$p(\boldsymbol{K}|\theta)$，求出 θ 的后验分布 $p(\theta|\boldsymbol{K})$，然后直接求总体分布。

$$p(\mathbf{x} \mid \boldsymbol{K}) = \int p(\mathbf{x}, \theta \mid \boldsymbol{K})\, d\theta$$

$$= \int p(\mathbf{x} \mid \theta, \boldsymbol{K})\, p(\theta \mid \boldsymbol{K})\, d\theta$$

$$= \int p(\mathbf{x} \mid \theta)\, p(\theta \mid \boldsymbol{K})\, d\theta;$$

  - 把类条件概率密度（总体分布）$p(\mathbf{x}|\boldsymbol{K})$ 和未知参数的后验概率密度 $p(\boldsymbol{\theta}|\boldsymbol{K})$ 联系起来。
  - 贝叶斯解的结果与最大似然估计的结果近似相等

$$p(\mathbf{x} \mid \boldsymbol{K}) \approx p(\mathbf{x} \mid \hat{\theta}_{\mathrm{ML}}).$$

贝叶斯学习

- 考虑学习样本个数 N，记样本集 $\boldsymbol{K}^N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$;
- 当 N>1时，

$$p(\boldsymbol{K}^N \mid \theta) = p(\mathbf{x}_N \mid \theta)\, p(\boldsymbol{K}^{N-1} \mid \theta);$$

- 因此，有递推后验概率公式:

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

贝叶斯学习

- 参数估计的递推贝叶斯方法(Recursive Bayes Incremental Learning)
  - 设 $p(\theta|K^0) = p(\theta)$，当样本数目增多，可得到后验概率密度函数序列：$p(\theta), p(\theta|\mathbf{x}_1), p(\theta|\mathbf{x}_1, \mathbf{x}_2), \cdots$
  - 贝叶斯学习(Bayesian Learning)性质：如果此序列收敛于以真实数值为中心的 δ 函数，即

$$p(\theta \mid \boldsymbol{K}^{N \to \infty}) = \delta(\theta - \theta_0);$$

$$p(\mathbf{x} \mid \boldsymbol{K}^{N \to \infty}) = p(\mathbf{x} \mid \hat{\theta} = \theta_0) = p(\mathbf{x}).$$

一元正态分布的贝叶斯学习

- 计算 μ 的后验分布

$$p(\mu \mid \boldsymbol{K}) = \frac{p(\boldsymbol{K} \mid \mu)\, p(\mu)}{p(\boldsymbol{K})}$$

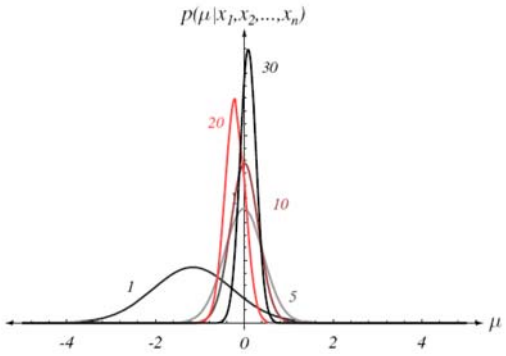
$$= \alpha \prod_{k=1}^{N} p(x_k \mid \mu)\, p(\mu) \sim N(\mu_N, \sigma_N^2);$$

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

其中 $m_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$ 为样本均值;

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

一元正态分布的贝叶斯学习



一元正态分布的贝叶斯学习

- 直接计算总体密度:

$$p(x|\boldsymbol{K}) = \int p(x|\mu)p(\mu|\boldsymbol{K})d\mu$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_N^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_N}{\sqrt{\sigma^2+\sigma_N^2}}\right)^2\right\}$$

$$\sim N(\mu_N, \sigma^2+\sigma_N^2).$$

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

非监督参数估计（简介）

- 样本类别未知，但各类条件概率密度函数的形式已知，根据所有样本估计各类密度函数中的参数。
- 非监督最大似然估计的思路:
  - 混合密度：分量密度的线性组合

$$p(\boldsymbol{x}|\theta) = \sum_{i=1}^{c} p(\boldsymbol{x}|\omega_i,\theta_i)P(\omega_i)$$

  - 似然函数和对数似然函数

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

$$H(\theta) = \ln[l(\theta)] = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i|\theta)$$

非监督参数估计（简介）

- 非监督最大似然估计的思路:
  - **最大似然估计**

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \prod_{k=1}^{N} p(x_k|\theta) = \arg\max_{\theta\in\Theta} \sum_{k=1}^{N} \ln p(x_k|\theta);$$

  - **可识别性问题**：未知参数的个数小于等于独立方程的个数。
  - **计算问题**：微分方程组

$$\nabla_{\theta_i} H(\hat{\theta}) = 0,\ i = 1,2,\cdots,c$$

    - 常采用迭代法进行参数估计。

总结：参数估计

- **最大似然估计**

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p(\boldsymbol{K}|\theta) = \arg\max_{\theta} \prod_{k=1}^{N} p(\mathbf{x}_k|\theta);$$

- **最大后验概率估计**

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\boldsymbol{K}|\theta)p(\theta) = \arg\max_{\theta} \sum_{k=1}^{N} \ln p(\mathbf{x}_k|\theta) + \ln p(\theta);$$

- **贝叶斯估计**

$$p(\theta|\boldsymbol{K}) = \frac{p(\boldsymbol{K}|\theta)p(\theta)}{\int p(\boldsymbol{K}|\theta)p(\theta)d\theta}, \hat{\theta}_{\mathrm{BE}} = E(\theta|\boldsymbol{K}) = \int_{\Theta} \theta p(\theta|\boldsymbol{K})d\theta;$$

- **贝叶斯学习**

$$p(\mathbf{x}|\boldsymbol{K}) = \int p(\mathbf{x}|\theta)p(\theta|\boldsymbol{K})d\theta.$$

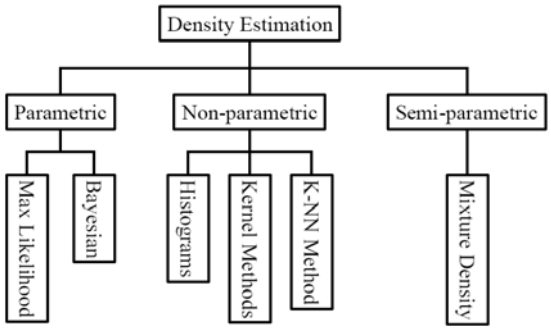
讨论：参数估计中的模型选择

- 实际工作中处理的大都是高维数据：d ⩾ 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。

非参数估计

- **非参数估计**：密度函数的形式未知，也不作假设，利用训练数据直接对概率密度进行估计。又称作模型无关方法。
  - **参数估计**需要事先假定一种分布函数，利用样本数据估计其参数。又称作基于模型的方法。
  - 任何非参数估计方法都需要选择平滑参数。
- 主要方法：
  - 直方图法
  - 核函数方法（Parzen窗法）
  - k_n-近邻法

概率密度函数的估计方法分类



直方图法

- 最简单的非参数概率密度估计方法。
- **基本思路**：$\hat{P}(\mathbf{x})$ 是 $p(\mathbf{x})$ 的一个离散近似。
  - 把观测向量 $\mathbf{x}$ 的每个分量分成 k 个等间隔小窗(bin)；$\mathbf{x} \in \boldsymbol{E}^d$，则形成 k^d 个小舱；
  - 统计落入各个小舱内的样本数

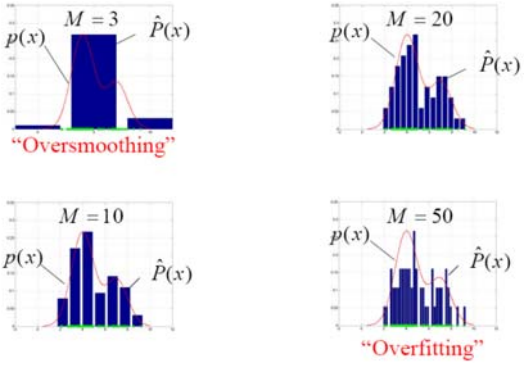
$$H(i) = \sum_{j=1}^{n} I(\mathbf{x} \in B_i). \forall i = 1.2. \cdots . m$$

  - 正规化

$$\hat{P}(i) = \frac{H(i)}{\sum_{j=1}^{m} H(j)}$$

直方图法

- 平滑参数：小窗个数/尺寸(bin size)



直方图法

- **优点**
  - 计算快速、直观易理解；
  - 直方图一旦建好，即不再需要训练数据；
  - 只保留与直方图小舱的位置和大小相关的信息；
  - 可顺序建立直方图，即每次考虑一个数据后丢弃。
- **缺点**
  - 估计的概率密度不平滑，在小舱边界不连续；
  - 对小窗个数/尺寸非常敏感；
  - 在高维空间的推广性不好（数据稀疏）。
    - 改进：Data-adaptive histogram，naive Bayes，Dependence trees...

非参数估计的基本方法

- **问题**：已知样本集 $\boldsymbol{K} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$，其中的样本均从服从 $p(\mathbf{x})$ 的总体中独立抽取，估计样本空间中任何一点的概率密度 $\hat{P}(\mathbf{x})$，以近似 $p(\mathbf{x})$ 。

- **基本思路**：用某种核函数表示某一样本对待估计的密度函数的贡献，所有样本所作贡献的线性组合（函数之和）视作对某点概率密度$p(\mathbf{x})$的估计。

$$\hat{p}_N(\mathbf{x}) = \sum_{i=1}^{n} \varphi(\mathbf{x} - \mathbf{x}_i).$$

非参数估计的基本方法

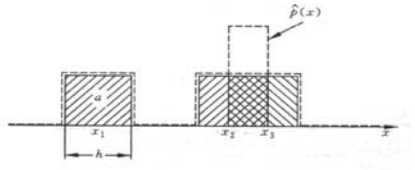


图 3.3 非参数估计的基本思路

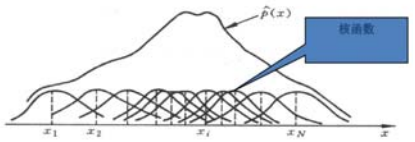


图 3.4 每个样本的贡献随距离变化的非参数估计

非参数估计的基本方法

- **基本事实**：一个向量 **x** 落在区域 ***R*** 中的概率为

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

➔ 可以通过估计概率 P 来估计概率密度函数 p。

- 如根据 $p(\mathbf{x})$ 抽取 n 个独立同分布的样本，则有

$$P(k) = \frac{n!}{k!(n-k)!}P^k(1-P)^{n-k} = B(n,P)$$

  - P：样本 **x** 落入区域 ***R*** 的概率；
  - $P(k)$：n 个样本中有 k 个落入区域 ***R*** 的概率；
  - $B(n, p)$：k 的二项分布。

非参数估计的基本方法

- $B(n, p)$ 的均值和方差

Mean:

$$\mu = E[k] = nP \Rightarrow P = E[k/n]$$

Variance:

$$\sigma^2 = E[(k-\mu)^2] = nP(1-P)$$

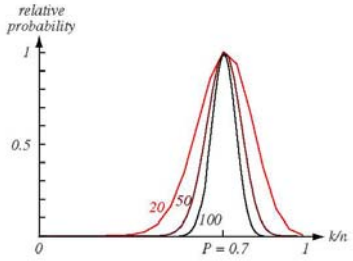
$$\Rightarrow E\left[(k/n - P)^2\right] = \frac{\sigma^2}{n^2} = \frac{P(1-P)}{n}$$

  - $E[k/n]$（即落入 ***R*** 中点的比例的期望）是 P 的一个很好的估计；
  - 样本个数 n 非常大时估计将非常准确（方差消失）；

非参数估计的基本方法

- 当 $n \to \infty$ 时，k/n 的分布逼近 δ 函数

➔ $P \simeq \frac{k}{n}$ Approximation 1

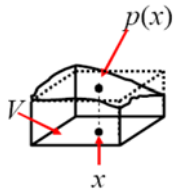


非参数估计的基本方法

- **假设**：$p(\mathbf{x})$ 是连续的，并且区域 ***R*** 足够小，以至于区间中 p 几乎没有变化，则有

$$P = \int_R p(\mathbf{x}')dx' \simeq p(\mathbf{x})V, \quad \text{Approximation 2}$$

其中，V 是区域 ***R*** 所包含的体积。



综合两个近似：

$$p(\mathbf{x}) \simeq \frac{k/n}{V}$$

非参数估计的基本方法

$$p(\mathbf{x}) \simeq \frac{k/n}{V}$$

- **讨论**：上述近似基于矛盾的假设
  - 区域 ***R*** 相对较大；即区域中包含很多的样本从而使得 P 的估计分布有非常显著的波峰。

Approximation 1

  - 区域 ***R*** 相对较小，使得 $p(\mathbf{x})$ 在积分区域内几乎没有变化。

Approximation 2

非参数估计的基本方法

- **讨论**：$\frac{k/n}{V}$ 总是 $p(\mathbf{x})$ 的空间平滑后的结果。
  - 如果希望得到 $p(\mathbf{x})$，须要求 V 趋近于零。
  - ➔ 区域 $\boldsymbol{R}$ 中可能不包含任何样本，即 $p(\mathbf{x}) = 0$。
  - 实际上，训练样本的个数 n 总是有限的，V 不能取得任意小。
  - ➔ k/n 总是有一定的变动，概率密度函数 $p(\mathbf{x})$ 总是存在一定程度的平滑效果。

非参数估计的基本方法

- **理论讨论**：假设可获得无限多的训练样本，如何估计点 $\mathbf{x}$ 处的概率密度函数？
  - 构造一系列包含 $\mathbf{x}$ 点的区域 $\boldsymbol{R}_1$，$\boldsymbol{R}_2$，…，$\boldsymbol{R}_n$。
  - 第一个区域用1个样本，第二个区域用2个样本…
  - 设 V_n 为区域 $\boldsymbol{R}_n$ 的体积，k_n 为落在区间 $\boldsymbol{R}_n$ 中的样本个数。
  - 对 $p(\mathbf{x})$ 的第 n 次估计为

$$p_n(\mathbf{x}) \simeq \frac{k_n/n}{V_n}.$$

非参数估计的基本方法

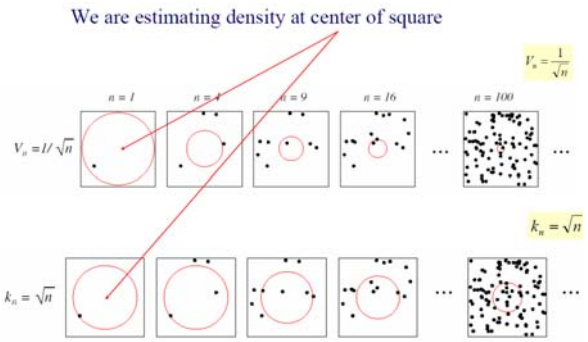
- $p_n(\mathbf{x})$ 收敛于 $p(\mathbf{x})$ 的三个必要条件

$$\lim_{n\to\infty} V_n = 0 \quad \text{— Approximation 2}$$
$$\lim_{n\to\infty} k_n = \infty \quad \text{— Approximation 1}$$
$$\lim_{n\to\infty} k_n/n = 0 \quad \text{— to allow } p_n(\mathbf{x}) \text{ to converge}$$

- 给定 n 个训练样本，如何估计 $p(\mathbf{x})$？（即如何获得满足三个必要条件的区域序列）
  - **核函数法（Parzen窗法）**：根据某个确定的体积函数来逐渐收缩一个给定的初始区间。
  - **k_N-近邻法**：确定 k_n 为 n 的某个函数，逐渐生长体积，直到最后能包含进 $\mathbf{x}$ 的 k_n 个相邻点。

非参数估计的基本方法

- 示例：Parzen窗法和k_N-近邻法



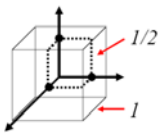
Parzen窗法

- 假设区域 $\boldsymbol{R}_n$ 是一个中心位于估计点 $\mathbf{x}$，棱长为 h_n 的 d 维超立方体，其体积为

$$V_n = h_n^d$$

- 定义**窗函数**

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



  - $\varphi(\mathbf{u})$ 表示一个中心在原点的单位超立方体。

Parzen窗法

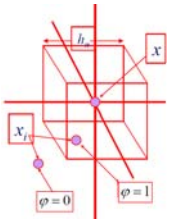
- 落入超立方体 $\boldsymbol{R}_n$ 内的样本数

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

equals 1 if $\mathbf{x}_i$ falls within hypercube centered at $\mathbf{x}$

- **Parzen窗估计**

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



  - 是一系列关于 $\mathbf{x}$ 和 $\mathbf{x}_i$ 的函数的叠加；即一种内插过程：每一个样本对估计所起的作用依赖于它到 $\mathbf{x}$ 的距离。

Parzen窗法

- Parzen窗估计 $p_n(\mathbf{x})$ 为合理的密度函数（值非负且积分为1）的条件是窗函数本身是合法的密度函数，即

$$\varphi(\mathbf{u}) \geq 0; \qquad \int \varphi(\mathbf{u})d\mathbf{u} = 1;$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)d\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n}\int \varphi(\mathbf{u})d\mathbf{u} = 1.$$

  - 窗函数可有更一般的形式，不限于超立方体函数。

Parzen窗法

- **常用窗函数**
  - 方窗
  - 正态窗： $\varphi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right) \sim N(0,1)$
  - 指数窗： $\varphi(u) = \exp(-|u|)$
  - 三角窗： $\varphi(u) = \begin{cases} 1-|u| & \text{if } |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$
  - 超球窗： $\varphi(\mathbf{u}) = \begin{cases} 1 & \text{if } \|\mathbf{u}\| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Parzen窗法

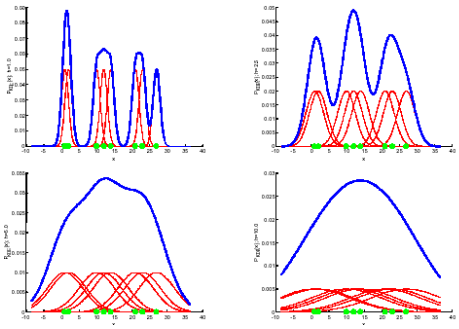
- **窗宽 h_n 的影响**
  - 定义函数及重写 $p_n(\mathbf{x})$

$$\delta_n(\mathbf{x}) = \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right); \qquad p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{x}_i);$$

  - $V_n = h_n^d$，h_n 会影响 $\delta_n(\mathbf{x})$ 的宽度和强度
    - 如 h_n 很大，δ_n 的强度很低，而且距离点 $\mathbf{x}$ 远近不同的样本的 δ_n 相差不大。$p_n(\mathbf{x})$ 是 n 个宽度较大且变化缓慢的函数的叠加，故是一个非常平滑、低分辨力的估计。
    - 如 h_n 很小，则 $\delta_n(\mathbf{x}-\mathbf{x}_i)$ 的峰值很大。$p_n(\mathbf{x})$ 是 n 个以样本点为中心的尖脉冲的叠加，统计变动很大，即是一个充满噪声的估计。

Parzen窗法

- **窗宽 h_n 的影响**
  - h_n 是平滑参数，需优化，根据样本的数量选择。



Parzen窗法

- **Parzen 估计量的统计性质**
  - $p_n(\mathbf{x})$是渐进无偏和平方误差一致估计的限制条件：
    - $p(\mathbf{x})$ 在 $\mathbf{x}$ 点连续；
    - 窗函数满足下列条件：

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

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

$$\sup_{\mathbf{u}} \varphi(\mathbf{u}) < \infty;$$

$$\lim_{\|\mathbf{u}\|\to\infty} \varphi(\mathbf{u})\prod_{i=1}^{d} u_i = 0;$$

    - 窗宽约束：

$$\lim_{n\to\infty} V_n = 0;$$

$$\lim_{n\to\infty} nV_n = \infty.$$

Parzen窗法

- **例1**： $p(\mathbf{x})$ 和 $\varphi(\mathbf{u})$ 均是正态分布

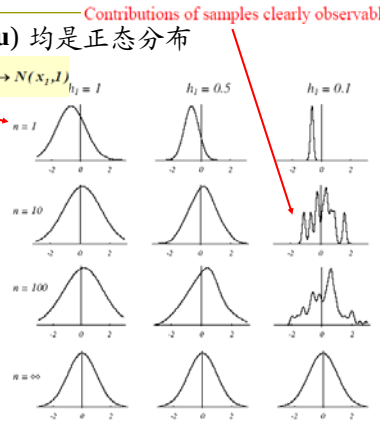
$$p_1(x) = \varphi(x-x_1) = \frac{1}{\sqrt{2\pi}}e^{-1/2(x-x_1)^2} \to N(x_1, 1)$$

$$p(x) \sim N(0,1)$$

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

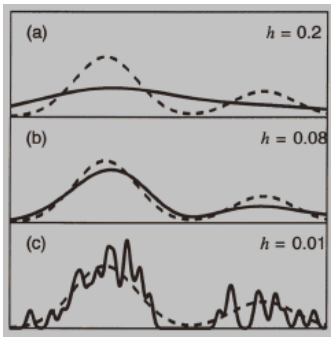
$$h_n = h_1/\sqrt{n}$$

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



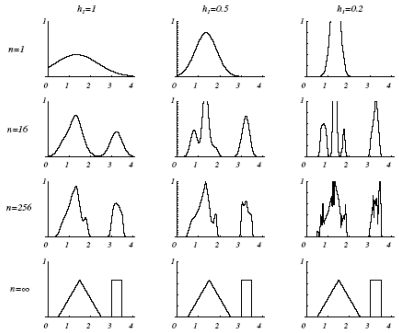
Parzen窗法

- **例2**：$\varphi(\mathbf{u})$ 是正态分布



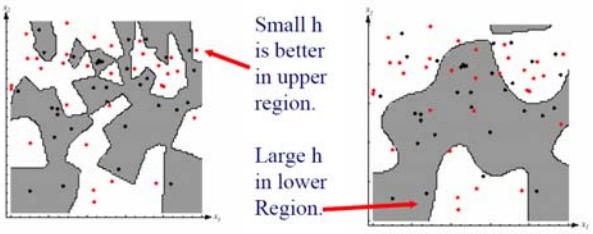
Parzen窗法

- **例3**： $p(\mathbf{x})$ 是一个均匀分布和一个三角形分布的混合分布，$\varphi(\mathbf{u})$ 是正态分布



Parzen窗法

- **讨论**
  - 基于Parzen窗估计的分类器中，我们对每一个类别都独立的估计概率密度，且根据“最大后验概率”（MAP）的原则进行分类。



Parzen窗法

- **讨论**
  - **通用性**：不需了解分布的形式都能够估计，且只要样本足够多，总可以得到可靠的收敛的结果。
  - **局限性**：时间消耗和存储器消耗惊人；有限样本的影响 — “维数灾难”：当维数较高时，样本数量无法达到精确估计的要求。
    (curse of diemensionality)

n	d	$n^{-4/(d+4)}$
16	1	0.1
32	2	0.1
178	5	0.1
3162	10	0.1
3E+13	50	0.1