第三章 概率密度函数的估计

苏智勇
可视计算研究组
南京理工大学
suzhiyong@njust.edu.cn
https://zhiyongsu.github.io

主要内容

3.1 引言

3.2 最大似然估计

3.3 贝叶斯估计与贝叶斯学习

3.4 概率密度估计的非参数方法

3.1 引言

- **模式分类的三种主要途径**
  - **估计类条件概率密度** $P(x \mid \omega_i)$

    通过$P(\omega_i)$和$P(x \mid \omega_i)$，利用贝叶斯规则计算后验概率$P(\omega_i \mid x)$，然后通过最大后验概率做出决策。

  - **直接估计后验概率** $P(\omega_i \mid x)$

    不需要先估计类条件概率密度$P(x \mid \omega_i)$。主要有K近邻方法等。

  - **直接计算判别函数**

    不需要估计$P(x \mid \omega_i)$或者$P(\omega_i \mid x)$。常见的方法有神经网络等。

3.1 引言

- **模式分类的三种途径**
  - **估计类条件概率密度** $P(x \mid \omega_i)$

    通过$P(\omega_i)$和$P(x \mid \omega_i)$，利用贝叶斯规则计算后验概率$P(\omega_i \mid x)$，然后通过最大后验概率做出决策。

    - **参数估计**：最大似然估计、贝叶斯估计等。
    - **非参数估计**：直方图法、Parzen窗方法等。

3.1 引言

- **基于样本的两步贝叶斯**

$$P(\omega_i \mid x) = \frac{P(x\omega_i)}{P(x)} = \frac{P(\omega_i)P(x \mid \omega_i)}{\sum_{j=1}^{n} P(\omega_j)P(x \mid \omega_j)}$$

  - 在上一章的学习中，我们一直**假设类的条件概率密度函数是已知的**，然后去设计贝叶斯分类器。但在实际中，这些知识往往是不知道的，这就需要**用已知的样本进行学习或训练**。也就是说利用统计推断理论中的估计方法，从样本集数据中估计这些参数。
  - **本章目的**：已知类别的**样本**（训练样本）→ **学习**或训练→ 获得类概率密度**函数** $P(x \mid \omega_i)$

3.1 引言

- **基于样本的两步贝叶斯**

$$P(\omega_i \mid x) = \frac{P(x\omega_i)}{P(x)} = \frac{P(\omega_i)P(x \mid \omega_i)}{\sum_{j=1}^{n} P(\omega_j)P(x \mid \omega_j)}$$

  - 根据训练样本估计**概率密度函数**、**先验概率**，记为$\hat{P}(x \mid \omega_i)$和 $\hat{P}(\omega_i)$
  - 根据估计的概率密度函数设计分类器
  - 希望$N \to \infty$时基于样本的估计收敛于理论结果

$$\hat{P}(x \mid \omega_i) \to P(x \mid \omega_i), \quad \text{when} \quad N \to \infty$$

$$\hat{P}(\omega_i) \to P(\omega_i), \quad \text{when} \quad N \to \infty$$

3.1 引言

- **概率密度函数估计的方法**
  - **参数估计**：总体分布/概率密度函数形式已知，用样本的统计量，去估计总体分布的未知或部分未知参数
    - **参数**是刻画总体某方面概率特性的数量。当此数量未知时，从总体抽出一个样本，用某种方法对这个未知参数进行估计就是**参数估计**。
    - 两类参数估计方法：**点估计**、**区间估计**
    - 例如，$X \sim N(\mu, \sigma^2)$：若μ，σ^2未知, 通过构造样本的函数, 给出它们的**估计值**或**取值范围**就是参数估计的内容
  - **非参数估计**：总体分布/概率密度函数形式未知、或者不符合现有任何分布模型

3.1 引言

- **参数估计**

  总体X的概率密度函数为$f(x \mid \theta)$，根据观测到的一组样本$(x_1, \ldots, x_n)$，去估计总体参数θ的过程。

  - 如果θ是已知确定的，x是变量，$f(x \mid \theta)$叫**概率函数**，描述对于不同的样本点x，其出现的概率是多少
  - 如果x是已知确定的，θ是变量，$f(x \mid \theta)$叫**似然函数**，描述对于不同的模型参数θ，出现x这个样本的概率是多少
  - **缺点**：估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。

相关概念

- **统计量**
  - **定义**：设$(X_1, X_2, \ldots, X_n)$为总体X的**样本**，$T(X_1, X_2, \ldots, X_n)$为n维实值**函数**（不含未知参数），T的取值记为$t = T(x_1, x_2, \ldots, x_n)$，称$T(X_1, X_2, \ldots, X_n)$为样本**统计量**，简称为**统计量**。
  - **常用统计量**
    - 样本均值：$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
    - 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
    - 样本k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k$

相关概念

- **参数空间**
  - 设总体分布的未知参数为θ，未知参数全部**可容许值组成的集合**称为参数空间，记为Θ。
- **点估计**
  - 利用样本数据，对未知的参数进行估计，所得到的具体数值。
    - $\hat{\theta}(X_1, X_2, \ldots, X_n)$为未知参数$\theta$的**估计量**
    - $\hat{\theta}(x_1, x_2, \ldots, x_n)$为未知参数$\theta$的**估计值**
    - 例如，已知分布是正态分布，用统计量样本均值$\bar{X}$来估计总体均值μ。
  - 常用点估计方法：**最大似然估计法**、**贝叶斯估计**、**最小二乘估计**、**矩估计法**

相关概念

- **区间估计**
  - 在推断总体参数时，根据统计量的抽样分布特征，估计出总体参数θ的一个**置信区间**(d_1, d_2)，而不是一个数值，并同时给出总体参数落在这一区间的概率。
- **估计量性质**
  - **无偏性**
    - **无偏估计量**：$E(\hat{\theta}) = \theta$，θ的估计量的数学期望等于θ
    - **渐近无偏估计量**：$\lim_{n\to\infty} E(\hat{\theta}) = \theta$，样本数趋于无穷时估计才具有无偏性

相关概念

- **估计量性质**
  - **有效性**
    - 设$\hat{\theta}_1$，$\hat{\theta}_2$是θ的两个无偏估计，如果$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$，对一切$\theta \in \Theta$成立，且不等号至少对某一$\theta \in \Theta$成立，则称$\hat{\theta}_1$比$\hat{\theta}_2$**有效**。
  - **一致性**
    - 当样本无限增多时，估计量$\hat{\theta}_n$依概率收敛于θ，则称$\hat{\theta}_n$为θ的**一致估计**，

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

3.2 最大似然估计

- 首先是由德国数学家高斯在1821年提出的。费歇在1922年重新发现了这一方法，并首先研究了这种方法的一些性质。
- **例**：假设在一个罐中放着许多白球和黑球，并假定已经知道两种球的数目之比是1:3，但不知道哪种颜色的球多。如果用放回抽样方法，从罐中取5个球，观察结果为：黑、白、黑、黑、黑，估计取到黑球的概率p.

  - $p = 1/4$：出现本次观察结果的概率为$\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right) = \frac{3}{1024}$
  - $p = 3/4$：出现本次观察结果的概率为$\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right) = \frac{81}{1024}$
  - 由于$\frac{3}{1024} < \frac{81}{1024}$，因此，认为$p = \frac{3}{4}$比$p = \frac{1}{4}$更有可能，于是 $\hat{p} = \frac{3}{4}$

这种选择一个参数使得实验结果具有最大概率的思想就是最大似然法的基本思想 。

3.2.1 最大似然估计的基本原理

- **基本假设**
  - 待估计参数θ未知但确定
  - 每一类的样本独立同分布
  - 类条件概率密度函数$P(x \mid \omega_i)$函数形式确定，只是参数θ未知
  - 不同类别的参数独立，可分别处理

3.2.1 最大似然估计的基本原理

- **构造似然函数** $L(\theta)$
  - 设总体为**离散型**，$X \sim p(x;\theta)$，$\theta \in \Theta$，θ未知。设$X_1, \cdots, X_n$是来自X的样本，$x_1, \cdots, x_n$是$X_1, \cdots, X_n$的一个样本值，则事件$\{X_1 = x_1, \cdots, X_n = x_n\}$发生的概率为

$$L(\theta) = P\{X_1 = x_1, \cdots, X_n = x_n\} = p(x_1;\theta)\cdots p(x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

似然函数

  - 设总体为**连续型**，其概率密度$f(x;\theta)$，$\theta \in \Theta$，θ未知。设$X_1, \cdots, X_n$是来自X的样本，$x_1, \cdots, x_n$是$X_1, \cdots, X_n$的一个样本值，则事件$\{X_1 = x_1, \cdots, X_n = x_n\}$发生的概率为

$$L(\theta) = P\{X_1 = x_1, \cdots, X_n = x_n\} = f(x_1;\theta)\cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

似然函数

3.2.1 最大似然估计的基本原理

- **最大似然原理**

  选择使$L(\theta)$达到最大的参数$\hat{\theta}$，作为参数θ的估计，即

$$L(\hat{\theta}(x_1, \ldots, x_n)) = \max_{\theta \in \Theta} L(\theta) \text{ 或 } \hat{\theta} = \operatorname{argmax} L(\theta)$$

  - $\hat{\theta}(x_1, \ldots, x_n)$为参数$\theta$的**最大似然估计值**
  - $\hat{\theta}(X_1, \ldots, X_n)$为参数$\theta$的**最大似然估计量**（maximum likelihood estimation, MLE）

3.2.2 最大似然估计的求解

- 一般，θ可由下式求得：

$$\frac{dL(\theta)}{d\theta}=0 \quad 或 \quad \frac{d}{d\theta}\ln L(\theta)=0$$

  - $\ln L(\theta)$称为**对数似然函数**，

- 未知参数θ可能不是一个，一般设为$\theta=(\theta_1,\theta_2,\cdots,\theta_k)$，则可由下式求得：

$$\frac{\partial L}{\partial \theta_i}=0,\quad i=1,\cdots,k \quad 或 \quad \frac{\partial \ln L}{\partial \theta_i}=0,\quad i=1,\cdots,k$$

- 用上述方法求参数的极大似然估计值有时行不通（$L(\theta)$不可导、无驻点），这时要用**最大似然原则**来求。

  - 若$L(\theta)$关于某个θ_i是单调增(减)函数，此时θ_i的最大似然估计在其边界取得。

3.2.2 最大似然估计的求解

- 例：设某种元件使用寿命X的概率密度函数为

$$f(x)=\begin{cases}2e^{-2(x-\theta)}, x\geq\theta\\ 0, \text{ otherwise}\end{cases}$$

其中 $\theta>0$ 是未知参数。设 $x_1,\ldots,x_n$ 是样本观测值，求 θ 的最大似然估计。

3.2.2 最大似然估计的求解

- **解**：
  - 似然函数$L(\theta)=\prod_{i=1}^{n}[2e^{-2(x_i-\theta)}]=2^n e^{-2\sum(x_i-\theta)}$，$x_i\geq\theta$
  - $\ln L(\theta)=n\ln 2-2\sum_{i=1}^{n}(x_i-\theta)$
  - $\frac{d\ln L(\theta)}{d\theta}=2n>0$
  - $L(\theta)$单调增加，$\theta\leq x_i$
  - θ的最大似然估计：$\hat{\theta}=\min\{x_1,x_2,\cdots,x_n\}$

3.2.3 正态分布下的最大似然估计

- **例**：设总体$X \sim N(\mu, \sigma^2)$，设 $x_1, \ldots, x_n$ 是X的样本观测值，求μ, σ^2的最大似然估计。
- **解**：

$$L(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^2(\sigma^2)^{\bar{n}}} e^{-\sum_{i=1}^{n(x_i-\mu)^2} 2\sigma^2}$$

$$\ln L = -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2)$$

$$\begin{cases} \left(\frac{\partial}{\partial \mu}\ln L\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\ \left(\frac{\partial}{\partial(\sigma^2)}\ln L\right) = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2(\sigma^2)} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{cases}$$

3.3 贝叶斯估计与贝叶斯学习

- **两大统计学派的起源**

  - **总体信息**

    进行统计推断时，总是假定总体是来自某个分布族的，并从总体中抽取一组样本。其中所假定的总体来自的分布族，就是所谓的**总体信息**。

  - **样本信息**

    从总体中抽取的样本，能给我们**样本信息**。

  - **先验信息**

    在进行试验之前，实际上已经对待估的参数有一定的了解。比如我们调查某厂的不合格率时，过去关于该厂不合格率的历史资料，这种在抽样前就已经存在的有关统计推断的信息，称为**先验信息**。

> **例** 一位常饮牛奶加茶的妇女声称，她能分辨出先倒进杯子里的是茶还是奶.对此做了十次试验，她都正确地说出来.
> 我们把分辨奶与茶次序看做是统计推断的过程，把妇女的经验看做是先验信息。如果该妇女完全没有经验，那么它十次都才对的概率非常小，在一次实验中几乎不可能发生，所以经验对她的判断起着至关重要的作用。

3.3 贝叶斯估计与贝叶斯学习

- **两大统计学派的起源**

  关于先验信息的争论，直接导致了两大统计学派的出现

  - **经典统计学派：**只利用**总体**信息和**样本**信息来进行统计推断
    - 参数虽然是未知的，但是它还是一个**数**，是固定不变的，需要做的只是找一个比较接近它的量来代替它
  - **贝叶斯统计学派：**除了利用**总体**信息和**样本**信息，还利用**先验信息**进行统计推断
    - 既然参数是不知道的，那么把它看做是**随机变量**应该是合理的，也就是不断变化的量。用**概率**来刻画未知参数的变动情况。
    - 例如，对于一个工厂而言，其次品率也不是一成不变的，在工人们精神好的时候次品率相对会低一些。

3.3 贝叶斯估计与贝叶斯学习

- **最大似然估计与贝叶斯估计的区别**
  - **参数**
    - **最大似然估计**：参数θ被当作是**固定形式的未知变量**。结合真实数据通过最大化似然函数来求解这个固定形式的未知变量。根据观测样本估计参数 θ 的**值**。
    - **贝叶斯估计**：参数θ被当作是某种**已知先验分布的随机变量**。通过贝叶斯规则将参数的先验分布转化成后验分布进行求解。根据观测样本估计参数 θ 的**分布**。
  - **计算复杂度**
    - **最大似然估计**：简单的微分运算
    - **贝叶斯估计**：复杂的多重积分
  - **准确性**

    当样本数据有限时，由于贝叶斯估计有很强的理论和算法基础，因此，贝叶斯估计的误差更小。

3.3.1 贝叶斯估计

- **基本思想**
  - **已知**：样本集 $\chi = \{x_1, x_2, \cdots, x_N\}$，待估计参数$\theta$看成具有先验分布密度$p(\theta)$的**随机变量**
  - **问题**：根据后验概率$p(\theta \mid \chi)$，求θ的贝叶斯估计θ^*
  - **方法**：利用样本概率密度分布$p(\chi \mid \theta)$；最优条件是最小错误率或者风险，**损失函数**记为$\lambda(\hat{\theta}, \theta)$
  - **目标**：θ的估计值θ^*应使估计损失的期望最小。

3.3.1 贝叶斯估计

- **贝叶斯估计步骤**

  - **核心**：利用贝叶斯公式求出参数θ的**后验概率**：$p(\theta \mid \chi) = \dfrac{p(\chi \mid \theta)p(\theta)}{\int_{\Theta} p(\chi \mid \theta)p(\theta)d\theta}$

  -确定参数θ的先验分布密度$p(\theta)$，待估参数为随机变量

  - 利用已知样本集χ，求出样本集的联合分布$p(\chi \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$，它是$\theta$的函数

  - 求出参数θ的**贝叶斯估计**（分布）：$\theta^* = \int_{\Theta} \theta p(\theta \mid \chi)d\theta$（证明过程见后）

3.3.1 贝叶斯估计

- **贝叶斯估计量 θ^*的证明**

  -设样本取值空间为E^d，参数的取值空间为Θ，用$\hat{\theta}$作为估计时的总期望风险：

  $$R = \int_{E^d}\int_{\Theta} \lambda(\hat{\theta},\theta)p(\theta,x)d\theta dx = \int_{E^d}\int_{\Theta} \lambda(\hat{\theta},\theta)p(\theta \mid x)p(x)d\theta dx$$

  - 定义在样本x下的**条件风险**为：$R(\hat{\theta} \mid x) = \int_{\Theta} \lambda(\hat{\theta},\theta)p(\theta \mid x)d\theta$

  - **总期望风险（贝叶斯风险）**：$R = \int_{E^d} R(\hat{\theta} \mid x)p(x)d\theta dx$

  - 在有限样本$\chi = \{x_1, x_2, \cdots, x_N\}$集合下：$\theta^* = \operatorname{argmin}\ R(\hat{\theta} \mid \chi) = \int_{\Theta} \lambda(\hat{\theta},\theta)p(\theta \mid \chi)d\theta$

3.3.1 贝叶斯估计

- **贝叶斯估计量 θ^*的证明**

  -决策之前需要事先定义**损失**，$\lambda(\hat{\theta}, \theta)$的具体定义不同，可得到不同的最佳贝叶斯估计

  -采用平方差损失：$\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$，则

  - 在给定样本x下θ的贝叶斯估计（分布）：$\theta^* = E[\theta \mid x] = \int_{\Theta} \theta p(\theta \mid x) d\theta$

  - 在给定样本集χ下θ的贝叶斯估计（分布）：$\theta^* = E[\theta \mid x] = \int_{\Theta} \theta p(\theta \mid \chi) d\theta$

3.3.1 贝叶斯估计

- 贝叶斯估计量 θ^*的证明

[证明]

$$R(\hat{\boldsymbol{\theta}}\mid \boldsymbol{X})=\int_{\Theta}\lambda(\hat{\boldsymbol{\theta}}\mid \boldsymbol{\theta})p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}=\int_{\Theta}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$=\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X})+E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$=\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}+\int_{\Theta}(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$+2\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

[证明]（续）

$$R(\hat{\boldsymbol{\theta}}\mid \boldsymbol{X})=\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}+\int_{\Theta}(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$+2\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$=(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

$$=(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})\left[\int_{\Theta}\boldsymbol{\theta}p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X})\int_{\Theta}p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}\right]$$

$$=(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})(E(\boldsymbol{\theta}\mid \boldsymbol{X})-E(\boldsymbol{\theta}\mid \boldsymbol{X}))=0$$

3.3.1 贝叶斯估计

[证明]（续）

$$R(\hat{\boldsymbol{\theta}}\mid \boldsymbol{X})=\int_{\Theta}(\boldsymbol{\theta}-E(\boldsymbol{\theta}\mid \boldsymbol{X}))^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}+\int_{\Theta}(E(\boldsymbol{\theta}\mid \boldsymbol{X})-\hat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$$

易见：第一项非负并且其取值与 $\hat{\boldsymbol{\theta}}$ 无关；

第二项也非负，但其取值与 $\hat{\boldsymbol{\theta}}$ 有关。

故：欲使贝叶斯风险最小化，需选择估计量使第二项最小化。

即 $\hat{\boldsymbol{\theta}}=E(\boldsymbol{\theta}\mid \boldsymbol{X})=\int_{\Theta}\boldsymbol{\theta}p(\boldsymbol{\theta}\mid \boldsymbol{X})\ d\boldsymbol{\theta}$ 证毕

结论：以上定理给出了估计待求参数的方法。

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

3.3.2 正态分布时的贝叶斯估计

- **例：总体$X \sim N(\mu, \sigma^2)$，设模型的均值μ是待估计的参数，方差σ^2已知，求μ的贝叶斯估计。**
- **解：**根据3.3.1节贝叶斯估计步骤求解

  - 利用贝叶斯公式，得到参数μ的后验概率分布：$p(\mu \mid \chi) = \dfrac{p(\chi \mid \mu)p(\mu)}{\int_{\Theta} p(\chi \mid \mu)p(\mu)d\mu}$

  - μ的先验分布为正态分布，假设其均值为μ_0，方差为σ_0^2，即$p(\mu) = \dfrac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\dfrac{1}{2\sigma_0^2}(x-\mu_0)^2\right)$

  - σ^2已知，求出样本集的联合分布$p(\chi|\mu) = \prod_{i=1}^{N} p(x_i|\mu)$，其中，各样本的概率密度函数

  $p(x_i \mid \mu) = \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{1}{2\sigma^2}(x_i-\mu)^2\right)$

3.3.2 正态分布时的贝叶斯估计

- **例：总体$X \sim N(\mu, \sigma^2)$，设模型的均值μ是待估计的参数，方差σ^2已知，求μ的贝叶斯估计。**

- 因为各样本独立抽取，所以

$$p(\mu \mid \chi) = a\prod_{i=1}^{N} p(x_i \mid \mu)p(\mu)，其中a = \frac{1}{\int_{\Theta} p(\chi \mid \mu)p(\mu)d\mu}$$为比例因子，只和χ有关，与μ无关，可视为对估计出的后验概率进行**归一化**的常数项

$$-\prod_{i=1}^{N} p(x_i \mid \mu)p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\right)$$

3.3.2 正态分布时的贝叶斯估计

- **例：总体$X \sim N(\mu, \sigma^2)$，设模型的均值μ是待估计的参数，方差σ^2已知，求μ的贝叶斯估计。**

- $$p(\mu \mid \chi) = a' \exp\left\{ -\frac{1}{2}\left[\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N} X_i + \frac{\mu_0}{\sigma_0^2} \right)\mu \right] \right\}$$

  - a'包含了所有和μ无关的因子
  - $p(\mu \mid \chi)$是关于μ的二次函数的指数函数，所以仍然是一个正态分布函数

3.3.2 正态分布时的贝叶斯估计

- **解：**

由于$p(\mu|\chi)$也满足正态分布，可以写为正态形式：$p(\mu|\chi)=\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2\right)$

$$p(\mu|\chi)=a'\exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}X_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

-比较上述两式，对应系数应该相等

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\frac{1}{N}\sum_{i=1}^{n}x_i+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_0$$

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

3.3.2 正态分布时的贝叶斯估计

- **解：**

将μ_N，σ_N^2带入$p(\mu|\chi)$，利用公式$\theta^* = \int_\Theta \theta p(\theta \mid \chi) d\theta$，得到参数$\mu$的贝叶斯估计：

$$\hat{\mu} = \int \mu p(\mu|\chi) \mathrm{d}\mu = \int \frac{\mu}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right] \mathrm{d}\mu = \mu_N$$

$$p(\mu|\chi) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_N^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_N}{\sqrt{\sigma^2+\sigma_N^2}}\right)^2\right] \sim N(\mu_N, \sigma^2+\sigma_N^2)$$

3.3.3 贝叶斯学习

- **样本概率密度函数的贝叶斯学习**

  求参数出θ的后验概率$p(\theta|\chi)$之后，利用贝叶斯估计量来估计样本概率密度函数的参数

  -已知：样本$\chi^N=\{x_1,x_2,\cdots,x_N\}$，$\theta^*=\int_\Theta \theta p(\theta\mid\chi^N)d\theta$，其中$p(\theta\mid\chi^N)=\dfrac{p(\chi^N\mid\theta)p(\theta)}{\int_\Theta p(\chi^N\mid\theta)p(\theta)d\theta}$

  -当$N>1$时，由于各次抽样**条件独立**，则$p(\chi^N\mid\theta)=p(x_N\mid\theta)p(\chi^{N-1}\mid\theta)$，因此

  $$p(\theta\mid\chi^N)=\frac{p(x_N\mid\theta)p(\chi^{N-1}\mid\theta)}{\int_\Theta p(x_N\mid\theta)p(\chi^{N-1}\mid\theta)d\theta}$$

  -递推的贝叶斯估计：$p(\theta)=p(\theta\mid\chi^0)$，$p(\theta\mid x_1)$，$p(\theta\mid x_1,x_2)$，…，$p(\theta\mid x_1,x_2,\cdots,x_N)$，…，可以得到一系列的对概率密度函数参数的估计，这一过程也成为贝叶斯学习

3.3.3 贝叶斯学习

采取分而治之的策略，只需要做：

$\mathcal{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_n\}$ ➡ $p(\boldsymbol{X}\mid\mathcal{X})$

求解思路 ➡ $p(\boldsymbol{\theta}\mid\mathcal{X})$（$\boldsymbol{\theta}$的后验概率密度）

$p(\boldsymbol{X},\boldsymbol{\theta})$ 联合概率密度

收敛？

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

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

$$p(\boldsymbol{X},\boldsymbol{\theta})=p(\boldsymbol{X}\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta}) \Rightarrow p(\boldsymbol{X},\boldsymbol{\theta}\mid\mathcal{X})=p(\boldsymbol{X}\mid\boldsymbol{\theta},\mathcal{X})\,p(\boldsymbol{\theta}\mid\mathcal{X})$$

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

引入标记 $\mathcal{X}^N=\{\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_N\}$

假定样本集合中各样本是独立抽取的

$$p(\mathcal{X}^N\mid\boldsymbol{\theta})=p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\boldsymbol{X}_{N-1}\mid\boldsymbol{\theta})\Lambda\,p(\boldsymbol{X}_2\mid\boldsymbol{\theta})\,p(\boldsymbol{X}_1\mid\boldsymbol{\theta})=p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\mathcal{X}^{N-1}\mid\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}\mid\mathcal{X}^N)=\frac{p(\mathcal{X}^N\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int_{\Theta}p(\mathcal{X}^N\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$ 贝叶斯公式

$$=\frac{p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\mathcal{X}^{N-1}\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int_{\Theta}p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\mathcal{X}^{N-1}\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$ 独立性

$$=\frac{p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\mathcal{X}^{N-1})\,p(\boldsymbol{\theta}\mid\mathcal{X}^{N-1})}{\int_{\Theta}p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\mathcal{X}^{N-1})\,p(\boldsymbol{\theta}\mid\mathcal{X}^{N-1})\,d\boldsymbol{\theta}}$$ 贝叶斯公式

$$=\frac{p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta}\mid\mathcal{X}^{N-1})}{\int_{\Theta}p(\boldsymbol{X}_N\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta}\mid\mathcal{X}^{N-1})\,d\boldsymbol{\theta}}$$ 约去 $p(\mathcal{X}^{N-1})$

3.3.3 贝叶斯学习

引入标记 $\mathcal{X}^N=\{\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_N\}$

假定样本集合中各样本是独立抽取的

$$p(\mathcal{X}^N|\boldsymbol{\theta})=p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{X}_{N-1}|\boldsymbol{\theta})\Lambda\ p(\boldsymbol{X}_2|\boldsymbol{\theta})\ p(\boldsymbol{X}_1|\boldsymbol{\theta})=p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\mathcal{X}^{N-1}|\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}|\mathcal{X}^N)=\frac{p(\mathcal{X}^N|\boldsymbol{\theta})\ p(\boldsymbol{\theta})}{\int_{\Theta}p(\mathcal{X}^N|\boldsymbol{\theta})\ p(\boldsymbol{\theta})\ d\boldsymbol{\theta}}$$ 贝叶斯公式

$$=\frac{p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\mathcal{X}^{N-1}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})}{\int_{\Theta}p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\mathcal{X}^{N-1}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})\ d\boldsymbol{\theta}}$$ 独立性

$$=\frac{p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\mathcal{X}^{N-1})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})}{\int_{\Theta}p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\mathcal{X}^{N-1})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})\ d\boldsymbol{\theta}}$$ 贝叶斯公式

$$=\frac{p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})}{\int_{\Theta}p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})\ d\boldsymbol{\theta}}$$ 约去 $p(\mathcal{X}^{N-1})$

实现参数$\boldsymbol{\theta}$在线学习的递推公式

$$p(\boldsymbol{\theta}|\mathcal{X}^N)=\frac{p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})}{\int_{\Theta}p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})\ d\boldsymbol{\theta}}$$

$N=1$时 $p(\boldsymbol{\theta}|\mathcal{X}^0)=p(\boldsymbol{\theta})$

$N=2$时 $p(\boldsymbol{\theta}|\mathcal{X}^1)=\frac{p(\boldsymbol{X}_1|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^0)}{\int_{\Theta}p(\boldsymbol{X}_1|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^0)\ d\boldsymbol{\theta}}$

.....

$N=N$时 $p(\boldsymbol{\theta}|\mathcal{X}^N)=\frac{p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})}{\int_{\Theta}p(\boldsymbol{X}_N|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1})\ d\boldsymbol{\theta}}$

.....

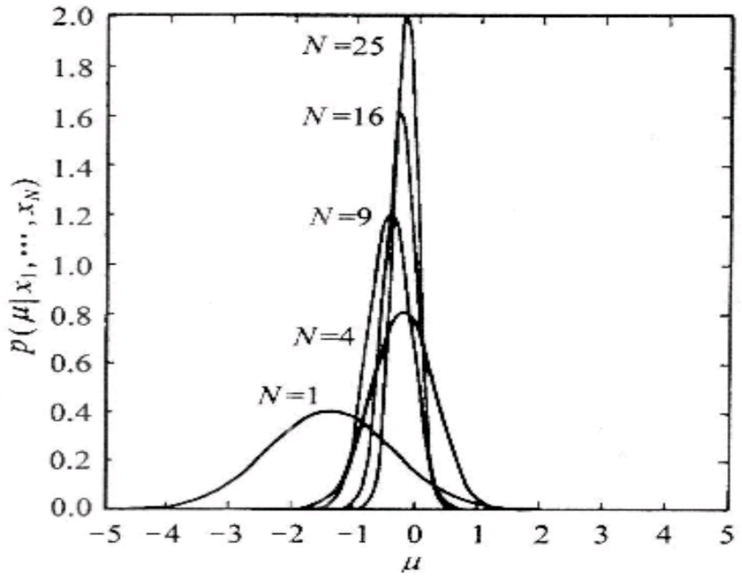
$$p(\boldsymbol{\theta}),\ p(\boldsymbol{\theta}|\mathcal{X}^1),\ p(\boldsymbol{\theta}|\mathcal{X}^2),\Lambda\ p(\boldsymbol{\theta}|\mathcal{X}^{N-1}),\ p(\boldsymbol{\theta}|\mathcal{X}^N),\Lambda$$

3.3.3 贝叶斯学习

$$p(\boldsymbol{\theta}), p(\boldsymbol{\theta}\mid\mathcal{X}^1), p(\boldsymbol{\theta}\mid\mathcal{X}^2),\Lambda\ p(\boldsymbol{\theta}\mid\mathcal{X}^{N-1}), p(\boldsymbol{\theta}\mid\mathcal{X}^N),\Lambda$$

随着N值的增加，相应后验概率密度一般会变得越来越尖锐。

若上述概率密度函数序列在 $N\to\infty$时收敛于以真实参数θ为中心的δ函数，则称相应的学习过程为贝叶斯学习。



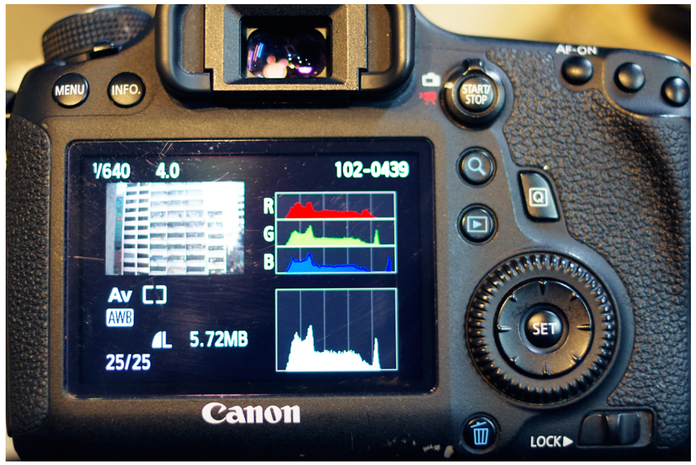
$$\lim_{N\to\infty} p(\boldsymbol{X}\mid\mathcal{X}^N)$$
$$= p(\boldsymbol{X}\mid\mathcal{X}^{N\to\infty})$$
$$= p(\boldsymbol{X}\mid\hat{\boldsymbol{\theta}}=\boldsymbol{\theta})$$
$$= p(\boldsymbol{X})$$

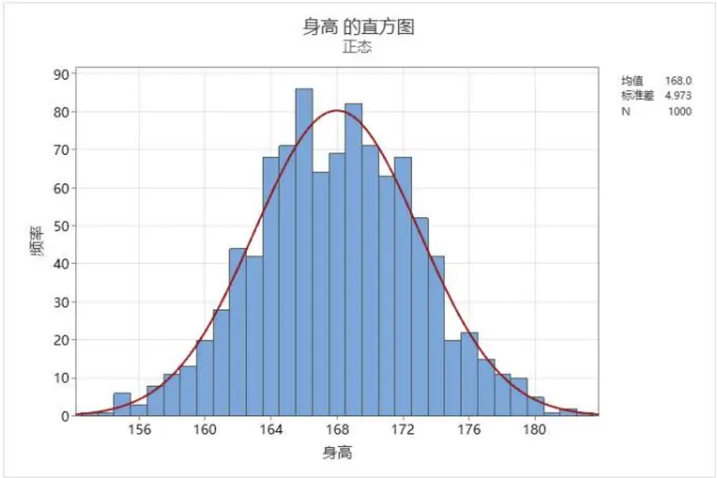
3.4 概率密度估计的非参数方法

- **参数估计方法的局限**
  - 已知总体分布/概率密度函数形式
  - 估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布
- **非参数方法基本思想**
  - 样本的密度函数形式未知
  - 有些情况样本的分布很难用函数形式描述
  - 直接用样本估计整个函数
  - 可以看作从所有可能的函数中进行选择

3.4.1 直方图法

- **直方图构造**
  - 把总数为N的样本x的每个分量分成k个等间隔小窗，每个小舱体积为V
  - 统计落入小舱内的样本数目q_i
  - 把小舱内的概率密度看作常数，用$\frac{q_i}{NV}$作为其估计值





3.4.1 直方图法

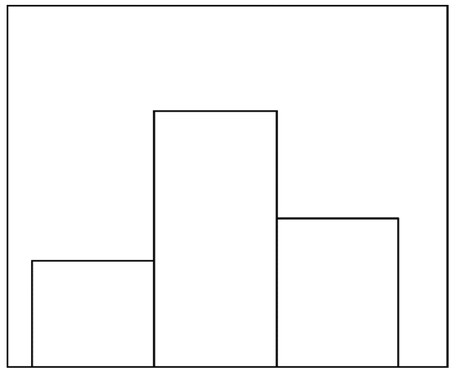
- **非参数估计**

问题：已知样本集$\chi = \{x_1, x_2, \cdots, x_N\}$中的样本(来自同一个类别)是从服从密度函数$p(x)$的总体中独立抽取出来的，求$p(x)$的估计$\hat{p}(x)$。

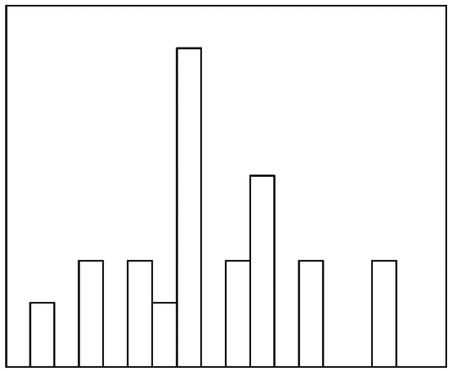
- 样本所在空间的某个区域R内，随机向量落入区域R范围的概率为：$P_R = \int_R p(x)dx$
- 在样本集χ中有k个落入区域R的概率为：$P_k = C_N^k P_R^k (1 - P_R)^{N-k}$
- k的期望值：$E[k] = NP_R$
- P_R的估计为：$\hat{P}_R = \frac{k}{N}$
- 当$p(x)$连续且V足够小时，假定$p(x)$在区域R内为常数：$P_R = \int_R p(x)dx = p(x)V$
- $\hat{P}_R = \frac{k}{N} = p(x)V \quad \Rightarrow \quad \hat{p}(x) = \frac{k}{NV}$

3.4.1 直方图法

- **非参数估计**
  - **小舱的选择原则**：与样本总数相适应
    - ✓过大，则估计的密度函数**粗糙**
    - ✓过小，则估计的密度函数**不连续**
    - ✓样本趋于无穷多时的收敛条件为：



(a) 小窗过宽

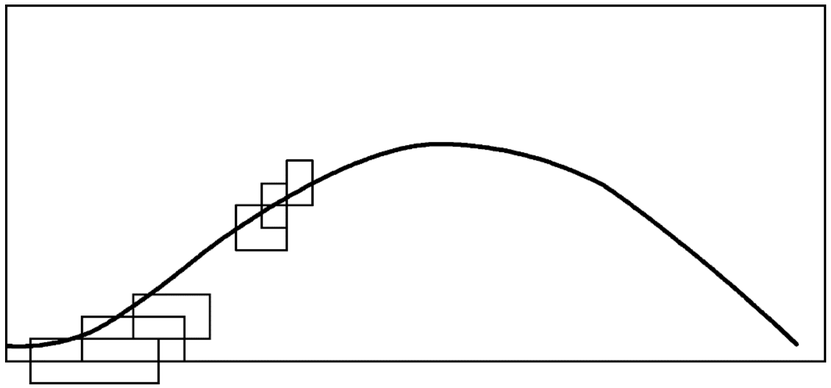


(b) 小窗过窄

$$\lim_{n\to\infty} V_n = 0, \quad \lim_{n\to\infty} k_n = 0, \quad \lim_{n\to\infty} \frac{k_n}{n} = 0$$

3.4.2 k_N近邻估计方法

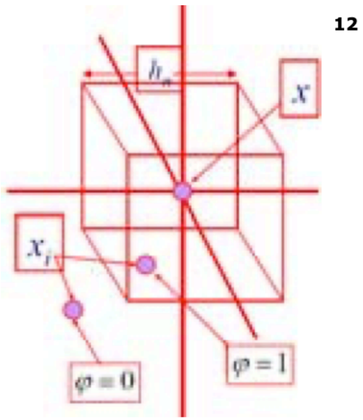
- **基本步骤**
  - 先确定k_N：总样本为N时每个小舱内的样本数
  - 求x处的密度估计$\hat{p}(x)$时，调整包含x的小舱体积，直到落入k_N个样本，估算 $\hat{p}(x) = \dfrac{k_N}{NV}$
  - 在x取值范围内以每一点为中心进行估计
  - 样本密度高的区域V较小、密度低的区域V增大
  - 需要确定合适的k_N和N的关系：如$k_N = a \times \sqrt{N}$



3.4.3 Parzen窗法

- **定义**：用窗函数（核函数）$\varphi(x)$估计概率密度的方法
- **基本步骤**
  - 固定小舱体积：$x \in R^d$，小舱为超立方体，h为棱长，$V = h^d$
  - 滑动小舱来估计每**点**(x_i)处的概率密度（直方图法估计小舱内的**平均**密度）
  - 对于任意一点的密度估计表达式：$\hat{p}(x) = \frac{1}{NV} \sum_{i=1}^{N} \varphi(\frac{x - x_i}{h})$

3.4.3 Parzen窗法



- **基本步骤**
  - 定义核函数（窗函数），衡量观测样本x_i对在x处的概率密度估计的贡献：

$$K(x, x_i) = \frac{1}{V}\varphi(\frac{x - x_i}{h})$$

    - $\varphi(\frac{x - x_i}{h})$为$d$维单位方窗函数，用于判断观测样本$x_i$是否落在以$x$为中心，$h$为棱长的小舱内：

$$\varphi(\left[u_1, u_2, \cdots, u_d\right]^{\top}) = \begin{cases} 1 & , \ |\mu_j| \le \frac{1}{2}, \ j = 1,2,\cdots,d \\ 0 & , \ \text{otherwise} \end{cases}$$

    - N个观测样本落入以x为中心的超立方体内的样本数：$k_N = \sum_{i=1}^{N} \varphi\left(\frac{x - x_i}{h}\right)$

3.4.3 Parzen窗法

- **基本步骤**
  - 概率密度估计则是每一点上观测样本贡献的平均：$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x, x_i)$
  - 常见核函数：方窗、高斯窗、超球窗等

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：$\varphi(u) = \begin{cases} 1-|u| & \text{if } |u| \le 1 \\ 0 & \text{otherwise} \end{cases}$
- 超球窗：$\varphi(\mathbf{u}) = \begin{cases} 1 & \text{if } \|\mathbf{u}\| \le 1 \\ 0 & \text{otherwise} \end{cases}$