第三章
概率密度函数的估计

任课教师：柳欣老师

email：starxliu@163.com

模式识别案例库小组任务

【案例库名称】

【案例描述】
该部分详细描述案例的背景、应用场景、面临的问题，以及案例可行性分析

【案例实现过程描述】
该部分详细描述案例实现的技术方案，包括但不限于实现的模型，框架，技术路线，步骤、算法描述等。

【案例结果展示】
该部分主要显示案例库实现的结果，包括结果图、表以及相应的介绍。

【提交的资料】

提交资料：案例库文档，源代码，开发环境、代码使用说明，运行截图；案例库介绍PPT （用于汇报）；

采集比对主界面

产品性能

测试的硬件环境是：CPU 主频 2.33GHZ，内存 4G

目标库容：1000 人；目标库来源：员工照片

规格项	描 述
最小人脸尺寸	60×60 px
人脸特征大小	5K Byte
人脸检测速度	10ms
眼睛定位速度	22ms
特征提取速度	98ms
比对速度	30 万次/秒
比对性能	正确识别率：85.2% 错误识别率：0.55% 错误拒绝率：14.7%
报警时间	小于 3 秒钟

3.1 引言

主要估计方法

- **参数估计**
  - 先假定研究问题具有某种数学模型，如正态分布，二项分布，再用已知类别的学习样本估计里面的参数（最大似然估计，bayes估计）
- **非参数估计**
  - 不假定数学模型，直接用已知类别的学习样本先验知识估计数学模型。（Parzen窗法和Kn-近邻方法）

3.1 引言

- **分类器设计**

◆ **贝叶斯决策需要已知两种知识：**

- 各类的先验概率$P(\omega_i)$
- 各类的条件概率密度函数$\boldsymbol{p(x|\omega_i)}$

◆ **实际问题**

- 已知有限数目样本

对未知样本分类

- **基于样本的两步贝叶斯决策**
  - 利用样本集估计$P(\omega_i)$和$p(x|\omega_i)$，记 $\hat{p}(x|\omega_i)$ 和 $\hat{P}(\omega_i)$
  - 基于上述估计值设计贝叶斯分类器
  - 希望：当样本数N→∞时，如此得到的分类器能收敛于理论上的最优解

3.1 引言

- **本章问题**
  - 如何利用样本集估计
  - 估计量的性质如何：收敛；无偏性、有效性、一致性
  - 利用样本集估计错误率的方法

- **类的先验概率$P(\omega_i)$的估计：**
  - 用训练数据中各类出现的频率来估计
  - 依靠经验

- **类条件概率密度函数的估计：两大类方法**
  - **参数估计：**概率密度函数的形式已知，而表征函数的参数未知，需要通过训练数据来估计
    - **矩估计**、最大似然估计、Bayes估计
  - **非参数估计：**概率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
    - Parzen窗法和k_n-近邻法

3.2 参数估计

- **参数估计的基本概念**
  - 统计量：样本集X={$\mathbf{x}_1$, $\mathbf{x}_2$,…, $\mathbf{x}_N$}的某种函数d(X)
  - 参数空间：总体分布的未知参数θ所有可能取值组成的集合 Θ
  - 点估计、估计量和估计值：
    - 点估计：构造一个统计量$d(\mathbf{x}_1, \mathbf{x}_2 ,\dots, \mathbf{x}_N)$作为参数$\theta$的估计

θ的**估计量**$\hat{\theta} = d(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$是样本集的函数

它对样本集的一次实现称为**估计值**

  - 区间估计：求参数的置信区间

3.2 参数估计

- 估计量的评价标准：
  - 无偏性：估计量的数学期望 $E(\hat{\theta})=\theta$

    渐进无偏：当样本数趋于无穷时估计才具有无偏性

    $$\lim_{N\to\infty}(E(\hat{\theta})-\theta)=0$$

  - 一致性：对任意ε样本数趋于无穷时，依概率趋于θ：

    $$\lim_{N\to\infty}P(\left|\hat{\theta}-\theta\right|>\varepsilon)=0$$

- 无偏性、有效性都只是说明对于多次估计来说，估计量能以较小的方差平均地表示其真实值，并不能保证具体的一次估计的性能；而一致性则保证当样本数无穷多时，每次的估计量都将在概率意义上任意地接近其真实值

3.2.1 最大似然估计

- **假设**
  - (1) 估计的**参数θ是确定而未知的**，而Bayes估计方法则视θ为随机变量
  - (2) 样本集可按类别分开，样本集$\mathbf{X}_i$中的样本都是从概率密度为$p(\mathbf{x}|\omega_i)$的总体中独立抽取出来的(i.i.d条件)
  - (3) $p(\mathbf{x}|\omega_i)$ 具有某种确定的函数形式，只是其参数θ_i未知。为表示$p(\mathbf{x}|\omega_i)$同θ_i有关，记作$p(\mathbf{x}|\omega_i,\theta_i)$
  - (4)各类样本只包含本类分布信息，不同类别的参数是独立的，这样就可以分别对每一类进行处理
- **每个类别都用相同的方法，表示类别的下标可以省略**

3.2.1 最大似然估计

- **问题描述**
  - 独立地按概率密度$p(\mathbf{x}|\theta)$抽取样本集X=$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$，用X估计未知参数$\theta$
- **似然函数（likelihood function）**
  - 相对于样本集的**θ的函数**，是N个随机变量的联合密度

$$l(\boldsymbol{\theta}) = p(X \mid \boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \mid \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

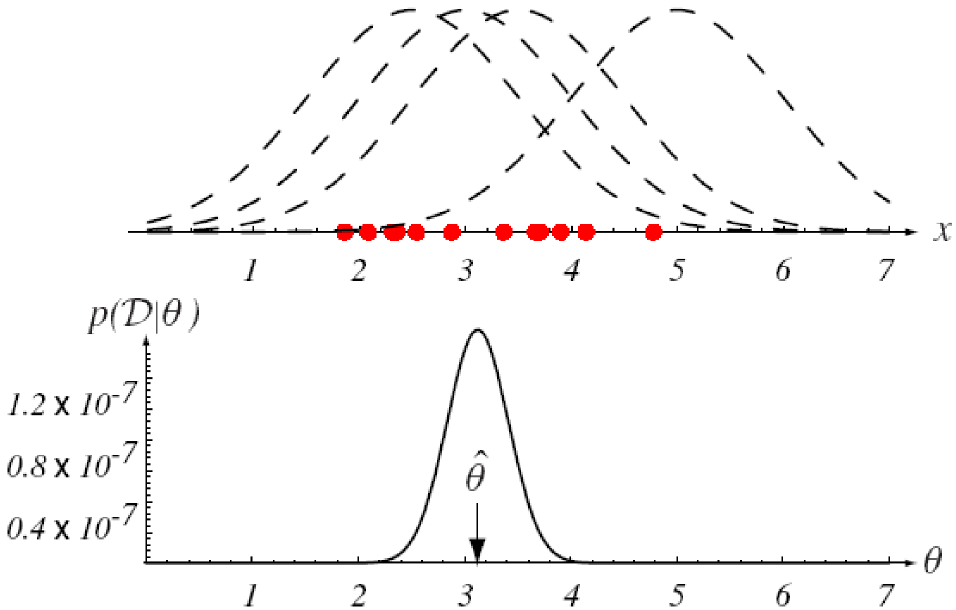
- **最大似然估计量**
  - 使似然函数极大化

$$\hat{\boldsymbol{\theta}} = d(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta})$$

  - 直观理解：参数θ的最大似然估计就是最符合给定的观测样本集的那一个

3.2.1 最大似然估计

- 一维正态分布示例



- 注意似然函数$p(X|\theta)$和条件概率密度函数$p(\mathbf{x}|\theta)$的区别：

前者是θ的函数，并不表示概率密度，后者是以θ为参数的关于x的函数

3.2.1 最大似然估计

- **最大似然估计量的求解**

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta})$$

- $\boldsymbol{\theta}$为标量，似然函数连续可微，最大似然估计量为微分方程的解

$$\frac{dl(\theta)}{d\theta} = 0$$

- 取似然函数的对数

$$H(\theta) = \ln l(\theta) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \theta)$$

$$\frac{dH(\theta)}{d\theta} = 0$$

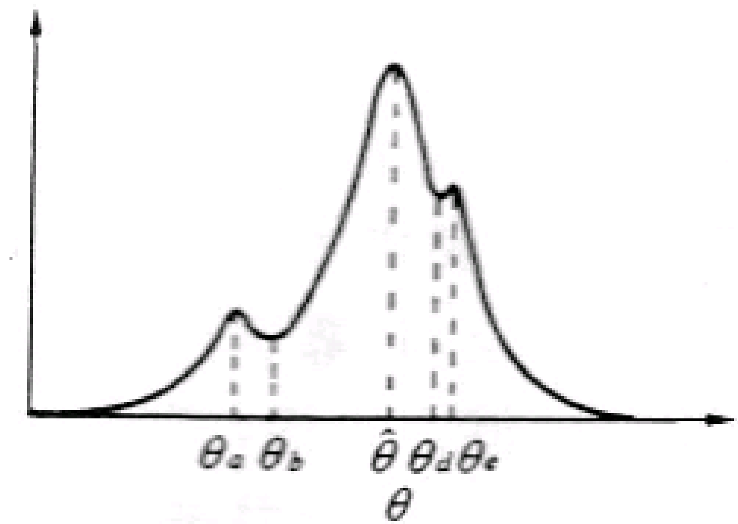
3.2.1 最大似然估计

- 当θ为有$\boldsymbol{S}$个分量的未知向量

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_S]$$

- 记梯度算子

$$\nabla_\theta = [\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_S}]^T$$



- 最大似然估计的必要条件

$$\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta})|_{\hat{\theta}_{ML}} = \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k | \boldsymbol{\theta})|_{\hat{\theta}_{ML}} = 0$$

- 如必要条件有多解，需从中求使似然函数最大者
- 若不满足连续可导，则无一般性方法，用其它方法求最大

3.2.1 最大似然估计

- 正态分布的最大似然估计

（1）**μ未知，Σ已知的情况**

- $p(x|\theta)$服从正态分布，待估参数为$\theta = \mu$

$$\ln p(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu)$$

$$\nabla_\mu \ln p(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu)$$

$$\nabla_\theta H(\theta) = \sum_{k=1}^{k=n} \Sigma^{-1}(x_k - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{k=1}^{k=n} x_k$$

- 即**均值的最大似然估计等于样本均值**

3.2.1 最大似然估计

（2）$\boldsymbol{\mu}$ 和$\boldsymbol{\Sigma}$均未知的情况

- $$p(x_k \mid \theta_1 = \mu, \theta_2 = \sigma^2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp(-\frac{(x_k - \theta_1)^2}{2\theta_2})$$

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

$$\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta})\big|_{\hat{\theta}_{ML}} = \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(x_k \mid \boldsymbol{\theta})\big|_{\hat{\theta}_{ML}} = 0$$

$$\left.\begin{aligned} \frac{\partial}{\partial\theta_1}\ln p(x_k \mid \theta_1, \theta_2) &= \frac{1}{\theta_2}(x_k - \theta_1) \\ \frac{\partial}{\partial\theta_2}\ln p(x_k \mid \theta_1, \theta_2) &= -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2{}^2} \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} \hat{\mu} &= \frac{1}{N}\sum_{k=1}^{N} x_k \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu})^2 \end{aligned}\right.$$

3.2.1 最大似然估计

- 多元正态分布参数最大似然估计

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k \qquad \hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

- **最大似然估计是一致估计**
- **均值估计是无偏的，协方差矩阵估计是有偏的。**
- **协方差矩阵的无偏估计是：**

如果一个变量的期望等于他的理想值，那么就称该变量无偏；否则称为有偏。

无偏估计的要求就是：**估计出来的参数**的数学期望等于被估计参数的真实值。

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

课后习题

均值估计是无偏的，方差估计是有偏的

$$p_i = P(X = x_i) \tag{1.1}$$

$$E(X) = \sum_{i=1}^{n} x_i p_i \tag{1.2}$$

$$D(X) = E\left(\left(X - E(X)\right)^2\right) = \sum_{i=1}^{n} \left(x_i - E(X)\right)^2 p_i \tag{1.3}$$

D(x)为样本的总体方差

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.4}$$

$$E(\overline{X}) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu \tag{1.5}$$

$$D(\overline{X}) = \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n} \tag{1.6}$$

Sample Variance is defined as equation(1.7):

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{1.7}$$

The unbiasedness of Sample Mean and Sample Variance is as follows:

$$\begin{cases} E(\overline{X}) = E(X) \\ E(S^2) = D(X) \end{cases} \tag{1.8}$$

$$D(X) = E(X^2) - (E(X))^2$$
$$E(X^2) = D(X) + (E(X))^2$$

$$E(X_i) = E(\overline{X}) = \mu$$
$$D(X_i) = \sigma^2, D(\overline{X}) = \frac{\sigma^2}{n}$$

$$E(X_i^2) = D(X_i) + \mu^2 = \sigma^2 + \mu^2$$
$$E(\overline{X}^2) = D(\overline{X}) + \mu^2 = \frac{\sigma^2}{n} + \mu^2$$

$$\begin{aligned} E(\overline{X}X_i) &= E\Big(\frac{X_1X_i + X_2X_i + \cdots + X_i^2 + \cdots + X_nX_i}{n}\Big) \\ &= \frac{1}{n}\Big(\sum_{\substack{k=0 \\ k\neq i}}^{n} E(X_kX_i) + E(X_i^2)\Big) \\ &= \frac{1}{n}((n-1)\mu^2 + (\sigma^2 + \mu^2)) \\ &= \frac{\sigma^2}{n} + \mu^2 \\ &= E(\overline{X}^2) \end{aligned}$$

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}E\Big(\sum_{i=1}^{n}(X_i - \overline{X})^2\Big) \\ &= \frac{1}{n-1}E\Big(\sum_{i=1}^{n}\big(X_i^2 - 2\overline{X}X_i + \overline{X}^2\big)\Big) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\Big(E\big(X_i^2 - 2\overline{X}X_i + \overline{X}^2\big)\Big) \\ &= \frac{1}{n-1}\Big(2n\mu^2 + (n+1)\sigma^2 - 2\sum_{i=1}^{n}\big(E(\overline{X}X_i)\big)\Big) \\ &= \frac{1}{n-1}\Big(2n\mu^2 + (n+1)\sigma^2 - 2n(\frac{\sigma^2}{n} + \mu^2)\Big) \\ &= \sigma^2 \\ &= D(X) \end{aligned}$$

当两个变量不相关时正确，乘积的期望等于期望的乘积

3.2.2 贝叶斯估计和贝叶斯学习

- **最大后验概率估计-Maximum a posteriori (MAP)**
  - 用一组样本集X $=\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ 估计未知参数 θ
  - 未知参数 θ 视为随机变量，先验分布为 $p(\theta)$，而在已知样本集X出现的条件下的后验概率为 $p(\theta|X)$

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\theta \,|\, \mathrm{X})$$

$$= \underset{\theta}{\operatorname{argmax}} \frac{p(\mathrm{X} \,|\, \theta)\, p(\theta)}{p(\mathrm{X})}$$

$$= \underset{\theta}{\operatorname{argmax}}\, p(\mathrm{X} \,|\, \theta)\, p(\theta)$$

3.2.2 贝叶斯估计和贝叶斯学习

- 贝叶斯决策问题与贝叶斯估计问题

◆ 贝叶斯决策问题:

样本$\mathbf{x}$

决策a_i

真实状态ω_j

状态空间$\mathcal{A}$是离散空间

先验概率$P(\omega_j)$

◆ 贝叶斯参数估计问题：

样本集$\mathbf{X}$

估计量$\hat{\theta}$

真实参数 θ

参数空间 Θ是连续空间

参数的先验分布$p(\theta)$

3.2.2 贝叶斯估计和贝叶斯学习

- 参数估计的条件风险：给定x条件下，估计量的期望损失

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

- 参数估计的贝叶斯风险：估计量的条件风险的期望

$$R = \int_{E^d} R(\hat{\theta}\,|\,\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- 贝叶斯估计：使风险最小的估计

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\arg\max}\, R(\hat{\theta}\,|\,\mathbf{x})$$

3.2.2 贝叶斯估计和贝叶斯学习

- 损失函数定义为误差平方：

$$\boxed{\lambda(\hat{\theta},\theta)=(\theta-\hat{\theta})^2}$$

$$\lambda(\hat{\theta},\theta)=\left|\theta-\hat{\theta}\right|$$

绝对误差损失函数

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

- **定理 3.1:** 如果定义损失函数为误差平方函数，则有：

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

3.2.2 贝叶斯估计和贝叶斯学习

- **贝叶斯估计的步骤**

1. 确定θ的先验分布 $p(\theta)$
2. 由样本集$\mathbf{X}=\{x_1, x_2, \ldots, x_N\}$求出样本联合分布：$p(\mathbf{X}|\theta)$
3. 计算θ的后验分布

$$p(\theta|\mathrm{X})=\frac{p(\mathrm{X}|\theta)p(\theta)}{\int_{\Theta}p(\mathrm{X}|\theta)p(\theta)d\theta}$$

4. 计算贝叶斯估计

$$\hat{\theta}_{\mathrm{BE}}=\int_{\Theta}\theta p(\theta|\mathrm{X})d\theta$$

3.2.2 贝叶斯估计和贝叶斯学习

- 贝叶斯学习
  - 利用**θ**的先验分布 ***p*(*θ*)**及样本提供的信息求出**θ**的后验分布***p*(*θ*| X)**，然后直接求总体分布***p*(x| X)**

$$p(\mathbf{x}|\mathcal{X})=\int p(\mathbf{x},\theta|\mathcal{X})d\theta=\int p(\mathbf{x}|\theta)p(\theta|\mathcal{X})d\theta$$

  - 递推贝叶斯方法—学习(收敛于真实参数为中心的δ函数)

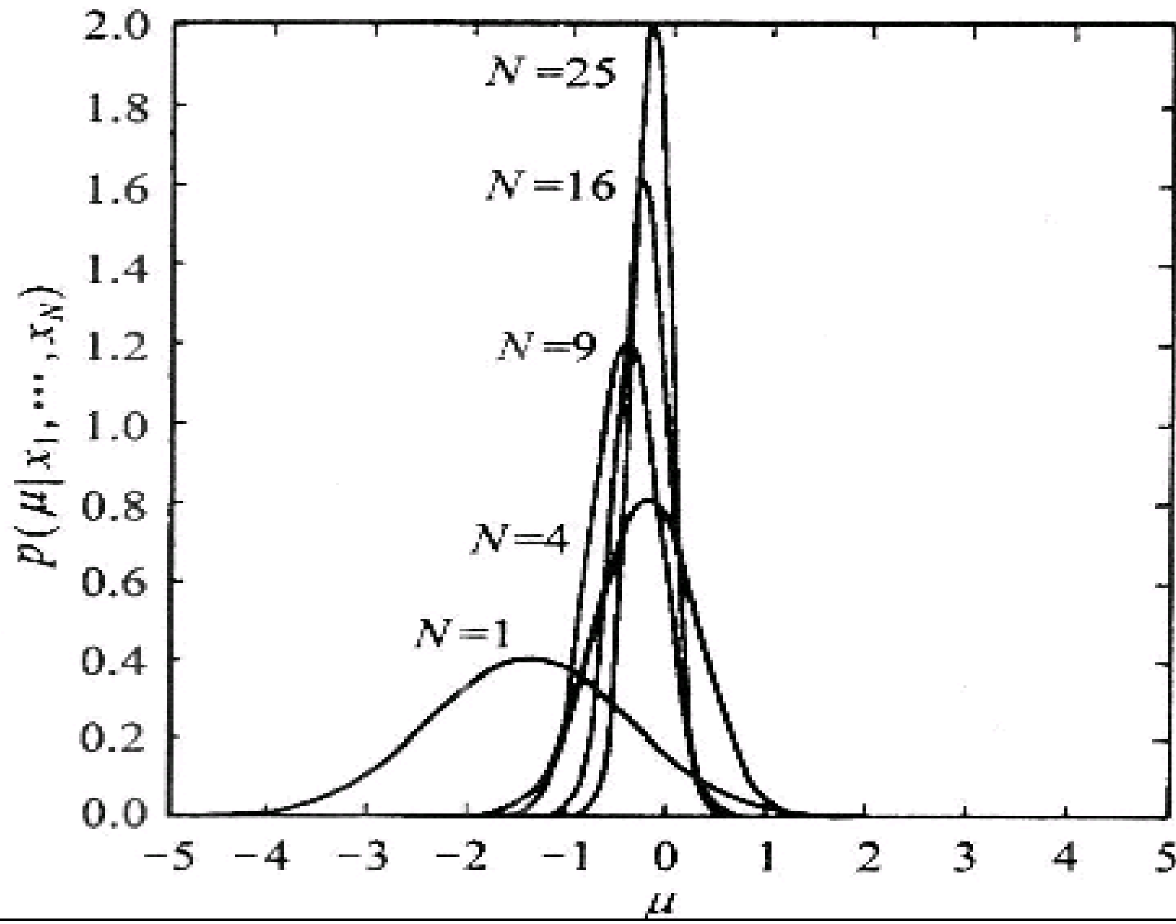
$$\mathcal{X}^N=\{x_1,x_2,\cdots x_N\}$$

$$p(\mathcal{X}^N|\theta)=\boxed{p(x_N|\theta)p(\mathcal{X}^{N-1}|\theta)}$$

$$p(\theta|\mathcal{X}^N)=\frac{p(x_N|\theta)p(\mathcal{X}^{N-1}|\theta)}{\int p(x_N|\theta)p(\mathcal{X}^{N-1}|\theta)d\theta}$$

3.2.2 贝叶斯估计和贝叶斯学习

- 正态分布的贝叶斯学习过程



3.2.2 贝叶斯估计和贝叶斯学习

- 正态分布的贝叶斯估计
  - **单变量，μ未知，σ^2已知的情况**
  - 总体分布密度为 $p(x|\mu) \sim N(\mu, \sigma^2)$
  - μ的先验分布为 $p(\mu) \sim N(\mu_0, \sigma_0^2)$
  - 后验分布 $$p(\theta|\mathrm{X}) = \frac{p(\mathrm{X}|\theta)p(\theta)}{\int_\Theta p(\mathrm{X}|\theta)p(\theta)d\theta}$$
  - 贝叶斯估计方法求μ的估计量

$$\boxed{\hat{\mu}_{\mathrm{BE}} = \int_\Theta \mu p(\mu|\mathrm{X})d\mu}$$

3.2.2 贝叶斯估计和贝叶斯学习

- 后验概率

$$p(\mu \mid X) = \frac{p(X \mid \mu) p(\mu)}{p(X)}$$

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

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \qquad \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2} \qquad m_N = \frac{1}{N} \sum_{k=1}^{N} x_k$$

- 计算$\boldsymbol{\mu}$的贝斯估计

$$\boxed{\hat{\mu} = \int \mu p(\mu \mid X) d\mu = \mu_N}$$

3.2.2 贝叶斯估计和贝叶斯学习

- 正态分布的贝叶斯学习

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

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

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

$N=0$时，$\hat{\mu}=\mu_0$； 若$\sigma_0^2=0$，则$\hat{\mu}\equiv\mu_0$

$N\to\infty$时，$\hat{\mu}\to m_N$ 若$\sigma_0>>\sigma$，则$\hat{\mu}=m_N$

3.3 总体分布的非参数估计

3.3.1 非参数估计的基本原理

- 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
- 常见的一些函数形式很难拟合实际的概率密度，经典的密度函数都是单峰的，而在许多实际情况中却是多峰的
- 两种常用的非参数估计方法：
  - Parzen窗法
  - k_N-近邻法

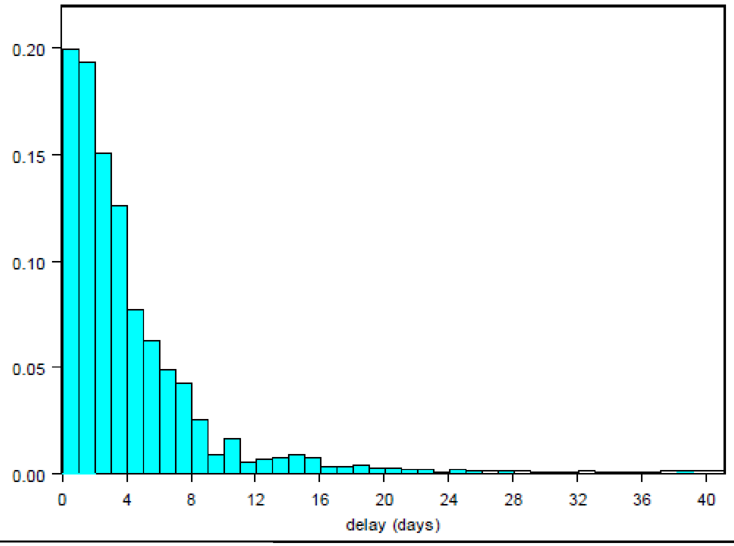
3.3.1 非参数估计的基本原理

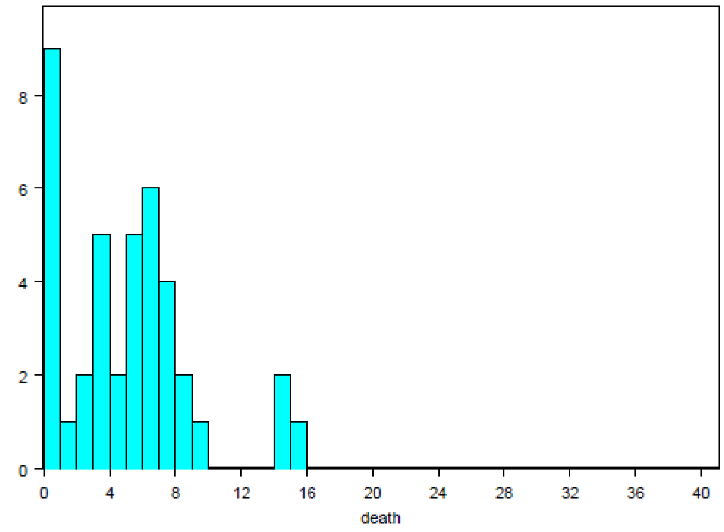
- **直方图方法**
  - 非参数概率密度估计的最简单方法

1）把x的每个分量分成k 个等间隔小窗

2）统计落入各个小舱内的样本数q_i

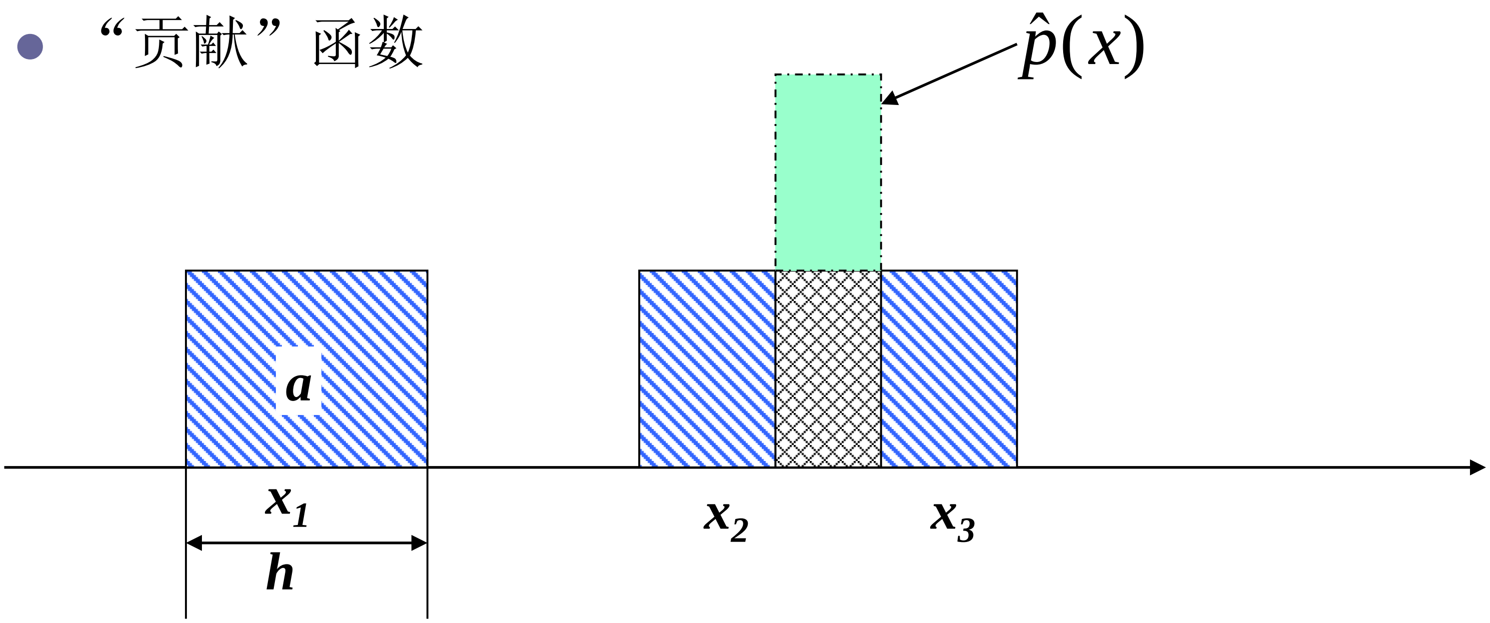
3）相应小舱的概率密度估计值为$q_i /(NV)$





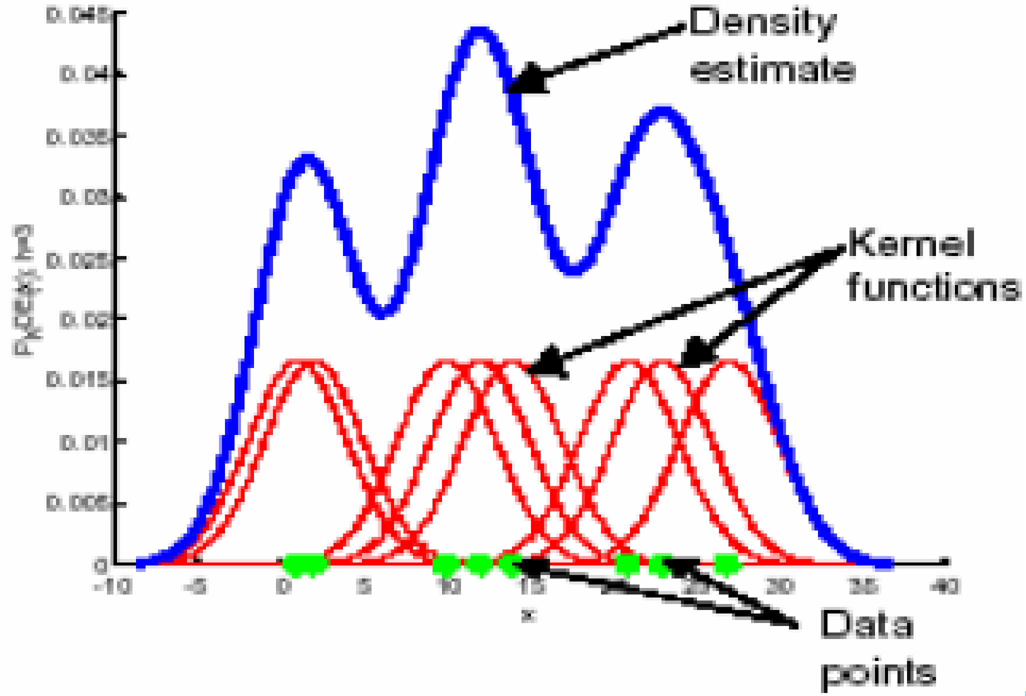
3.3.1 非参数估计的基本原理

- 估计的目的：从样本集$K=\{x_1, x_2,\ldots, x_N\}$估计样本空间中任何一点的概率密度$p(x)$
- 基本思路的例子
  - 样本对概率密度的“贡献”
  - “贡献”函数



3.3.1 非参数估计的基本原理

- 基本方法：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视作对某点概率密度$p(x)$的估计
- 一个样本对自己所在位置的分布贡献最大，离得越远贡献越小



3.3.1 非参数估计的基本原理

- 样本集X={$x_1, x_2, \ldots, x_N$}中的样本，均从概率密度为$p(x)$的总体中独立抽取

- 从k个样本落入区域R的概率P→小区域的概率密度(当区域R的体积足够小，区域内$p(x)$变化很小）

$$\hat{p}(\mathbf{x}) = \frac{P}{V} = \frac{k/N}{V}$$

3.3.1 非参数估计的基本原理

- 构造包含x的区域序列$\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N$，采用N个样本的估计为

$$\hat{p}_N(\mathbf{x}) = \frac{P_N^{(k)}}{V_N} = \frac{k_N / N}{V_N}$$

- $\hat{p}_N(\mathbf{x})$收敛于$p(\mathbf{x})$的条件:

$$(1)\ \lim_{N\to\infty} V_N = 0 \quad (2)\ \lim_{N\to\infty} k_N = \infty \quad (3)\ \lim_{N\to\infty} \frac{k_N}{N} = 0$$

- 区域序列的两种选择方法
  - Parzen窗法

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

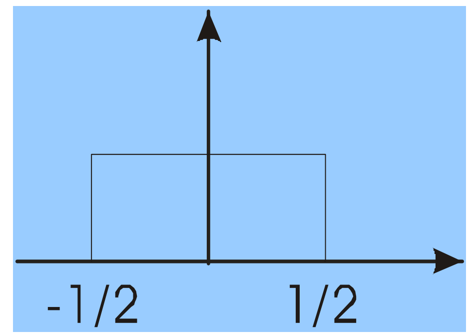
  - k_N-近邻法

$$k_N = k_1\sqrt{N}$$

3.3.2 Parzen窗法

- 区域R_N是一个d维超立方体，棱长h_N，体积$V_N = h_N^d$
- 定义窗(核)函数

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



计数函数

- 落入超立方体内样本数k_N

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_N}\right)$$

- Parzen窗法估计的基本公式

$$\boxed{\hat{p}_N(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{V_N} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_N}\right)}$$

3.3.2 Parzen窗法

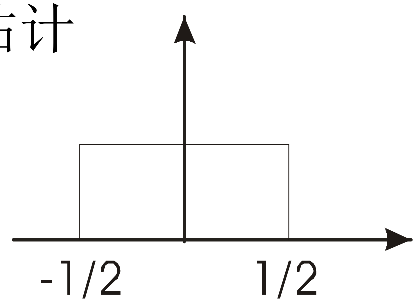
- 窗函数需满足归一化条件（密度函数条件）

$$\varphi(u) \geq 0$$
$$\int \varphi(u)du = 1$$

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

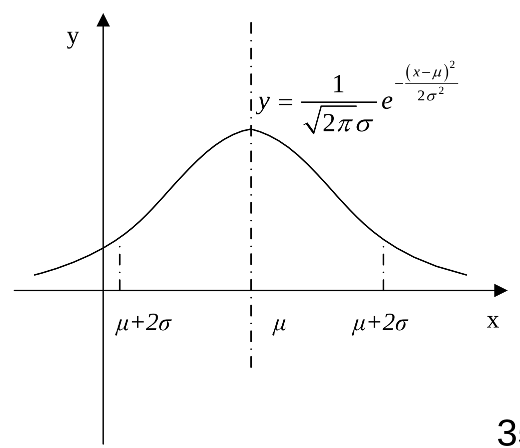
可作为密度函数的估计

- 常用的窗函数
  - 方窗



  - 正态窗

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



  - 指数窗

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

3.3.2 Parzen窗法

- h_N是控制“窗”宽度的参数，根据样本的数量选择
  - 太大：平均化，分辨力低
  - 太小：统计变动大
- 定义窗函数（核函数）的方式

$$\delta_N(\mathbf{x}) = \frac{1}{V_N}\varphi(\frac{\mathbf{x}}{h_N})$$

$$\boxed{\delta_N(\mathbf{x},\mathbf{x}_i) = \frac{1}{V_N}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_N})}$$

$$\hat{p}_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta_N(\mathbf{x}-\mathbf{x}_i)$$

不同窗宽的估计效果有一定不同

3.3.2 Parzen窗法

- 估计密度函数是渐进无偏和平方误差一致的。其充要条件

$p(\mathbf{x})$在$\mathbf{x}$点连续

窗函数满足密度函数的条件

$$\lim_{N\to\infty} h_N = 0$$

$$\lim_{N\to\infty} N{h_N}^d = \infty$$

$$\Longrightarrow \quad h_N = \frac{h_1}{\sqrt[d+1]{N}}$$

3.3.2 Parzen窗法

- 例1：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 $X=(x_1, x_2, \ldots. x_6)=(3.2, 3.6, 3, 6, 2.5, 1.1)$
  - 估计$p(x|\omega_1)$即$p_N(x)$

解：选正态窗函数

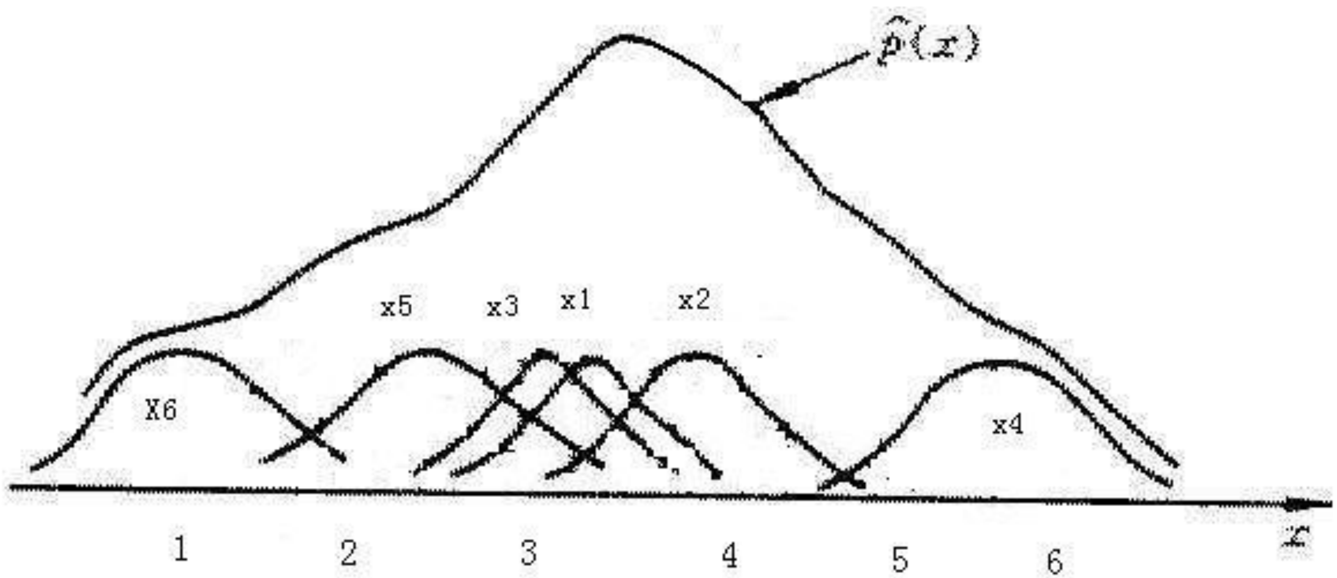
$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2) \qquad \therefore \varphi(u)=\varphi(\frac{|x-x_i|}{h_N})=\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{|x-x_i|}{h_N})^2]$$

$$\therefore \mathrm{V_N}=\mathrm{h_N}=\frac{\mathrm{h_1}}{\sqrt{\mathrm{N}}},\text{其中选}\mathrm{h_1}=0.5\sqrt{6},\ \ N=6 \qquad \therefore V_N=\frac{0.5\sqrt{6}}{\sqrt{6}}=0.5$$

$$\therefore P_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi(\frac{|x-x_i|}{h_N})=0.134\exp[-\frac{1}{2}(\frac{|x-3.2|}{05})^2]+\ldots 0.134\exp[-\frac{1}{2}(\frac{|x-1.1|}{05})^2]$$

3.3.2 Parzen窗法

用图形表示是6个分别以3.2，3.6，3，6，2.5，1.1为中心的丘形曲线(正态曲线)，而$p_N(x)$则是这些曲线的线性组合。



3.3.2 Parzen窗法

例2：设待估计的是个均值为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,试用Parzen窗法估计$p_N(x)$

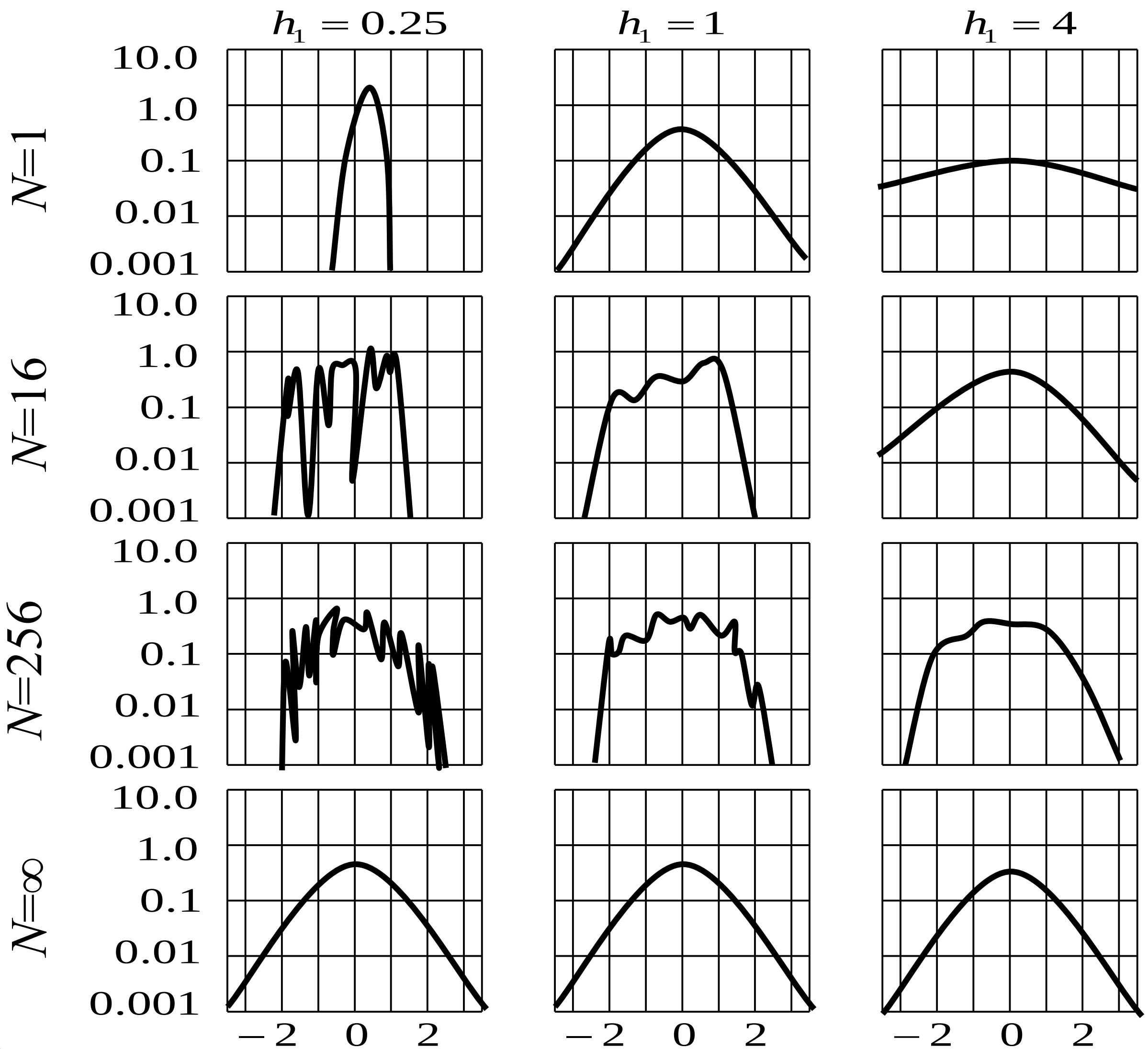
$$设 h_N = h_1 / \sqrt{N}$$

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

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

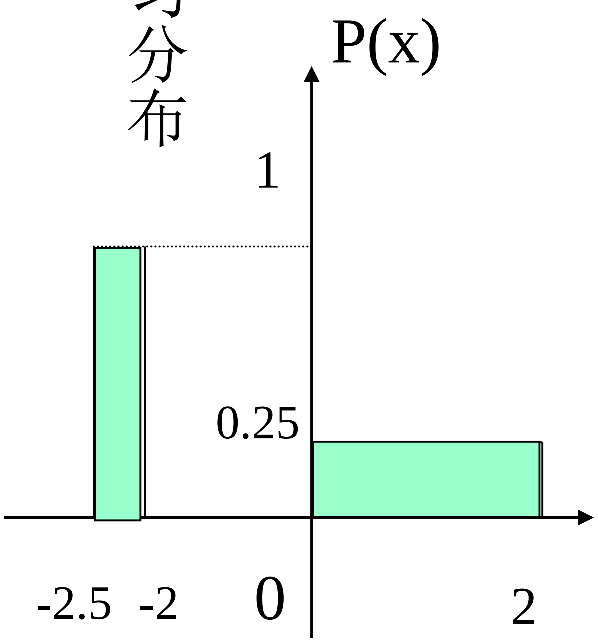
3.3.2 Parzen窗法

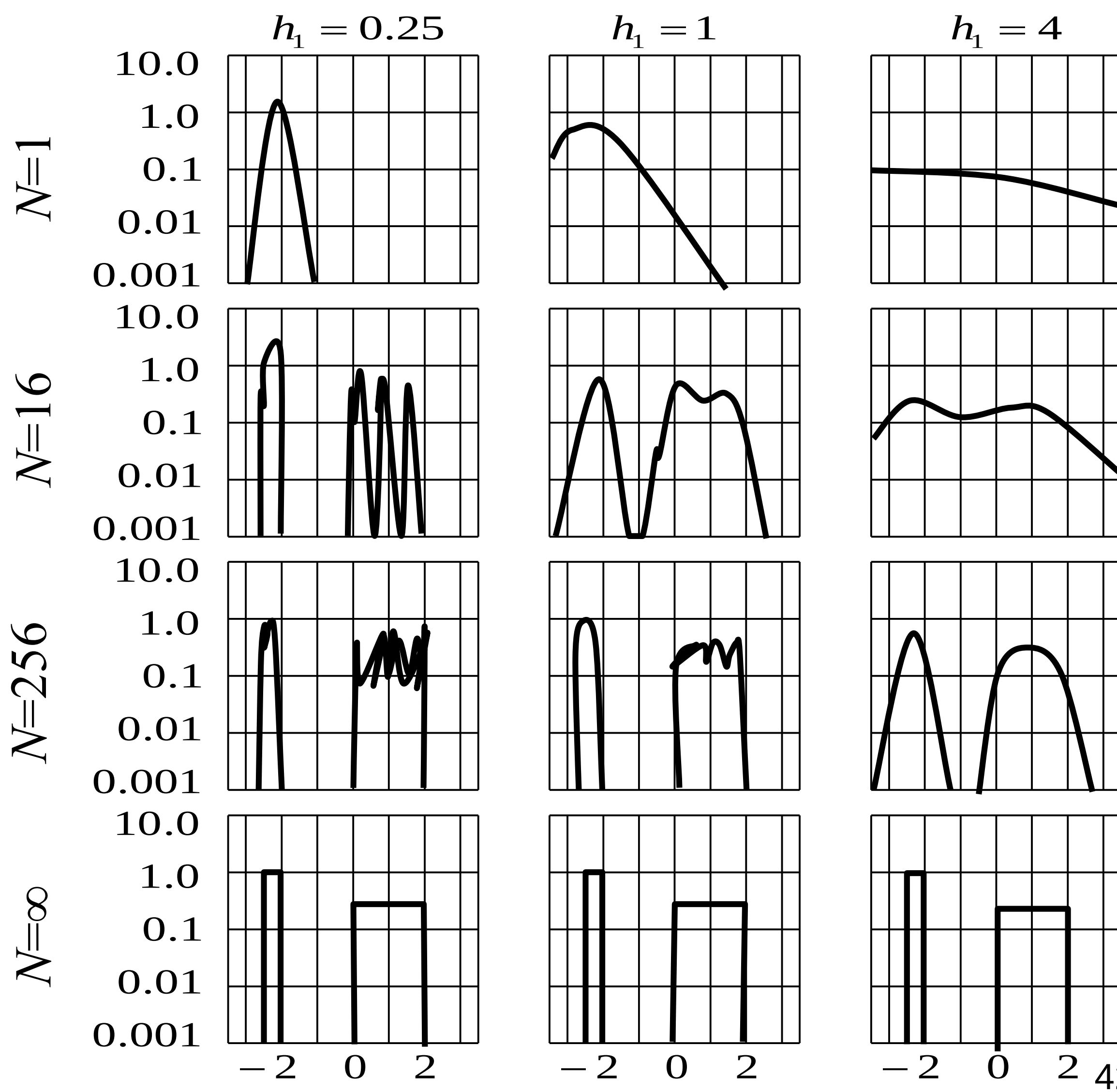
估计单一正态分布



3.3.2 Parzen窗法

估计两个均匀分布





3.3.2 Parzen窗法

- 由上例知窗口法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
- 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

3.3.2 Parzen窗法

- 有限样本的影响
  - 密度估计的均方误差：

$$E(\hat{p}_N(x) - p(x))^2 \cong O(N^{-\frac{4}{d+4}})$$

- 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

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

3.3.3 k_N-近邻估计

- **Parzen**窗估计：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
- 概率密度估计表达式

$$\hat{p}_N(\mathbf{x}) = \frac{1}{V_N}\frac{k_N}{N}$$

- 收敛条件

$$\lim_{N\to\infty} k_N = \infty$$

$$\lim_{N\to\infty} k_N / N = 0$$

经验值

$$k_N = k_1\sqrt{N}$$

3.3.3 k_N-近邻估计

- k_N-近邻估计概率密度函数，所需样本多、计算量大，因此有时并不利用它来进行密度函数估计，而是直接用它来进行样本分类——最近邻规则
  - 将x归入到k_N个最近邻中某一类样本最多的那个类别中
- 其它非参数估计方法
  - 有限项正交函数级数逼近法

- 有限样本下，概率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——→直接设计分类器

3.4 高斯混合模型（GMM）

- GMM的表达
- GMM参数计算
- GMM应用举例

3.4.1 高斯模型

- 是一种参数化模型，用高斯密度函数估计目标的分布

  - 单高斯（Single Gaussian Model）

$$N(x;\mu,\sum)=\frac{1}{(2\pi)^{d/2}\sqrt{|\sum|}}\exp\left\{-\frac{1}{2}(x-\mu)^T\sum^{-1}\left(x-\mu\right)\right\}\quad x\in R^d$$

  - 高斯混合模型（Gaussian mixture model）

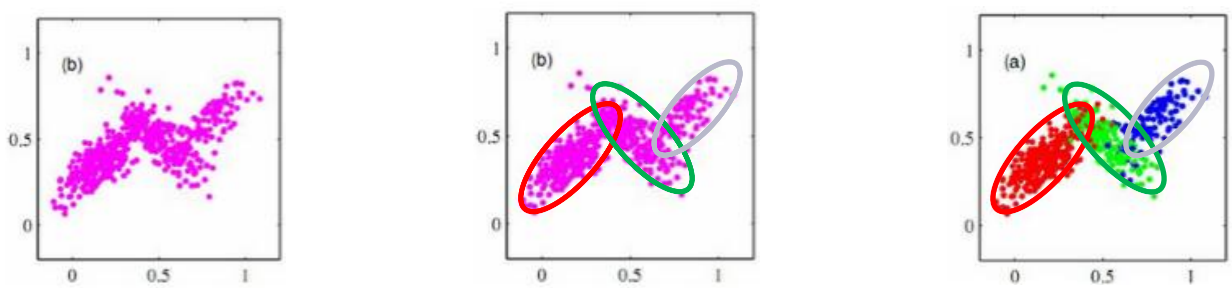
$$\Phi(x;\Theta)=\sum_{i=1}^{k}\pi_i N(x;\mu_{i,}\Sigma_i)\qquad \Theta=\{\mu_1,\Sigma_1,\ldots,\mu_i,\Sigma_i,\ldots,\mu_k,\Sigma_k\}$$

3.4.2 高斯混合模型(GMM)

- 公式表达： $\Phi(x;\Theta)=\sum_{i=1}^{k}\pi_i N(x;\mu_{i,}\Sigma_i)$

由k个加权(π_i)的高斯函数($N(\cdot)$)的线性组合构成，
其中：

参数空间： $\Theta=\{\pi_1,\mu_1,\Sigma_1,\ldots,\pi_i,\mu_i,\Sigma_i,\ldots,\pi_k,\mu_k,\Sigma_k\}$ $\quad |\Theta|=3*k$



高斯混合模型（GMM）

- GMM的表达式
- GMM参数计算
- GMM应用举例

3.4.3 GMM参数学习

- 给定一些观察数据**X**={x}, 假设{x}符合如下的混合高斯分布

$$p(x;\Theta)=\sum_{i=1}^{K}\pi_i N(x;\mu_i,\Sigma_i)$$

求解一组混合高斯模型的参数Θ，使得*：

$$\underset{\Theta}{\operatorname{argmax}} P(X;\Theta)=\underset{\Theta}{\operatorname{argmax}}\prod_i^N p(x_i;\Theta),$$

$$s.t.\ \sum_{k}^{K}\pi_k=1,\ 0\le\pi_k\le1$$

*注意：对于观察集{x}中的各个观察值x_i，这里认为相互之间独立。

3.4.3 GMM参数学习

- 对目标函数取对数：

$$\ln(p(X;\Theta)) = \sum_{i=1}^{N} \ln p(x_i;\Theta) = \sum_{i=1}^{N} \ln(\sum_{k=1}^{K} \pi_k N(x_i;\mu_k,\Sigma_k))$$

- 可以看出目标函数是和的对数，优化问题麻烦，简化的问题：某混合高斯分布一共有K个分布，对于每一个观察到的x，如果我们同时还知道它是属于K中哪一个分布的，则可以直接求解出各个高斯分布的参数。
- 因此引入隐变量Z,用于表示样本{x}输入哪一个高斯分布

3.4.3 GMM参数学习

- **定义$Z_i=\{z_{i1},\ldots, z_{iK}\}$, z_{ik}表示x_i是否属于第k个高斯函数，z_{ik}只有两个取值0、1，即$z_{ik}=1$表示x_i属于第k个高斯函数， $z_{ik}=0$表示x_i不属于第k个高斯函数。**

- 那么，有：

$$\sum_{k=1}^{K} z_{ik} = 1$$

$$p(Z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$$

3.4.3 GMM参数学习

- 引入Z后

$$p(x_i \mid z_{ik}=1;\Theta) = N(x_i \mid \mu_k, \Sigma_k)$$

- 从而得到

$$p(x_i \mid Z_i;\Theta) = \prod_{k=1}^{K} N(x_i \mid \mu_k, \Sigma_k)^{z_{ik}}$$

注意：这里z_{ik}只有0和1的选择

3.4.3 GMM参数学习

- **在简化问题中，我们实际的观察变量是{X,Z},根据一下两个公式**

$$p(Z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}} \qquad p(x_i \mid Z_i;\Theta) = \prod_{k=1}^{K} N(x_i \mid \mu_k, \Sigma_k)^{z_{ik}}$$

- 可以得到

$$p(X, Z;\Theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} N(x_i \mid \mu_k, \Sigma_k)^{z_{ik}}$$

注意：这里N是X={x}集合的大小。

3.4.3 GMM参数学习

- 比较原问题和简化问题

$$\ln(p(X;\Theta)) = \sum_{i=1}^{N} \ln p(x_i;\Theta) = \sum_{i=1}^{N} \ln(\sum_{k=1}^{K} \pi_k N(x_i;\mu_k,\Sigma_k))$$

$$\ln p(\mathrm{X},\mathrm{Z};\Theta) = \ln\{\prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{ik}} N(x_i \mid \mu_k,\Sigma_k)^{z_{ik}}\}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{ik}\{\ln \pi_k + \ln N(x_i \mid \mu_k,\Sigma_k)\}$$

- 后者的ln直接作用于正态分布，使正态分布由乘的e指数形式变为加的简单形式

3.4.3 GMM参数学习

$$\ln p(\mathrm{X},\mathrm{Z};\Theta)=\sum_{i=1}^{N}\sum_{k=1}^{K}z_{ik}\{\ln\pi_k+\ln N(x_i\mid\mu_k,\Sigma_k)\}$$

- **为了最大化上式，由于z_{ik}已知，我们可以把上式按观察到的(x,z) 分为K组，即按照所属的高斯函数进行分组**

$$\ln p(\mathrm{X},\mathrm{Z};\Theta)=\sum_{i\in C_1}(\ln\pi_1+\ln N(x_i\mid\mu_1,\Sigma_1))+\sum_{i\in C_2}(\ln\pi_2+\ln N(x_i\mid\mu_2,\Sigma_2))$$
$$+\ldots+\sum_{i\in C_K}(\ln\pi_K+\ln N(x_i\mid\mu_K,\Sigma_K))$$

注意：用到z_{ik}的取值，所以z_{ik}不会再出现在公式中。

3.4.3 GMM参数学习

- 因为我们假定z_{ik}已知，因而最大化某一个高斯函数

$$\sum_{i \in C_k} (\ln \pi_k + \ln N(x_i \mid \mu_k, \Sigma_k))$$

是可以数值求解的。

- 假定C_k中含有N_k个样本，则

$$\sum_{i \in C_k} (\ln \pi_k + \ln N(x_i \mid \mu_k, \Sigma_k)) = N_k \ln \pi_k + \ln N(X \mid \mu_k, \Sigma_k)$$

注意：这里X表示的是x_i的联合，X 表示的是仅属于C_k的样本的联合。

3.4.3 GMM参数学习

- 对于单高斯函数

$$\ln N(X \mid \mu_k, \Sigma_k) = -\frac{N_k D}{2}\ln(2\pi) - \frac{N_k}{2}|\Sigma| - \frac{1}{2}\sum_{i=1}^{N_k}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)$$

- 对上式μ_k求偏导

$$\frac{\partial}{\partial \mu_k}\ln N(X \mid \mu_k, \Sigma_k) = \sum_{i=1}^{N_k}\Sigma_k^{-1}(x_i - \mu_k)$$

- 令上式等于0，则有 $\mu_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_i$

- 同理，可以得到 $\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N_k}(x_i - \mu_k)(x_i - \mu_k)^T$

3.4.3 GMM参数学习

- 在z_{ik}已知的情况下，我们求出了高斯函数的数值解：

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} z_{ik} x_i$$

$$\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N} z_{ik}(x_i - \mu_k)(x_i - \mu_k)^T$$

其中

$$N_k = \sum_{i=1}^{N} z_{ik}$$

3.4.3 GMM参数学习

- 现在，假定已知所有μ_k和Σ_k，来求解π_k，因为π_k满足$\sum_{k=1}^{K}\pi_k = 1$，可以对lnp(X, Z|Θ)应用拉格朗日乘法：

$$L = \ln p(\mathrm{X},\mathrm{Z};\Theta) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\{\ln \pi_k + \ln N(x_i \mid \mu_k, \Sigma_k)\} + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

- 对L中π_k和λ求导并令其为0，得到

$$\left.\begin{aligned} \frac{\partial L}{\partial \lambda} &= \sum_{k=1}^{K}\pi_k - 1 = 0 \\ \frac{\partial L}{\partial \lambda} &= \frac{\sum_{i=1}^{N} z_{ik}}{\pi_k} + \lambda = 0 \end{aligned}\right\} \quad \begin{aligned} \lambda &= -\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} = -N \\ \pi_k &= \frac{\sum_{i=1}^{N} z_{ik}}{N} = \frac{N_k}{N} \end{aligned}$$

3.4.3 GMM参数学习

- 引入隐变量Z后，我们得到了最大化目标函数的结果

$$\ln p(\mathrm{X},\mathrm{Z};\Theta)=\sum_{i=1}^{N}\sum_{k=1}^{K}z_{ik}\{\ln\pi_k+\ln N(x_i\mid\mu_k,\Sigma_k)\}$$

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}z_{ik}x_i$$

$$\Sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}z_{ik}(x_i-\mu_k)(x_i-\mu_k)^T$$

$$\pi_k=\frac{\sum_{i=1}^{N}z_{ik}}{N}=\frac{N_k}{N}$$

$$N_k=\sum_{i=1}^{N}z_{ik}$$

⇒

通过上面的公式，可以看出如果能够知道每一个z_{ik}的取值，那么就能求解出最大化目标函数的参数取值，但事实上给定一组观察数据{x}后，是无法获取z_{ik}的，因此我们将用z_{ik}的均值$E\{z_{ik}\}$来代替。

⇒

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}\boxed{E(z_{ik})}x_i$$

$$\Sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}\boxed{E(z_{ik})}(x_i-\mu_k)(x_i-\mu_k)^T$$

$$\pi_k=\frac{\sum_{i=1}^{N}\boxed{E(z_{ik})}}{N}=\frac{N_k}{N}$$

$$N_k=\sum_{i=1}^{N}\boxed{E(z_{ik})}$$

3.4.3 GMM参数学习

进一步解释

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} E(z_{ik}) x_i$$

$$\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N} E(z_{ik})(x_i-\mu_k)(x_i-\mu_k)^T$$

$$\pi_k = \frac{\sum_{i=1}^{N} \mathrm{E}(z_{ik})}{N} = \frac{N_k}{N}$$

$$N_k = \sum_{i=1}^{N} E(z_{ik})$$

现在的问题转化为如何求解$E\{z_{ik}\}$？

为了求解$E\{z_{ik}\}$，我们假定已知一组参数取值（Θ的取值），从而求解z_{ik}的后验数学期望$E\{z_{ik}|\Theta\}$,得到$E\{z_{ik}\}$我们再计算Θ。

显然求得的Θ是给定$E\{z_{ik}\}$情况下的结果，并不一定是目标函数的解，为此：

引入一种迭代思想，迭代计算$E\{z_{ik}\}$和Θ

$$E\{z_{ik}\} \rightleftarrows \Theta \Rightarrow$$ EM算法

3.4.3 GMM参数学习

$$E(z_{ik}\mid\Theta)=\sum_{n=1}^{N}\sum_{z_{ik}} z_{ik}\,p(z_{ik}\mid x_n,\Theta)$$

$$=\sum_{z_{ik}} z_{ik}\,p(z_{ik}\mid x_i,\Theta)$$

z_{ik}只与x_i有关

$$=\sum_{z_{ik}} z_{ik}\frac{p(z_{ik}\mid\Theta)\,p(x_i\mid z_{ik},\Theta)}{p(x_\mathrm{i}\mid\Theta)}$$

贝叶斯公式

$$=1\bullet\frac{p(z_{ik}=1\mid\Theta)\,p(x_i\mid z_{ik}=1,\Theta)}{p(x_\mathrm{i}\mid\Theta)}+0\bullet\frac{p(z_{ik}=0\mid\Theta)\,p(x_i\mid z_{ik}=0,\Theta)}{p(x_\mathrm{i}\mid\Theta)}$$

Z_{ik}只有0、1取值

$$=\frac{p(z_{ik}=1\mid\Theta)\,p(x_i\mid z_{ik}=1,\Theta)}{p(x_\mathrm{i}\mid\Theta)}$$

$$=\frac{\pi_\mathrm{k}\mathrm{N}(x_i\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_\mathrm{j}\mathrm{N}(x_i\mid\mu_j,\Sigma_j)}$$

公式展开

3.4.3 GMM参数学习

- 小结

E-step:给定Θ计算E{z_{ik}| Θ}

$$E(z_{ik} \mid \Theta) = = \frac{\pi_k \mathrm{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathrm{N}(x_i \mid \mu_j, \Sigma_j)}$$

E-step

M-step

M-step:给定E{z_{ik}| Θ}更新Θ

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} E(z_{ik}) x_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} E(z_{ik})(x_i - \mu_k)(x_i - \mu_k)^T$$

$$\pi_k = \frac{\sum_{i=1}^{N} \mathrm{E}(z_{ik})}{N} = \frac{N_k}{N}$$

$$N_k = \sum_{i=1}^{N} E(z_{ik})$$

3.4.3 GMM参数学习

算法流程

1.初始化一组参数Θ，如K-means方法；

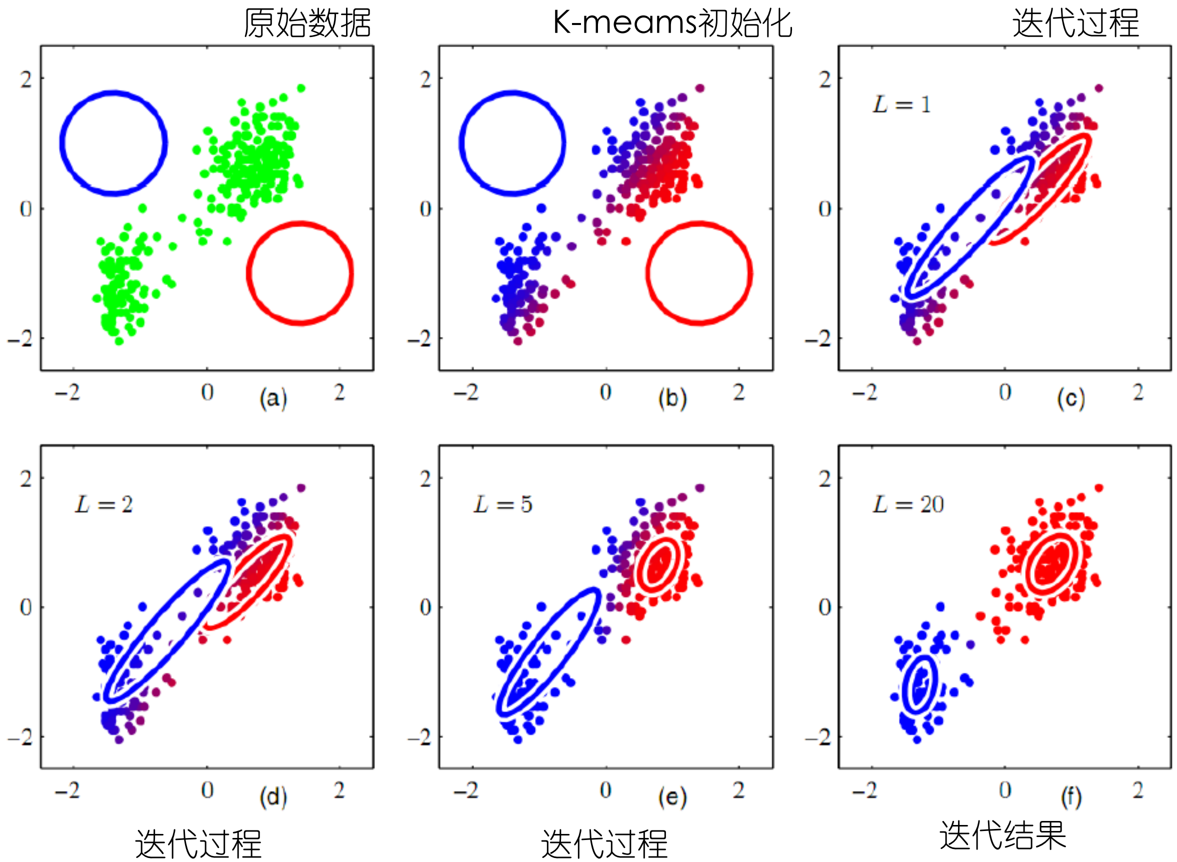
2.根据参数Θ计算$E\{z_{ik}|\,\Theta\}$；

3.根据$E\{z_{ik}|\,\Theta\}$更新参数Θ；

4.迭代2.3.直到达到某种阈值条件，如$|\Theta^{new}-\Theta^{old}|$，$|p(X|\Theta^{new})-p(X|\Theta^{old})|$；

3.4.3 GMM参数学习

- GMM，K=2



3.4.4 GMM应用

- 背景模型：对图像中的场景进行建模，从而进行运动检测。

要点1：将图像中的每个图像单位(像素，块等)看成是从混合高斯分布样本中采样得到的随机变量；

要点2：根据先验知识，每个像素点是前景或背景的先验概率可以估值；

要点3：考虑到背景的多模态和复杂度，一般的混合高斯模型采用3-5个单高斯模型进行混合。

视频帧　背景图像　前景图像

3.4.4 GMM应用

- 用K个高斯模型来表征图像中各个像素点的特征
- $I_{xy,t}$属于背景区域的概率密度：

$$f(I_{xy,t}) = \sum_{k=1}^{K} \pi_{xy,k,t} \bullet N(I_{xy,t}, \mu_{xy,k,t}, \sigma_{xy,k,t}), \qquad \sum_{k=1}^{K} \pi_{xy,k,t} = 1$$

$I_{xy,t}$ 表示t时刻图像I中的xy位置的像素特征。

模型主要步骤：

1，模型初始化；
2，背景描述；
3，前景判决；
4，模型更新。