第三章

概率密度函数的估计

目　录

3.1 引言

- Bayes分类
  - 已知先验概率 $p(\omega_i)$ 与类条件概率 $p(\mathbf{x} \mid \omega_i)$，可以设计一个最优分类器。
- 问题
  - 实际情况中，$p(\mathbf{x} \mid \omega_i)$ 的确切分布很难知道，这就需要根据已有样本作出参数估计。
  - 特定条件下，可以合理地假设 $p(\mathbf{x} \mid \omega_i)$ 是均值为 μ_i，协方差矩阵为 Σ_i 的正态分布，将问题缩小为估计 μ_i Σ_i 的值。

参数估计

- 参数估计是知道概率密度的分布形式，但其中的部分未知或全部未知。概率密度函数估计就是通过样本来估计这些参数。
- 本章介绍：
  - 最大似然估计
  - 贝叶斯估计
  - EM估计方法

非参数估计

- 非参数估计是既不知道分布形式，也不知道分布里的参数，通过样本的分布把概率密度函数值数值化估计出来。

- 本章介绍：
  - Parzen窗法
  - Kn近邻法

3.2 最大似然估计

- 一般原则：条件
  - 设已知样本集有样本类$X_1, X_2, \cdots, X_c$，其中X_j类有样本$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$，是按概率密度$p(\mathbf{x} \mid \omega_j)$从总体中独立地抽取的，但是其中某一参数$\mu$或参数矢量$(\mu, \sigma)$不知道，记作参数$\theta_j$。
  - 假设1：参数θ_j唯一地是由$p(\mathbf{x} \mid \omega_j)$决定

最大似然估计

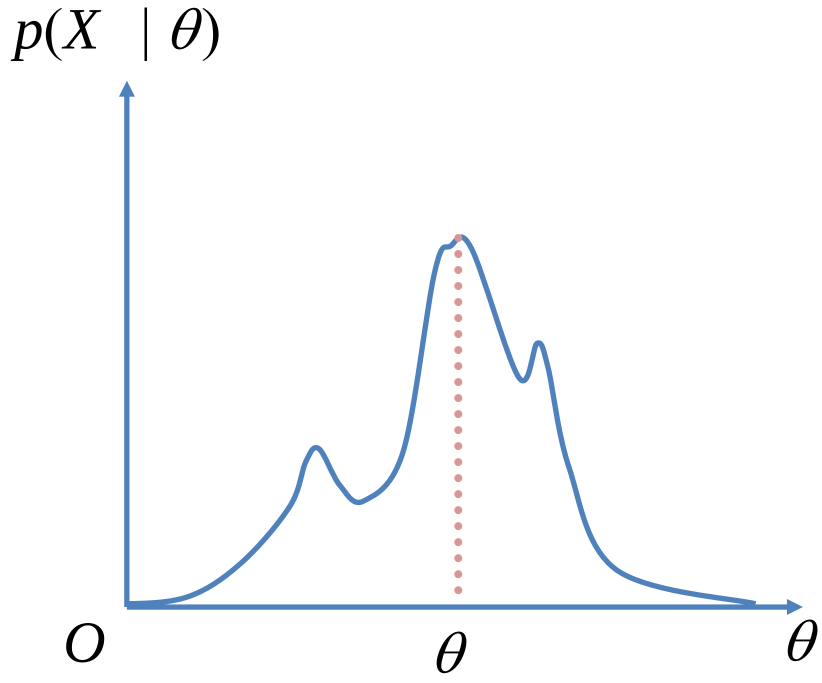
- 似然函数： $p(X \mid \theta)$
  - 同一类的样本子集 $X = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ ，它们具有概率密度 $p(\mathbf{x}_k \mid \theta), k = 1, 2, \cdots n$ ，且样本是独立抽取的，因此

$$p(X \mid \theta) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \theta),$$

$$L(\theta) = p(X \mid \theta)$$

$$\hat{\theta} = \arg\max L(\theta)$$

下页



最大似然估计

- 对数似然函数：$\log p(X \mid \theta)$

$$L(\theta) = \log p(X \mid \theta) = \sum_{k=1}^{n} \log p(\mathbf{x}_k \mid \theta),$$

$$\hat{\theta} = \arg\max L(\theta)$$

- 计算：

$$\nabla_\theta L = \frac{\partial}{\partial \theta}\left(\log p(X \mid \theta)\right) = \sum_{k=1}^{n} \frac{\partial}{\partial \theta}\left[\log p(\mathbf{x}_k \mid \theta)\right] = 0$$

$$\nabla_\theta = \left\{ \begin{array}{c} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{array} \right\}$$

最大似然估计

- 问题：$\nabla_\theta L = 0$ 并不一定能够得到解。
- 举例：**x**服从均匀分布，参数 θ_1, θ_2 未知

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

假设从总体中独立地抽取N个样本，则

$$L(\theta) = p(X \mid \theta) = \begin{cases} \dfrac{1}{\left(\theta_2 - \theta_1\right)^N} \\ 0 \end{cases}$$

最大似然估计

- 对数似然函数

$$L(\theta) = \log p(X \mid \theta) = -N \ln\left(\theta_2 - \theta_1\right)$$

$$\frac{\partial L(\theta)}{\partial \theta_1} = N \cdot \frac{1}{\theta_2 - \theta_1}$$

$$\frac{\partial L(\theta)}{\partial \theta_2} = -N \cdot \frac{1}{\theta_2 - \theta_1}$$

?

正态分布下的最大似然估计

- 均值未知的d维正态情况
  - 设X 中的某一样本$\mathbf{x}_k = (x_{k1}, x_{k2}, \cdots, x_{kd})^T$ 具有正态形式，参数μ未知，

$$p(\mathbf{x}_k \mid \mu) = \frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{x}_k - \mu)^T \Sigma^{-1} (\mathbf{x}_k - \mu) \right]$$

log

$$\log p(\mathbf{x}_k \mid \mu) = -\frac{1}{2} \log\left[(2\pi)^d \mid \Sigma \mid \right] - \frac{1}{2} (\mathbf{x}_k - \mu)^T \Sigma^{-1} (\mathbf{x}_k - \mu)$$

$$\boxed{\begin{aligned} \nabla_\theta \log p(\mathbf{x}_k \mid \mu) &= \frac{\partial}{\partial \mu} (\mathbf{x}_k - \mu)^T \Sigma^{-1} (\mathbf{x}_k - \mu) \\ &= \ldots \end{aligned}}$$

正态分布下的最大似然估计

- 进一步地

$$
\begin{aligned}
\nabla_{\mu} \log p(\mathbf{x}_k \mid \mu) &= \frac{\partial}{\partial \mu}(\mathbf{x}_k - \mu)^T \Sigma^{-1} \mathbf{x}_k - \mu) \\
&= \frac{\partial}{\partial \mu}(\mathbf{x}_k - \mu)^T \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right] + (\mathbf{x}_k - \mu)^T \frac{\partial}{\partial \mu}\left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right] \\
&= [-1]^T \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right] + \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right]^T [-1]^T \\
&= 2[-1]^T \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right]
\end{aligned}
$$

$$\nabla_{\mu} L = 2[-1]^T \left[\Sigma^{-1}(\mathbf{x}_k - \hat{\mu})\right] = 0 \quad \Rightarrow \quad \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\mu}) = 0$$

- 结论

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

正态分布下的最大似然估计

- 均值、方差未知的一维正态情况

$$\theta_1 = \mu, \quad \theta_2 = \sigma^2$$

$$p(\mathbf{x}_k \mid \theta) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2} \frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2} \right]$$

log

$$\log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2} \log 2\pi\theta_2 - \frac{1}{2} \frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}$$

正态分布下的最大似然估计

$$\log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}$$

∇_{θ_1}

$$\nabla_{\theta_1} \log p(\mathbf{x}_k \mid \theta) = -\left[\frac{1}{2\theta_2}\cdot 2(\mathbf{x}_k - \theta_1)\cdot(-1)\right] = \frac{\mathbf{x}_k - \theta_1}{\theta_2}$$

- 均值

$$\sum_{k=1}^{n} \nabla_{\theta_1} L = \frac{1}{\theta_2}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\theta}_1) = 0 \qquad \sum_{k=1}^{n}(\mathbf{x}_k - \hat{\theta}_1) = 0$$

$$\boxed{\hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k}$$

正态分布下的最大似然估计

$$\log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}$$

∇_{θ_2}

$$\nabla_{\theta_2} \log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\left[(\frac{1}{2\theta_2}\cdot 2\pi) - \frac{(\mathbf{x}_k - \theta_1)^2}{2}\cdot(-1)\theta_2^{-2} \right]$$

$$= -\frac{1}{2\theta_2} + \frac{(\mathbf{x}_k - \theta_1)^2}{2\theta_2^2}$$

- 方差：有偏估计

$$\sum_{k=1}^{n} \nabla_{\theta_2} L = \frac{1}{2\theta_2}\left[\sum_{k=1}^{n}(-1) + \sum_{k=1}^{n}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2} \right] = 0$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \theta_1)^2 \approx \sigma^2 = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \mu)^2$$

正态分布下的最大似然估计

- 多变量情况

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

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

3.3 Bayes估计

- Bayes估计与最大似然估计的区别

最大似然估计是把待估计的参数当作未知但固定的量；而贝叶斯估计则把待估计的参数本身看作是随机变量，要做的是根据观测数据对参数的分布进行估计，除了测量数据外，还可以考虑参数的先验分布

  – 目的：把待估参数θ 看成具有先验分布密度$p(\theta)$ 的随机变量，其取值与样本集$X = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$有关，我们要做的是根据$X$ 估计最优的θ^* 。

Bayes估计

- 最优的条件可设定为最小风险
- 损失函数：假定把连续变量 θ 估计成 $\hat{\theta}$ 的损失为 $\lambda(\hat{\theta},\theta)$ 。
- 定义：在样本 $\mathbf{x}$ 下的条件风险为：

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

- 目标：对期望风险求最小，等价于对所有的样本求条件风险最小：

$$\theta^* = \arg\min R(\hat{\theta}\,|\,X\ ) = \int_{\Theta} \lambda(\hat{\theta},\theta)\, p(\theta\,|\,X\ )\mathrm{d}\theta$$

Bayes估计

- 通常情况下我们使用的损失函数为平方误差损失函数 $\lambda(\hat{\theta},\theta)=(\hat{\theta}-\theta)^2$,
- 在平方误差损失函数与样本集X 下，贝叶斯估计量$\hat{\theta}$为θ在X下的条件期望，即：

$$\theta^* = E[\theta \mid X\ ] = \int_{\Theta} \theta p(\theta \mid X\ ) d\theta$$

Bayes估计

- 贝叶斯估计的步骤是：

1. 确定 θ 的先验分布 $p(\theta)$
2. 求出样本集的联合分布为 $p(X\ |\theta)=\prod_{i=1}^{N}p(\mathbf{x}_i\,|\,\theta)$
3. 利用贝叶斯公式，求 θ 的后验概率分布：

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

4. θ的贝叶斯估计量是：$\theta^*=\int_{\Theta}\theta p(\theta|X\ )d\theta$

正态分布下的Bayes估计

设 ω_j 类：$p(\mathbf{x} \mid \mu) \sim N(\mu, \sigma^2)$，$\mu$ 为未知随机参数

条件1：$X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ 为已知类别为 ω_j 的n个同类样本，并且是独立抽取的。

条件2：考虑 $\mathbf{x}$ 是一维的情况。

条件3：把 μ 均看作是随机变量，遵循如下分布

$$p(\mathbf{x}_k \mid \mu) \sim N(\mu, \sigma^2)$$

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

正态分布下的Bayes估计

- 推导过程

条件1

条件3

$$p(\mu\mid X)=\frac{p(X\mid\mu)p(\mu)}{\int p(X\mid\mu)p(\mu)d\mu}=\alpha p(X\mid\mu)p(\mu)=\alpha\left[\prod_{k=1}^{n}p(\mathbf{x}_k\mid\mu)\right]p(\mu)$$

$$p(\mu\mid X)=\alpha\left\{\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\mathbf{x}_k-\mu)^2}{2\sigma^2}\right]\right\}\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

$$=\alpha'\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{n}\left(\frac{\mathbf{x}_k-\mu}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]\right\}$$

$$=\alpha'\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{n}\left(\frac{\mathbf{x}_k{}^2}{\sigma^2}-\frac{2\mathbf{x}_k\mu}{\sigma^2}+\frac{\mu^2}{\sigma^2}\right)+\frac{\mu^2}{\sigma_0{}^2}-\frac{2\mu\mu_0}{\sigma_0{}^2}+\frac{\mu_0{}^2}{\sigma_0{}^2}\right]\right\}$$

$$=\alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0{}^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}\mathbf{x}_k+\frac{\mu_0}{\sigma_0{}^2}\right)\mu\right]\right\}$$

正态分布下的Bayes估计

- $p(\mu \mid X)$仍是一个正态函数，称为再生密度。

$$p(\mu \mid X) = \alpha'' \exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^{\ 2}}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}\mathbf{x}_k + \frac{\mu_0}{\sigma_0^{\ 2}}\right)\mu\right]\right\}$$

假设 $p(\mu \mid X) \sim N(\mu_n, \sigma_n^{\ 2})$ ，即

$$p(\mu \mid X) = \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma_n^{\ 2}}\right]$$

$$= \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^{\ 2}}\mu^2 - \frac{2\mu_n}{\sigma_n^{\ 2}}\mu + \frac{\mu_n^{\ 2}}{\sigma_n^{\ 2}}\right)\right]$$

比较

正态分布下的Bayes估计

- μ_n, σ_n的求解

$$\begin{cases} \dfrac{1}{\sigma_n^{\ 2}} = \dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^{\ 2}} \\ \dfrac{\mu_n}{\sigma_n^{\ 2}} = \dfrac{1}{\sigma^2}\sum_{k=1}^{n}\mathbf{x}_k + \dfrac{\mu_0}{\sigma_0^{\ 2}} = \dfrac{n}{\sigma^2} m_n + \dfrac{\mu_0}{\sigma_0^{\ 2}} \end{cases} \qquad m_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\Downarrow$$

$$\mu_n = \frac{n\sigma_0^{\ 2}}{n\sigma_0^{\ 2} + \sigma^2} m_n + \frac{\sigma^2}{n\sigma_0^{\ 2} + \sigma^2}\mu_0$$

$$\sigma_n^{\ 2} = \frac{\sigma_0^{\ 2}\sigma^2}{n\sigma_0^{\ 2} + \sigma^2}$$

正态分布下的Bayes估计

- 根据 $\theta^* = \int_\Theta \theta p(\theta | X)d\theta$

  计算 μ 的贝叶斯估计

$$\mu^* = \int \mu p(\mu | X)d\mu = \int \mu \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma_n^2}\right]d\mu = \mu_n$$

$$= \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}m_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

正态分布下的Bayes估计

- 分析： $\mu_n = \dfrac{n\sigma_0^2}{n\sigma_0^2+\sigma^2} m_n + \dfrac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \qquad \sigma_n^2 = \dfrac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$

  1. 再生密度的均值是样本均值和先验均值的线性组合。

  2. 一般情况下 $\sigma_0 \neq 0$，则当$n \to \infty, \mu_n \to m_n$。

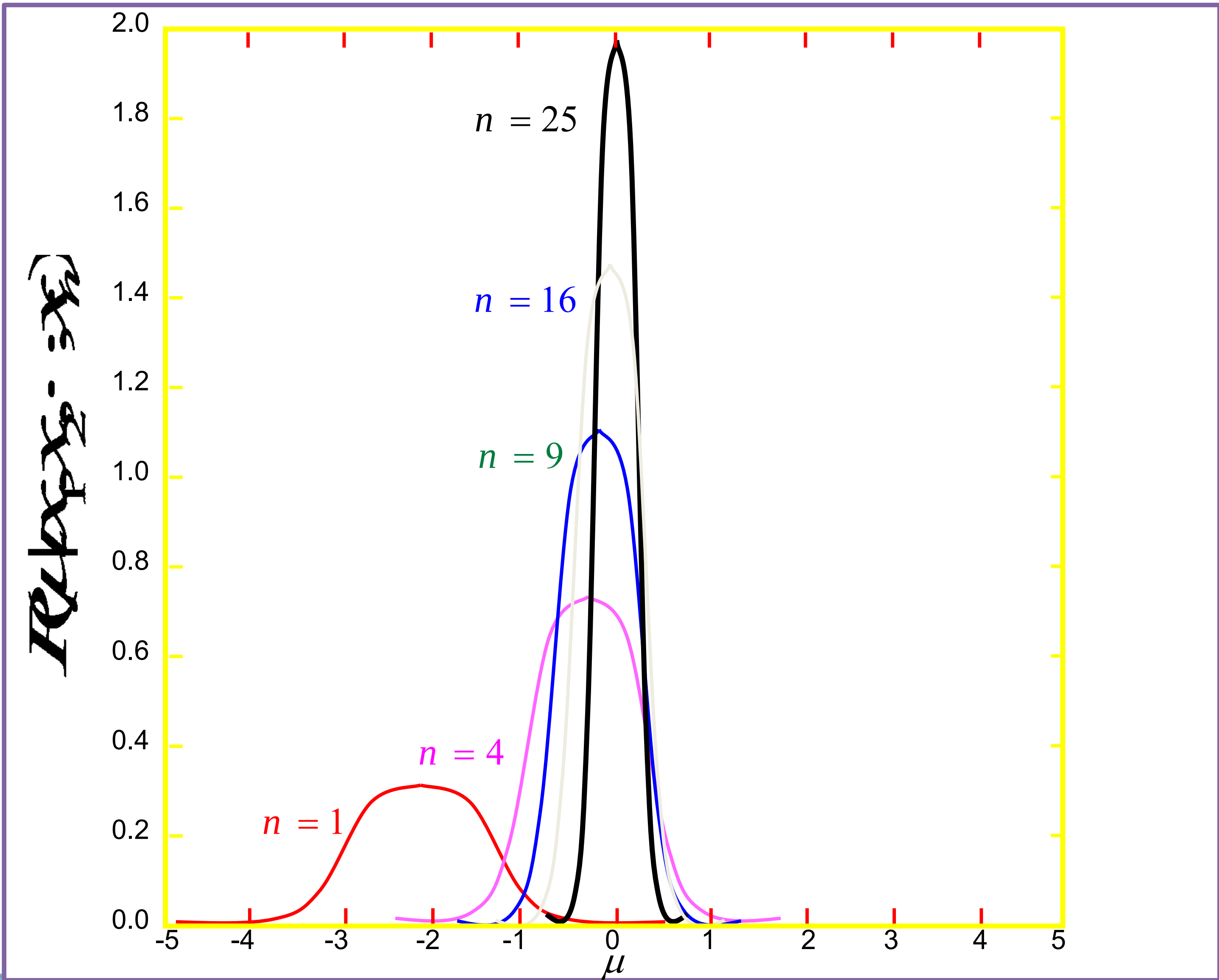
  极端情况1：$\sigma_0 = 0 \Rightarrow \mu_n = \mu_0, \forall n$，说明先验值$\mu_0$十分可靠。

  极端情况2：$\sigma_0 \gg \sigma \Rightarrow \mu_n = m_n$，说明先验值十分没有把握。

  3. σ_n^2随n的增加而减小，说明 σ_n^2趋于$\dfrac{\sigma^2}{n}$。

  – 参见下页图示

正态分布下的Bayes估计



正态分布下的Bayes估计

- 求类条件密度

$$p(\mathbf{x}|\mathcal{X}) = \int p(\mathbf{x}|\theta)p(\theta|\mathcal{X})d\theta \quad + \quad \begin{matrix} p(\mathbf{x}|\mu) \sim N(\mu,\sigma^2) \\ p(\mu|\mathcal{X}) \sim N(\mu_n,\sigma_n^2) \end{matrix}$$

$$p(\mathbf{x}|\mathcal{X}) = \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(\mathbf{x}-\mu)^2}{\sigma^2}\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma_n^2}\right]d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{(\mathbf{x}-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]\cdot f(\sigma,\sigma_n)$$

其中，$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{(\sigma^2+\sigma_n^2)}{\sigma^2}\left(\mu-\frac{\sigma^2 x+\sigma_n^2\mu_n^2}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu$

正态分布下的Bayes估计

- 分析

  1. $$p(\mathbf{x} \mid X) \propto \exp\left[-\frac{(\mathbf{x}-\mu_n)^2}{\sigma^2+\sigma_n{}^2} \right]$$

     $$p(\mathbf{x} \mid X) \sim N(\mu_n, \sigma^2+\sigma_n{}^2)$$

  2. 条件概率 $p(\mathbf{x} \mid X)$ 的均值和后验概率$p(\theta \mid X)$的均值相等。

  3. 条件概率$p(X \mid X)$的方差比后验概率$p(\theta \mid X)$的方差大。 $\sigma^2+\sigma_n{}^2$ $\sigma_n{}^2$

  4. 多维正态分布

     $$p(X \mid X) \sim N(\mu_n, \Sigma+\Sigma_n)$$

3.4 EM估计方法

- EM（Expectation Maximization）算法是一种对概率模型寻找隐藏参数的最大似然解的技术。
- 在概率模型中有两个变量，其一是可以观测到的变量**x**，另一个是隐藏变量***z***。概率模型的联合分布 $p(\mathbf{x},\mathbf{z}\mid\theta)$ 由参数θ控制。我们的目标是最大化下面函数的似然函数：

$$p(\mathbf{x}\mid\theta)=\sum_{\mathbf{z}}p(\mathbf{x},\mathbf{z}\mid\theta)$$

EM 算法

- 直接最优化 $p(\mathbf{x}|\theta)$ 是相当困难的，但最优化完备数据集 $p(\mathbf{x},\mathbf{z}|\theta)$ 的似然函数会简单一些。引入一个分布*q*(*z*) 把对数似然函数

$\ln p(\mathbf{x}|\theta)$ 分解成以下形式：

$$\ln p(\mathbf{x}|\theta) = L(q,\theta) + KL(q\,\|\,p)$$

其中 $$L(q,\theta) = \sum_{Z} q(\mathbf{z}) \ln\{\frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})}\}$$

$$KL(q\,\|\,p) = -\sum_{Z} q(\mathbf{z}) \ln\{\frac{p(\mathbf{z}|\mathbf{x},\theta)}{q(\mathbf{z})}\}$$

EM 算法

式中 $KL(q\|p)\geq 0$ ， 因为 $\ln p(\mathbf{x}|\theta)$ 与*q*(***z***)无关，所以最大化 $L(q,\theta)$ 只需要让 $KL(q\|p)$=0 即可。当且仅当 $q(\mathbf{z})=p(\mathbf{z}|\mathbf{x},\theta^{old})$时，$KL(q\|p)$ 取到最小值0，于是：

$$L(\mathrm{q},\theta)=\sum_{Z}p(\mathbf{z}|\mathbf{x},\theta^{old})\,\mathrm{lnp}(\mathbf{x},\mathbf{z}|\theta)-\sum_{Z}p(\mathbf{z}|\mathbf{x},\theta^{old})\,\mathrm{lnp}(\mathbf{z}|\mathbf{x},\theta^{old})$$

$$=Q(\theta,\theta^{old})+\mathrm{const}$$

由上式可得到一个新的 θ 。

EM 算法

- **EM算法**是一个寻找参数的最大似然解的两阶段迭代优化技术，第一段求期望（E步）和第二段最大化（M步）。
- **E步**：令 $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}, \theta^{old})$ 后，最大化$L(q,\theta)$得到一个新的参数 θ。
- **M步**：更新参数 $\theta^{new} = \arg\max(\mathrm{Q}(\theta, \theta^{old}))$ 。

混合正态分布的EM估计

- 混合正态分布（Gaussian Mixture Distribution）是指：

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x} \mid \mu_k, \Sigma_k)$$

- 其中K可以理解为是这个混合正态分布中正态分布的个数。隐变量**z**的分布为:

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \quad \sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$$

混合正态分布的EM估计

- $\mathbf{x}$关于z的条件分布为：

$$p(\mathbf{x}\mid\mathbf{z})=\prod_{k=1}^{K}N(\mathbf{x}\mid\mu_k,\Sigma_k)^{z_k}$$

似然函数：

$$\ln p(X\mid\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(\mathbf{x}_n\mid\mu_k,\Sigma_k)\}$$

1. 估计μ_k：对似然函数关于μ_k求偏导并另其为零，可得：

$$0=-\sum_{n=1}^{N}\underbrace{\frac{\pi_k N(\mathbf{x}_n\mid\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_j N(\mathbf{x}_n\mid\mu_j,\Sigma_j)}}_{\gamma_{z_{nk}}}\Sigma_k(\mathbf{x}_n-\mu_k)$$

混合正态分布的EM估计

- 解得 $\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$

2. 估计 Σ_k： 对似然函数求关于 Σ_k 求偏导并令其等于零可得：

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

3. 估计 π_k： 考虑到 $\sum_{k=1}^{K}\pi_k = 1$的约束。似然函数变为：

$$\ln p(X \mid \pi, \mu, \Sigma) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

混合正态分布的EM估计

对其求 π_k 的偏导得到：

$$0=\sum_{n=1}^{N}\frac{N(\mathbf{x}_n\mid\mu_k,\Sigma_k)}{\sum_j\pi_j N(\mathbf{x}_n\mid\mu_j,\Sigma_j)}+\lambda$$

利用隐函数的条件概率可以得到 $\lambda=-N$

化简得：

$$\pi_k=\frac{N_k}{N}$$

混合正态分布的EM估计

- 算法步骤：

1. 初始化均值μ_k，协方差Σ_k，以及混合系数π_k，并估计初始对数似然函数值；
2. E步：计算在这组参数下z的后验概率下 z_{nk} 的期望$\gamma(z_{nk})$；
3. M步：使用z_{nk}的期望$\gamma(z_{nk})$重新估计参数的最大值。

$$\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

混合正态分布的EM估计

- 算法步骤：

$$\Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k^{new} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

4. 估计对数似然函数：

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(\mathbf{x}_n \mid \mu_k, \Sigma_k)\}$$

并检测是否达到收敛条件，若没有达到则继续执行第二步。

3.5 非参数方法

- 前几节的结论是基于概率密度的分布形式已知的假设。
- 实际问题并不一定满足这个假设。
  - 经典的参数密度是单峰的。
  - 实际的问题包含多峰的密度。
- 模式分类的非参数方法
  - 根据样品模式估计密度函数$P(\mathbf{x} \mid \omega_j)$，然后利用Bayes公式；
  - 直接估计后验概率$P(\omega_j \mid \mathbf{x})$。

非参数方法

- 非参数方法
  - 概率密度的估计
  - Parzen窗估计法
  - 近邻估计

3.6 概率密度的估计

- 基础：
  - 一个样本X落在区域R里的概率P为

$$P = \int_R p(\mathbf{x}) d\mathbf{x}$$

概率P是密度函数p($\mathbf{x}$)的一种经过平均后的形式，对P作估计就是估计出$p(\mathbf{x})$的这个平均值。

- 概率密度估计
  - 设样本 $\mathbf{x}_1, \cdots, \mathbf{x}_n$ 是按照概率密度 $p(\mathbf{x})$ 独立抽取的，n个样本中有k个落在区域R里的概率符合二项定律。

$$P_k = C_n^k P^k (1 - P)^{n-k}$$

其中，P是1个样本落在区域R里的概率。

概率密度的估计

k是一个随机变量，k的期望值是

$$E(k) = nP$$

由于k的二项分布在均值附近有一个峰值，所以k/n是P的一个很好的估计。

假设$p(\mathbf{x})$连续，且R小到$p(\mathbf{x})$在R上几乎没有什么变化，则，

$$P = \int_R p(\mathbf{x})d\mathbf{x} \approx p(\mathbf{x}) \cdot \int_R 1 d\mathbf{x} = p(\mathbf{x}) \cdot V$$

其中，$\mathbf{x}$是R中的一点，V是被R包围的体积。

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

概率密度的估计

- 讨论(1) $p(\mathbf{x}) \approx \dfrac{k/n}{V}$
  - 体积V固定，如果样本取得越来越多，则比值k/n将在概率上按预计的收敛，因此得到一个$p(\mathbf{x})$的空间平均估计值，

$$\frac{P}{V} = \frac{\int_R p(\mathbf{x})d\mathbf{x}}{\int_R 1 d\mathbf{x}}$$

  - 若要想得到$p(\mathbf{x})$，必须让V趋于0。
    - 如果固定样本数n，让V趋于0，则区域不断缩小，以至最后不包含任何样本，而$p(X) \approx 0$ 的估计没有意义。若恰好有几个样本和$\mathbf{x}$重合，则估计值就发散到无穷大，同样也没有意义。

概率密度的估计

- 讨论(2)
  - 实际上，样本的数目有限，所以体积不允许任意小，因此密度函数是一定范围内的平均值。
  - 理论上，假设可以利用的样本数无穷，可以利用极限的方法来研究密度函数的估计。即，

构造一个包含$\mathbf{x}$在内的区域序列$R_1,\cdots,R_n$，设R_n的体积是V_n，其中的样本数为k_n，则

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

什么条件？ make → $p_n(\mathbf{x}) \to p(\mathbf{x})$

概率密度的估计

- 三个条件：

1. $\lim_{n \to \infty} k_n = \infty$
2. $\lim_{n \to \infty} V_n = 0$
3. $\lim_{n \to \infty} k_n / n = 0$

$$\Rightarrow \lim_{n \to \infty} p_n(\mathbf{x}) = p(\mathbf{x})$$

  - n增大时，落入V_n中样本数k_n也要增加；
  - 同时，V_n应不断减少，以使$p_n(\mathbf{x})$趋于$p(\mathbf{x})$；
  - 在小区域V_n中尽管落入了大量样本，但相对于样本总数，这个数量仍然很小；
  - 为了防止V_n下降太快，必须控制使之下降比V_n/n的下降慢一些，例如$V_n = 1/\sqrt{n}$。

概率密度的估计

- 概率密度估计的结论及方法的演变
  - Parzen窗：在具有一定数量的样本时，可以选定一个中心在**x**处的体积V_n，然后计算落入其中的样本数k_n来估计局部密度$p_n(\mathbf{x})$的值。
  - k_n近邻估计：选定一个k_n值，以**x**为中心建立一个体积V_n，让V_n不断增大，直到它能捕获k_n个样本，这是的体积V_n即用来计算$p_n(\mathbf{x})$的估值。
- 问题
  - 样本有限时，上述两种方法的性能难以估计。

3.6.1 Parzen窗估计法

- Parzen窗函数

$$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)=\begin{cases}1 & |\mathbf{x}-\mathbf{x}_i|\le\dfrac{h_n}{2}, j=1,2,\cdots,d\\ 0 & \textbf{otherwise}\end{cases}$$

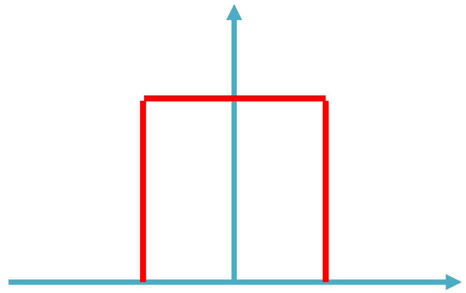
其中， $\mathbf{x}$是d维空间中要估计概率密度值$p_n(\mathbf{x})$的点，V_n是以$\mathbf{x}$为中心边长为h_n的超立方体。 $\mathbf{x}_i$是样本，落在V_n中的样本数k_n是

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

Parzen窗函数

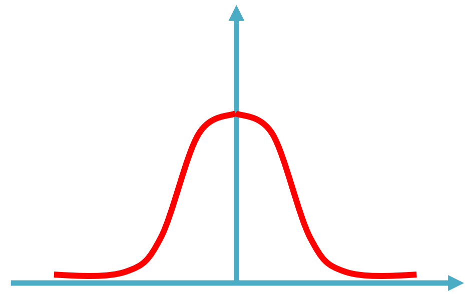
- 方窗函数

$$\varphi(u)=\begin{cases}1 & |u|\le\frac{1}{2}\\ 0 & \textbf{otherwise}\end{cases}$$



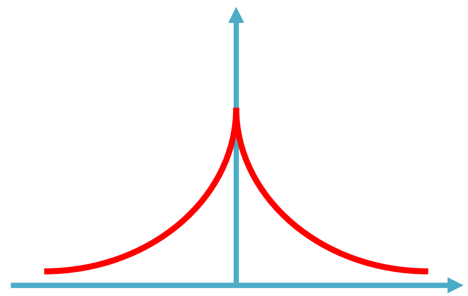
- 正态窗函数

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



- 指数窗函数

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



Parzen窗估计法

- h_n对的$p_n(\mathbf{x})$影响

  若

  $$\delta_n(\mathbf{x}) = \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right) \qquad V_n = h_n^{\ d}$$

  则

  $$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{x}_i)$$

  – h_n既影响 $\delta_n(X)$ 的幅度，又影响它的宽度
  – V_n或h_n太大，估计的分辨率太低，平滑的结果
  – V_n或h_n太小，估计的统计变动太大，不稳定的“噪声性”的估计

Parzen窗估计法

- $p_n(\mathbf{x})$收敛性的讨论

如果$p_n(\mathbf{x})$满足，

$$\lim_{n\to\infty} p_n(\mathbf{x}) = p(\mathbf{x})$$

$$\lim_{n\to\infty} \delta_n^{\,2}(\mathbf{x}) = 0$$

则称$p_n(\mathbf{x})$均方收敛于$p(\mathbf{x})$。

Parzen窗估计法

- 例子：正态分布

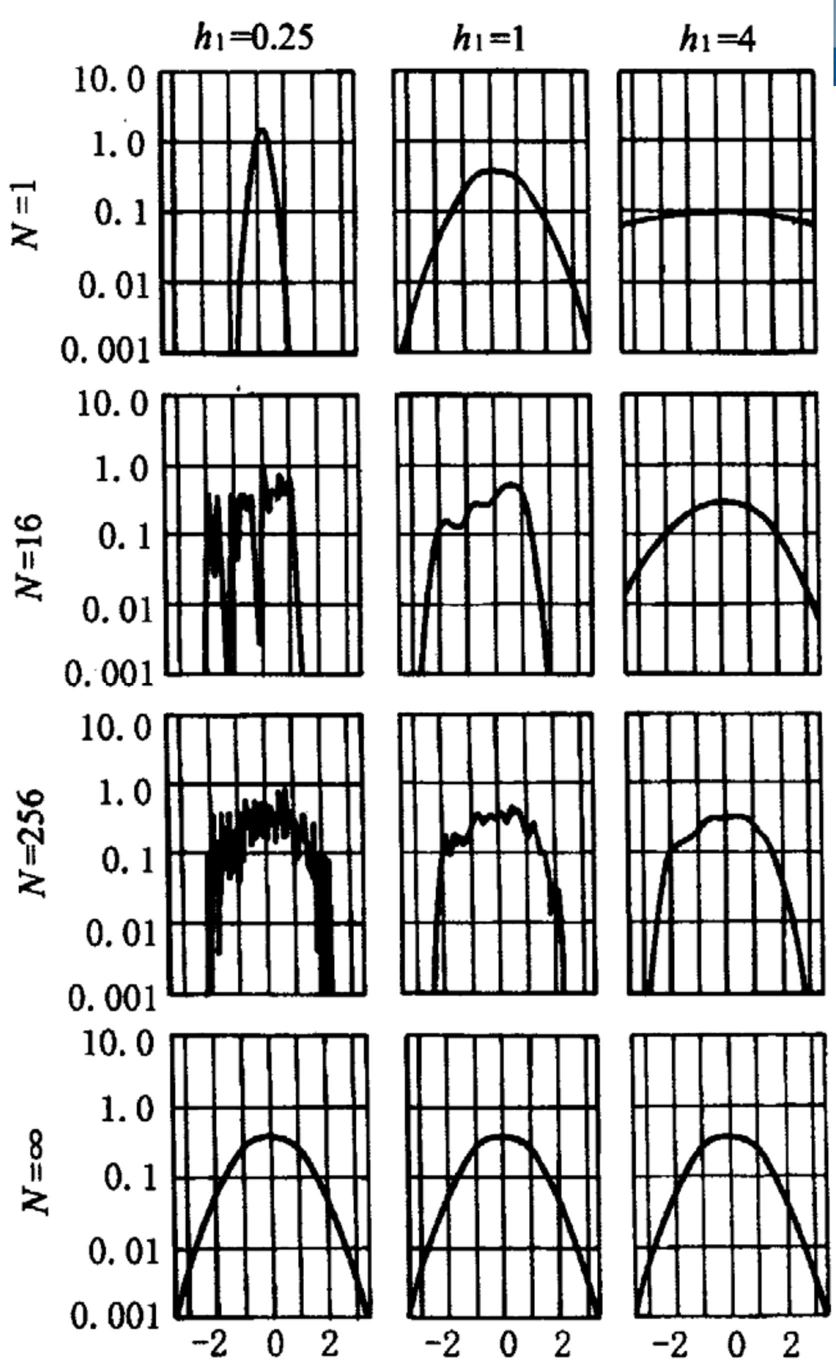
$$p(X) \sim N(0,1)$$

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

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

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

平滑的正态曲线



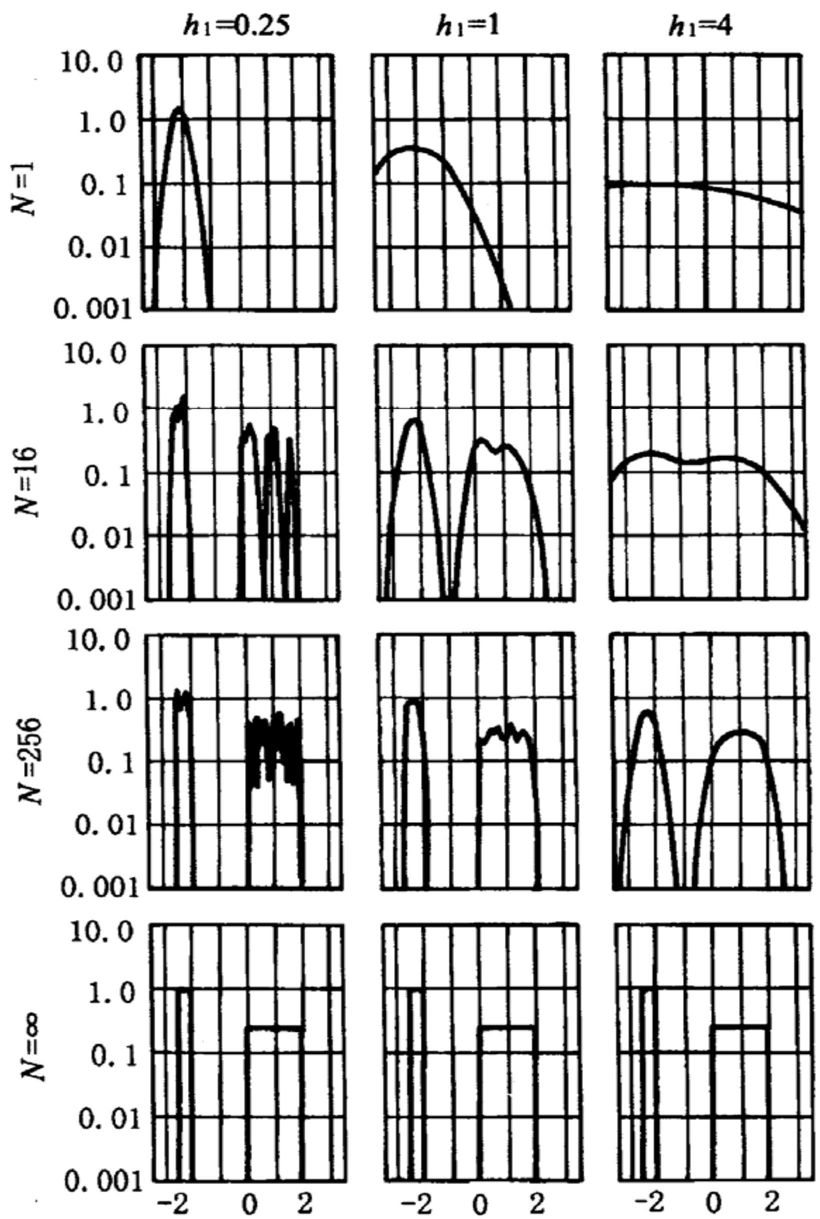
Parzen窗估计法

- 例子：二个均匀分布密度的混合。

$$p(\mathbf{x}) = \begin{cases} 1, & -2.5 < x < -2 \\ 0.25, & 0 < x < 2 \\ 0 & otherwise \end{cases}$$

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

混合的方波分布



Parzen窗估计法

- 结论
- 优点
  - 一般情况下，无论对单峰分布或双峰混合情况，非参数方法都适用
- 缺点
  - 要求的样本数目很大，同时增加计算的成本
  - 如果特征维数增加，样本的数目按指数速度增长，导致“维数灾难”

3.6.2 近邻估计

- Parzen法的问题
  - 单元序列 $V_1,\cdots,V_n$ 的选择问题。对于某组数据适用的V不一定适用于其他数据。
- 策略
  - 建立单元序列和数据之间的函数关系，而不是简单地和样本数目相关。
  - 在数据X的周围建立一个单元并让它不断地增大直至捕获k_n个样品——$\mathbf{x}$的k_n个近邻。

近邻估计

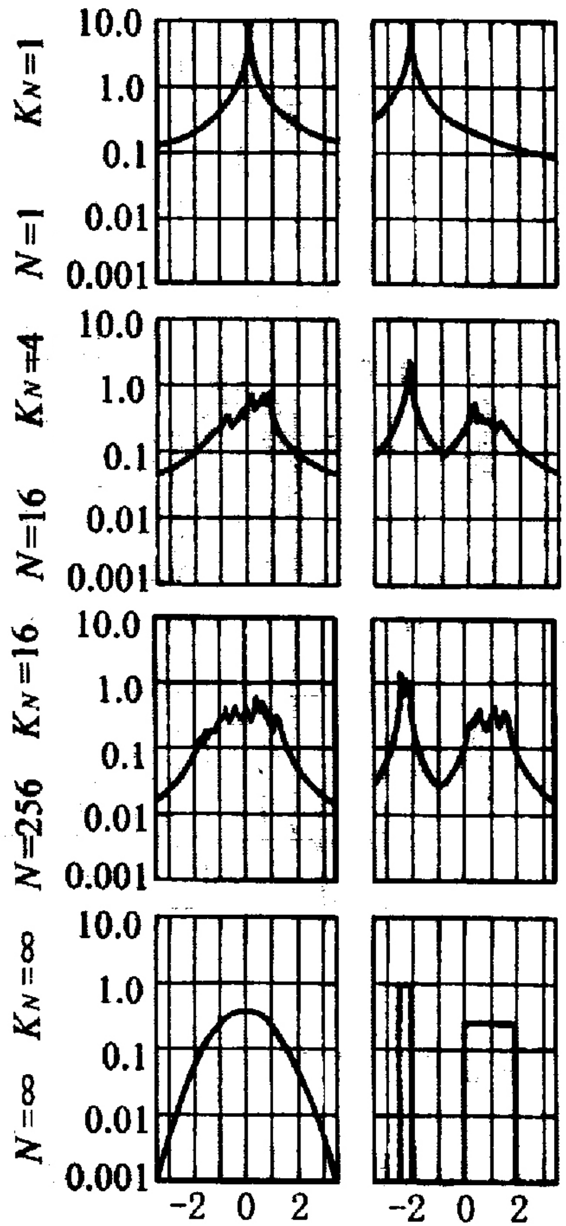
- k_n近邻的选择方法
  利用欧氏距离作准则，根据$\mathbf{x}$到它的第k个近邻(k-NN)的欧氏距离$r_n(\mathbf{x})$来估计体积V_n

$$p_n(\mathbf{x}) = \frac{A}{\left[r_k(\mathbf{x})\right]^d}$$

d是特征空间的维数，A是常数，由k_n和n决定

近邻估计

- 例子
  - 单一正态分布
  - 两个均匀分布的估计



小 结

- 最大似然估计
- 贝叶斯估计
- EM估计
- 非参数估计