贝叶斯基础与分类

郑 伟 诗

http://www.isee-ai.cn/~zhwshi/

中山大学

机器智能与先进计算
教育部重点实验室

贝叶斯基础与分类

一. 贝叶斯分类原理

二. 概率密度估计的参数化方法

三. 非参数方法

四. 因果发现与推断初步

第一部分：贝叶斯分类原理

贝叶斯分类原理

- 假设存在c类对象，ω_j表示第j类标签，$j = 1, 2, \cdots, c$，$\mathbf{x}$表示对某一待分类对象观察到的特征向量。
- 似然值：类条件概率密度函数$p(\mathbf{x}|\omega_j)$。
- 证据因子：

$$p(\mathbf{x}) = \sum_{j=1}^{c} [p(\mathbf{x}|\omega_j) \cdot P(\omega_j)]$$

- 后验概率：贝叶斯分类的准则

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j) \cdot P(\omega_j)}{p(\mathbf{x})}$$

即：后验概率 = （似然值 · 先验概率）/证据因子

- 贝叶斯：先验&似然&后验
- 贝叶斯公式：$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$
- 贝叶斯公式（全概率）：$P(B_i|A_j) = \frac{\prod_{j=1}^{n} P(B_i)P(A_j|B_i)}{\sum_{i=1}^{n}\prod_{j=1}^{n} P(B_i)P(A_j|B_i)}$ 其中先验概率是P(\theta)，似然是P(D|\theta)，后验是P(\theta|D)
- 朴素贝叶斯模型（NBM）：基于贝叶斯定理和特征条件独立假设（对条件概率分布做了条件独立性的假设）的方法。假设给定目标值时属性之间相互条件独立。

贝叶斯分类原理

- 以两类为例（即假设$c = 2$）
  - 如果将$\mathbf{x}$判定为ω_2，则分类错误的概率为：
    $P(\text{error}|\mathbf{x}) = P(\omega_1|\mathbf{x}) = 1 - P(\omega_2|\mathbf{x})$
  - 如果将$\mathbf{x}$判定为ω_1，则分类错误的概率为：
    $P(\text{error}|\mathbf{x}) = P(\omega_2|\mathbf{x}) = 1 - P(\omega_1|\mathbf{x})$

- 贝叶斯决策：最小化错误概率
  若$P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$，判定为$\omega_1$，否则判定为$\omega_2$。此时：

$$P(\text{error}|\mathbf{x}) = \min[P(\omega_1|\mathbf{x}),\ P(\omega_2|\mathbf{x})]$$

- 多类情况下最小化错误概率的判定类别：

$$\omega^* = \underset{j}{\text{argmax}}\, P(\omega_j|\mathbf{x})$$

贝叶斯分类原理

- 行动风险：设$\{\omega_1, \omega_2, \ldots, \omega_c\}$为有限的$c$个类别集，$\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$表示有限的$a$种可能采取的行动集，$\lambda(\alpha_i \mid \omega_j)$表示类别状态$\omega_j$下采取行动$\alpha_i$的风险。
- 观察到对象特征$\mathbf{x}$时采取行动$\alpha_i, i = 1, 2, \cdots, a$，的条件风险：

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

- 总风险为$\mathbf{x}$空间中条件风险的总和，即：

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

其中$\alpha(\mathbf{x})$表示针对$\mathbf{x}$所采取的行动。

- 最小化R ⇔最小化每一个$R(\alpha(\mathbf{x})|\mathbf{x})$，即针对每一个$\mathbf{x}$，选择使$R(\alpha_i|\mathbf{x})$最小的$\alpha_i$。最小化后的总风险$R$称为贝叶斯风险，它是可获得的最优效果。

贝叶斯分类原理

- 对于包含二类的风险最小化问题
  $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$：类别为$\omega_j$而采取行动$\alpha_i$时产生的损失

条件风险：

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11}P(\omega_1 \mid \mathbf{x}) + \lambda_{12}P(\omega_2 \mid \mathbf{x})$$

$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21}P(\omega_1 \mid \mathbf{x}) + \lambda_{22}P(\omega_2 \mid \mathbf{x})$$

降低风险的决策规则：
如果$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$，则采取行动$\alpha_1$。

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$$

$$\Leftrightarrow (\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$$

$$\Leftrightarrow \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$ （似然比）

贝叶斯分类原理

- 最小化错误率的分类：当采用下面的“0-1损失”函数时，风险最小化的决策结果就是最小的分类错误率

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i,j = 1,\ldots,c$$

其中α_i 表示“判定类别为ω_i”。此时，条件风险就是分类错误的概率：

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

$$= \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

最小化风险决策就是最小化分类错误率的决策，需要最大化后验概率$P(\omega_i \mid \mathbf{x})$，即：

如果$P(\omega_i \mid \mathbf{x}) \geq P(\omega_j \mid \mathbf{x})$，$\forall j \neq i$，则将$\mathbf{x}$判定为$\omega_i$。

贝叶斯分类原理

- 决策域：令$\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}.\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda$，则

  当似然比$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}>\theta_\lambda$时，判定$\mathbf{x}$的类别为$\omega_1$。

- 如果λ是“0-1损失”函数，即$\lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$，则$\theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}$（记为$\theta_a$）。

  否则决策域会随风险调节，例如

  当$\lambda=\begin{pmatrix}0 & 1.2\\ 1 & 0\end{pmatrix}$时，$\theta_\lambda=\frac{1.2P(\omega_2)}{P(\omega_1)}$（记为$\theta_b$）。

贝叶斯分类原理

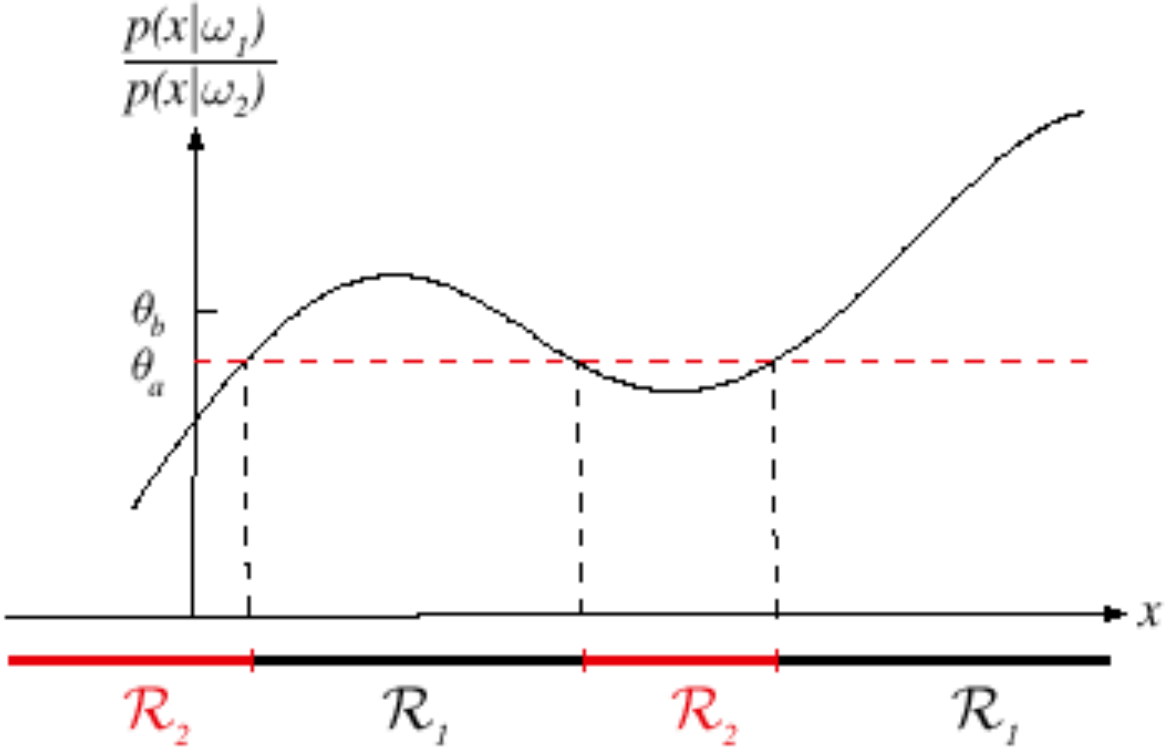


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

第二部分：概率密度估计的参数化方法

- 最大似然估计
- 贝叶斯估计
- 期望最大化

概率密度估计

- 设计贝叶斯最优分类器需要
  - $P(\omega_i)$（先验概率）
  - $p(\mathbf{x}|\omega_i)$ (类条件概率密度)

很不幸，绝大多数情况下信息并不完整。

由训练样本设计分类器时，通常容易得到先验概率，困难往往在于准确估计类条件概率密度，特别是在特征空间维数过高时。

在概率密度采用参数化的函数形式时，对应的概率密度估计问题成为参数估计问题。

以一个简单的d维正态密度为例：

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

为准确估计其密度，需对其中参数$\boldsymbol{\mu}$（均值向量）、$\boldsymbol{\Sigma}$（协方差矩阵）进行估计。而参数量按d的平方增长。

概率密度估计

- 常用参数估计方法：最大似然估计法和贝叶斯估计法。两种方法结果通常很接近，但是做法有所不同：
  - 最大似然估计法将待估计参数视为固定但未知，通过最大化观察到训练样本的概率确定最佳参数
  - 贝叶斯估计法将待估计参数看成是符合某种已知先验概率分布的随机变量，但需要根据观察到的训练样本进一步明确其分布

概率密度估计——最大似然估计法

- 最大似然估计
  - 随着训练样本容量增大时收敛效果好
  - 通常比其他方法更为简单

- 最大似然估计的基本原理：
  背景：假设我们有c个类别，为表达方便，记

$$p(\mathbf{x} \mid \omega_j) \equiv p(\mathbf{x} \mid \omega_j, \theta_j)\ ,\ j = 1, \ldots, c。$$

  根据已有的训练样本来估计参数向量$\theta = (\theta_1, \theta_2, \ldots, \theta_c)$，其中$\theta_i$，$i = 1, 2, \ldots, c$是第$c$类的参数向量。

概率密度估计——最大似然估计法

- 设训练集D含有来自某类的n个样本$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$，如果假设这些样本是独立的，则

$$p(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

为简化后续推导，上式中类标被省略。$p(D|\boldsymbol{\theta})$被称为给定数据集下参数向量$\boldsymbol{\theta}$的似然函数。

- 参数向量$\boldsymbol{\theta}$的最大似然估计，就是最大化$p(D|\boldsymbol{\theta})$的$\boldsymbol{\theta}$取值，记为$\hat{\boldsymbol{\theta}}$。其思想是寻找最符合观察到的训练样本的$\boldsymbol{\theta}$。

概率密度估计——最大似然估计法

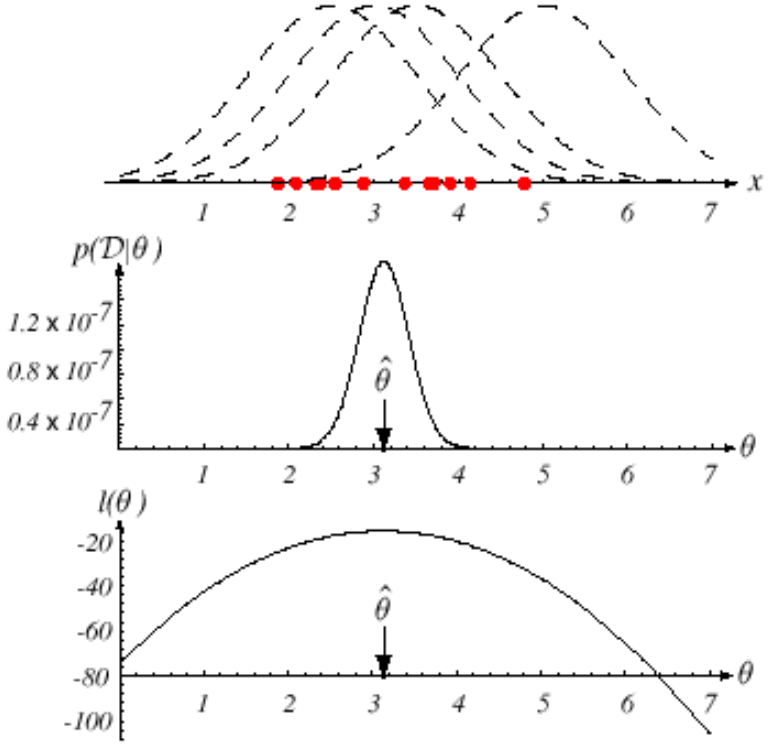


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

概率密度估计——最大似然估计法

- 最优估计
  - 设$\boldsymbol{\theta}$的维数为p，即$\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$，$\nabla_{\boldsymbol{\theta}}$为如下的梯度算子

$$\nabla_{\boldsymbol{\theta}}=\left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_p}\right]^{\mathrm{T}}$$

  - 定义对数似然函数

$$l(\boldsymbol{\theta})=\ln p\,(D|\boldsymbol{\theta})$$

  - 将最大似然估计问题转换为：求最大化对数似然函数$l(\boldsymbol{\theta})$的参数$\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})=\arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{n} \ln p\,(\mathbf{x}_k|\boldsymbol{\theta})$$

从而

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})=\sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p\,(\mathbf{x}_k|\boldsymbol{\theta})$$

最优解的必要条件：$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})=\mathbf{0}$

概率密度估计——最大似然估计法

- 以一个特殊情况为例：前述正态密度中μ未知而Σ已知($\theta = \mu$)，待估计密度函数为$p(\mathbf{x}_k \mid \boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

而
$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

因此，μ 的最大似然估计必须满足：

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0}$$

两边左乘∑并整理得：

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

即μ的最大似然估计为训练样本均值。

概率密度估计——最大似然估计法

- 更一般的情况：正态分布中μ和Σ均未知
  考虑单变量情况：$\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l_k(\boldsymbol{\theta}) = \ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l_k(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}(\ln p(x_k|\theta)) \\ \frac{\partial}{\partial\theta_2}(\ln p(x_k|\theta)) \end{bmatrix} = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

全体训练样本的对数似然函数的极值条件：

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

解得：$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x}$；$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$

概率密度估计——最大似然估计法

- 该极大似然估计$\hat{\sigma}^2$是有偏估计量，因：

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

协方差矩阵Σ的无偏估计量：

$$\hat{\Sigma} = \underbrace{\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^{\mathrm{T}}}_{\text{样本协方差矩阵}}$$

概率密度估计——贝叶斯估计法

- 贝叶斯估计：待估计参数θ被视为随机变量，用后验概率密度$p(\theta \mid D)$来估计θ，从而获得$p(\mathbf{x}|\omega, D)$（D为某一类别的训练数据）
- 以单变量正态分布为例：$p(\mu \mid D)$，μ是唯一的未知参数

$$p(x|\mu) \sim N(\mu, \sigma^2), \qquad p(\mu) \sim N(\mu_0, \sigma_0^2)$$

（μ_0, σ_0，和σ均已知）

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu} = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

再生密度函数：$p(\mu|D) \sim N(\mu_n, \sigma_n^2)$

其中$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \quad (\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k)$

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

概率密度估计——贝叶斯估计法

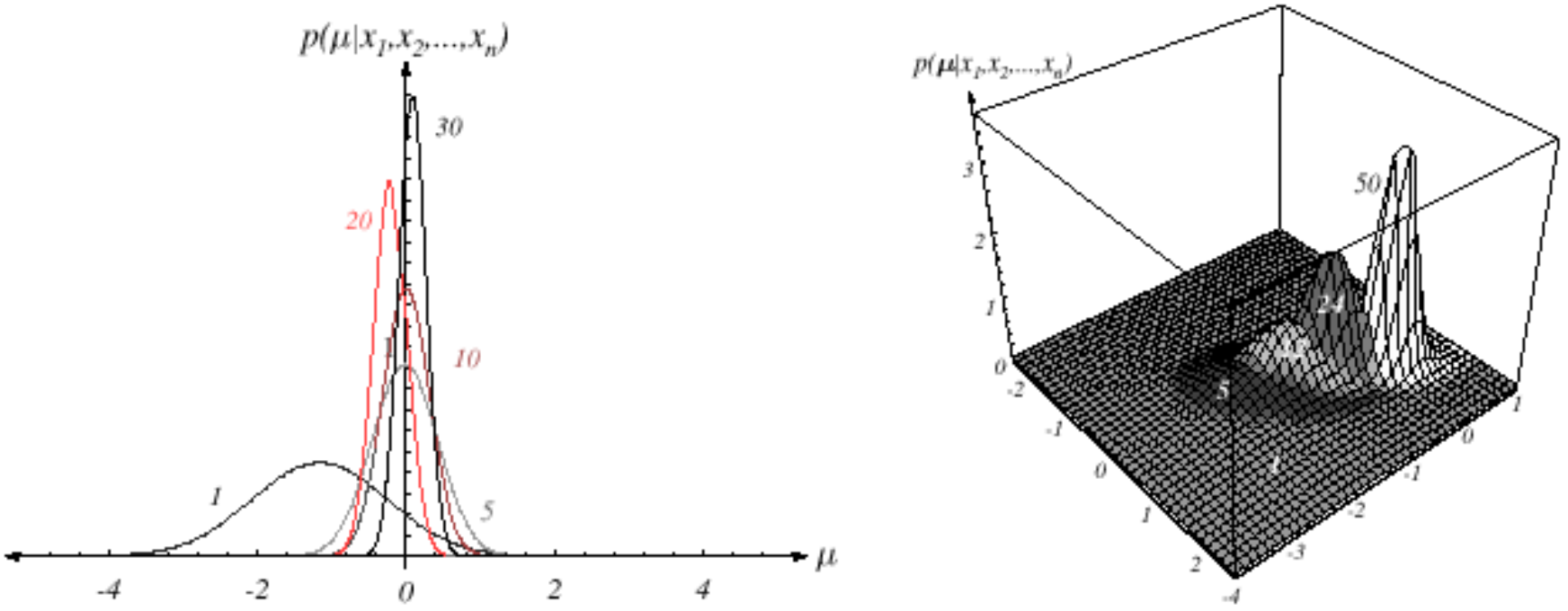


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

概率密度估计——贝叶斯估计法

- $p(\mu|D)$已获得，接下来估计概率密度$p(x|\omega,D)$，简记为$p(x|D)$

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

积分结果为一正态分布：

$$p(x|D)\sim N(\mu_n,\sigma^2+\sigma_n^2)$$

（实际上是我们希望获得的类条件概率密度$p(x|\omega_j,D_j)$）

- 至此已得到$p(x|\omega_j,D_j)$与$P(\omega_j)$。利用贝叶斯公式，可获得贝叶斯分类准则：

$$\omega^*=\arg\max_{\omega_j}P(\omega_j|x,D)\equiv\arg\max_{\omega_j}p(x|\omega_j,D_j)P(\omega_j)$$

概率密度估计——贝叶斯估计法

- 贝叶斯参数估计的一般理论：该方法可应用于未知概率密度具有参数化的函数形式时。其中的基本假设包括：

  1. 概率密度函数$p(\mathbf{x} \mid \boldsymbol{\theta})$的形式已知，但参数向量$\boldsymbol{\theta}$具体取值未知；

  2. 关于参数向量$\boldsymbol{\theta}$的知识包含在已知的先验分布$p(\boldsymbol{\theta})$中；

  3. 其余关于参数向量$\boldsymbol{\theta}$的信息包含在观察到的训练样本$D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$中。

  此时的基本问题是：

  1. 计算关于$\boldsymbol{\theta}$的后验概率密度$p(\boldsymbol{\theta}|D) = \dfrac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$，

  其中由于样本之间的独立性假设，$p(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$；

  2. 计算指定样本$\mathbf{x}$的概率密度$p(\mathbf{x}|D)$。

概率密度估计——期望最大化

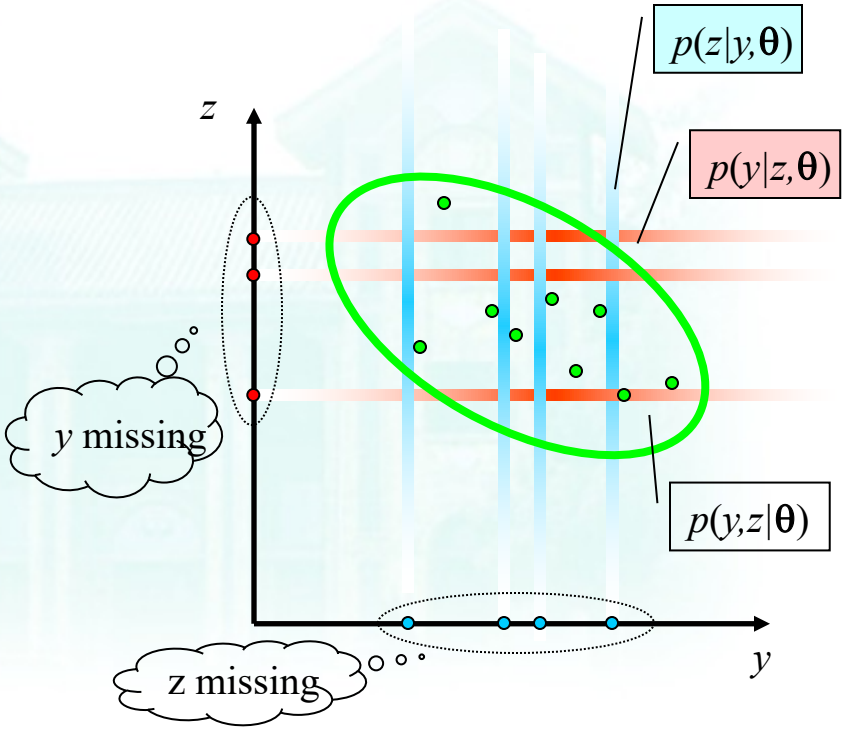
- 期望最大化可用于在数据不完整时进行概率密度估计。
- 数据不完整的一些情况：
  - 问题固有的部分信息无法获得，例如：
    - 高斯混合模型：数据点属于哪个聚类？
  - 数据丢失/出错，例如：
    - 产生数据时出现故障或噪声、错误擦除部分数据

- 问题描述：
  - 完整的训练样本集$D = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$来自某一待估计概率分布；
  - 假设部分特征丢失，为便于分析，将样本点表示为$\mathbf{x}_k = \{\mathbf{x}_{k\mathrm{g}}, \mathbf{x}_{k\mathrm{b}}\}$，表示特征包含“好”、“坏”（丢失）两部分；
  - 同理，将训练集D按“好”、“坏”特征分为D_g和D_b两个集合，全部的特征集合则为$D = D_\mathrm{g} \cup D_\mathrm{b}$。

概率密度估计——期望最大化

- 给定来自某待估计参数化分布的样本$\mathbf{x}_k \in R^d$
- 在某些$\mathbf{x}_k$中，部分维度丢失，知道丢失的位置
- 如何估计概率分布的参数？
- 如何替换丢失的数据？

- 期望最大化的基本思想：
  - 假如我们已获得对“好”、“坏”联合密度的估计，则条件密度将给出缺失数据的分布；
  - 假如对缺失数据的分布有所估计，则我们可以用它来估计联合密度。
- 有一种方法可以对上述两个步骤进行迭代，从而稳步提高总的似然函数。



概率密度估计——期望最大化

- 我们希望能在数据不完整的情况下进行最大似然密度估计。假如我们知道缺失的数据，我们就可以直接最大化对数似然函数(g = 已知, b = 丢失)：

$$L_{\mathrm{gb}}(\boldsymbol{\theta}) = \sum_k \ln p\left(\mathbf{x}_{kg}, \mathbf{x}_{kb} | \boldsymbol{\theta}\right)$$

- 遗憾的是，我们并不知道丢失的数据D_{b}，我们能做的只能是最大化“好”数据D_{g}的对数似然：

$$L_{\mathrm{g}}(\boldsymbol{\theta}) = \ln p\left(D_{\mathrm{g}} | \boldsymbol{\theta}\right) = \sum_k \ln p\left(\mathbf{x}_{k\mathrm{g}} | \boldsymbol{\theta}\right)$$

$$= \sum_k \ln \int p(\mathbf{x}_{k\mathrm{b}}, \mathbf{x}_{k\mathrm{g}} | \boldsymbol{\theta}) d\mathbf{x}_{k\mathrm{b}}$$

- 不过这个积分的对数形式给我们的优化问题带来麻烦。
- 解决方法：采用一种迭代策略，求解逐步优化的$(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots)$

概率密度估计——期望最大化

- 为简化表达，将$p(\cdot|\boldsymbol{\theta}^i)$记为$p^i(\cdot)$
- 重写对数似然的表达式：

$$L_{\mathrm{g}}(\boldsymbol{\theta}^i) = \ln p^i(D_{\mathrm{g}}) = \int p^i(D_{\mathrm{b}}|D_{\mathrm{g}}) \ln p^i(D_{\mathrm{g}}) dD_{\mathrm{b}}$$

$$= \int p^i(D_{\mathrm{b}}|D_{\mathrm{g}}) \ln \left[p^i(D_{\mathrm{g}}) \frac{p^i(D_{\mathrm{b}}|D_{\mathrm{g}})}{p^i(D_{\mathrm{b}}|D_{\mathrm{g}})} \right] dD_{\mathrm{b}}$$

$$= \int p^i(D_{\mathrm{b}}|D_{\mathrm{g}}) \ln p^i(D_{\mathrm{b}}, D_{\mathrm{g}}) dD_{\mathrm{b}}$$

$$- \int p^i(D_{\mathrm{b}}|D_{\mathrm{g}}) \ln p^i(D_{\mathrm{b}}|D_{\mathrm{g}}) dD_{\mathrm{b}} \equiv Q(\boldsymbol{\theta}^i; \boldsymbol{\theta}^i) + H(\boldsymbol{\theta}^i)$$

概率密度估计——期望最大化

- 类似地：

$$L_{\mathrm{g}}(\boldsymbol{\theta}^i) = \ln p^i(D_{\mathrm{g}}) = \ln \int p^i(D_{\mathrm{b}}, D_{\mathrm{g}}) dD_{\mathrm{b}}$$

$$= \ln \int \frac{p^i(D_{\mathrm{b}}, D_{\mathrm{g}})}{p^{i-1}(D_{\mathrm{b}}|D_{\mathrm{g}})} p^{i-1}(D_{\mathrm{b}}|D_{\mathrm{g}}) dD_{\mathrm{b}}$$

- 利用Jensen不等式：如果 $\int a(x)dx = 1$，$a(x) \geq 0$，则

$$\ln \int a(x)g(x)dx \geq \int a(x) \ln g(x) dx$$

可得：

$$L_{\mathrm{g}}(\boldsymbol{\theta}^i)$$

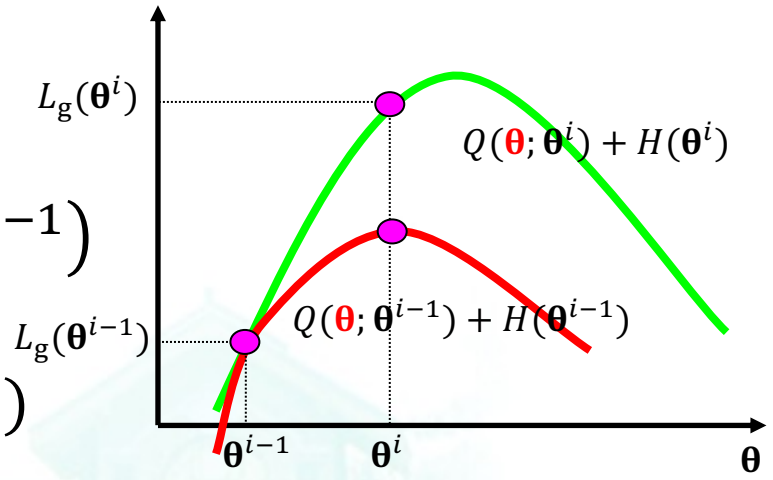
$$\geq \int p^{i-1}(D_{\mathrm{b}}|D_{\mathrm{g}}) \ln p^i(D_{\mathrm{g}}, D_{\mathrm{b}}) dD_{\mathrm{b}}$$

$$- \int p^{i-1}(D_{\mathrm{b}}|D_{\mathrm{g}}) \ln p^{i-1}(D_{\mathrm{b}}|D_{\mathrm{g}}) dD_{\mathrm{b}} \equiv Q(\boldsymbol{\theta}^i; \boldsymbol{\theta}^{i-1}) + H(\boldsymbol{\theta}^{i-1})$$

概率密度估计——期望最大化

- 优化策略：

$$L_{\mathrm{g}}(\boldsymbol{\theta}^{i-1}) = Q(\boldsymbol{\theta}^{i-1};\boldsymbol{\theta}^{i-1}) + H(\boldsymbol{\theta}^{i-1})$$
$$\leq \quad \Leftarrow \quad \downarrow \text{maximize}$$
$$L_{\mathrm{g}}(\boldsymbol{\theta}^{i}) \geq Q(\boldsymbol{\theta}^{i};\boldsymbol{\theta}^{i-1}) + H(\boldsymbol{\theta}^{i-1})$$



- 这表明，如果我们对$\boldsymbol{\theta}^{i}$最大化$Q(\boldsymbol{\theta}^{i};\boldsymbol{\theta}^{i-1})$，则似然函数只会增加
- 总之，在每次迭代时，我们都搜索最大化以下函数的$\boldsymbol{\theta}^{i}$

$$Q(\boldsymbol{\theta}^{i};\boldsymbol{\theta}^{i-1}) = \int p(D_{\mathrm{b}}|D_{\mathrm{g}};\boldsymbol{\theta}^{i-1}) \ln p(D_{\mathrm{g}}, D_{\mathrm{b}}|\boldsymbol{\theta}^{i}) dD_{\mathrm{b}}$$

也就是最大化联合似然函数对数的期望$Q(\boldsymbol{\theta}^{i};\boldsymbol{\theta}^{i-1})$，其中的期望是使用先前估计参数下的条件分布来计算的。

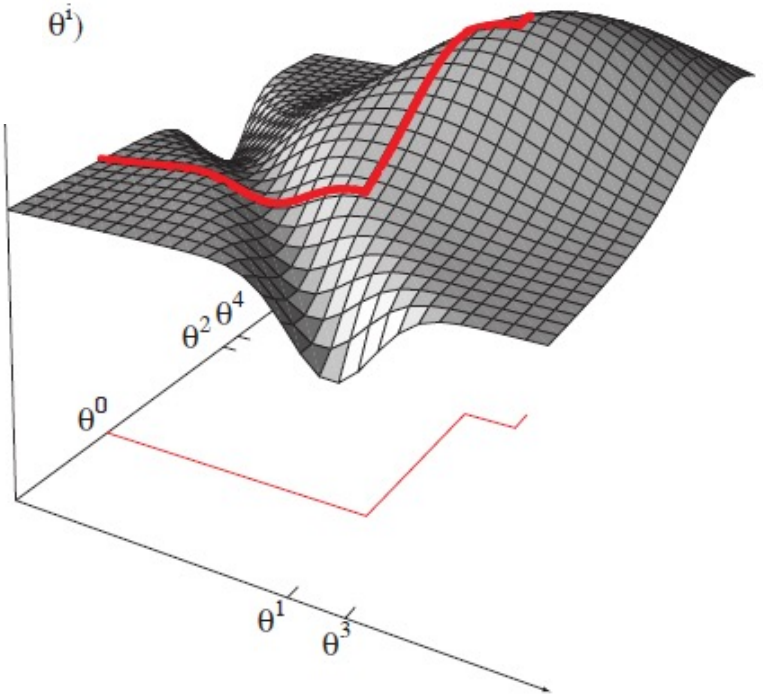
概率密度估计——期望最大化

- 期望最大化算法

```
1 begin initialize θ^0, T, i = 0
2         do i ← i + 1
3             E step : compute Q(θ; θ^i)
5             M step : θ^{i+1} ← arg max_θ Q(θ; θ^i)
6         until Q(θ^{i+1}; θ^i) − Q(θ^i; θ^{i−1}) ≤ T
7   return θ̂ ← θ^{i+1}
8 end
```

运用期望最大化算法寻找最佳模型的过程如下：

从某一个初始的模型参数$\boldsymbol{\theta}^0$开始，然后通过“M步”，找到此时最佳的$\boldsymbol{\theta}^1$。接下来固定$\boldsymbol{\theta}^1$，求出使得$Q(\boldsymbol{\theta}^2;\boldsymbol{\theta}^1)$最优的值$\boldsymbol{\theta}^2$。该过程迭代进行，直到$Q(\boldsymbol{\theta}^i;\boldsymbol{\theta}^{i-1})$不能再增大为止。



概率密度估计——期望最大化

- $Q(\boldsymbol{\theta};\boldsymbol{\theta}^i) = \int p(D_\mathrm{b}|D_\mathrm{g},\boldsymbol{\theta}^i)\ln p\,(D_\mathrm{g},D_\mathrm{b}|\boldsymbol{\theta})dD_\mathrm{b} = E_{D_\mathrm{b}}\left[\ln p\,(D_\mathrm{g},D_\mathrm{b}|\boldsymbol{\theta})\right]$
- 以多维正态分布为例：

$$E_{D_\mathrm{b}}\left[\ln p\,(D_\mathrm{g},D_\mathrm{b}|\boldsymbol{\theta})\right] = \sum_k E_{\mathbf{x}_{k\mathrm{b}}}\left[\ln p\,(\mathbf{x}_{k\mathrm{g}},\mathbf{x}_{k\mathrm{b}}|\boldsymbol{\mu},\boldsymbol{\Sigma})\right]$$

- $\boldsymbol{\mu}$的更新规则：

$$\frac{\partial}{\partial\boldsymbol{\mu}}E_{D_\mathrm{b}}\left[\ln p\,(D_\mathrm{g},D_\mathrm{b}|\boldsymbol{\theta})\right]$$

$$= \sum_{k_{\overline{\pi}}1}^{n} E_{\mathbf{x}_{k\mathrm{b}}}\left[\frac{\partial}{\partial\boldsymbol{\mu}}\left(C_0 - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right)\right]$$

$$= \sum_{k=1} E_{\mathbf{x}_{kb}}[\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})] = \mathbf{0}$$

$$\Rightarrow \boldsymbol{\mu} = \frac{1}{n}\sum_{k=1}^{n} E_{\mathbf{x}_{kb}}[\mathbf{x}_k] = \frac{1}{n}\sum_{k=1}^{n}\begin{bmatrix}\mathbf{x}_{kg}\\ E_{\mathbf{x}_{kb}}[\mathbf{x}_{kb}]\end{bmatrix}$$

概率密度估计——期望最大化

- $\boldsymbol{\Sigma}$的更新规则可类似获得。

- 最终$\boldsymbol{\mu}$与$\boldsymbol{\Sigma}$的更新策略为：

$$\boldsymbol{\mu}^{i+1} = \frac{1}{n}\sum_{k=1}^{n}\begin{bmatrix}\mathbf{x}_{k\mathrm{g}} \\ E[\mathbf{x}_{k\mathrm{b}}]\end{bmatrix}$$

$$\boldsymbol{\Sigma}^{i+1} = \frac{1}{n}\sum_{k=1}^{n}\begin{bmatrix}\mathbf{x}_{k\mathrm{g}}\mathbf{x}_{k\mathrm{g}}^{\mathrm{T}} & \mathbf{x}_{k\mathrm{g}}E[\mathbf{x}_{k\mathrm{b}}]^{\mathrm{T}} \\ E[\mathbf{x}_{k\mathrm{b}}]\mathbf{x}_{k\mathrm{g}}^{\mathrm{T}} & E[\mathbf{x}_{k\mathrm{b}}\mathbf{x}_{k\mathrm{b}}^{\mathrm{T}}]\end{bmatrix} - \boldsymbol{\mu}^{i+1}[\boldsymbol{\mu}^{i+1}]^{\mathrm{T}}$$

第三部分：非参数方法

- Parzen窗法
- K_n近邻估计法

非参数方法

- 典型的参数化概率密度函数是单峰的(具有单一的局部最大值)，而许多实际问题涉及多峰的概率密度。

- 非参数方法适用于任意分布，而不需要假设密度形式已知。

- 有两种非参数方法：
  ——估计$p(\mathbf{x}|\omega_j)$
  ——直接逼近后验概率$P(\omega_j \mid \mathbf{x})$

非参数方法——Parzen窗法

- 基本思想：

  设密度函数为$p(\mathbf{x})$，则向量$\mathbf{x}$落在区域R中的概率为：

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

  P是密度函数$p(\mathbf{x})$的平滑（或平均）版本。

  假设样本大小为n。则其中k个点落在R中的概率为：

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

  k的期望值为：

$$E[k] = nP$$

非参数方法——Parzen窗法

- P的最大似然估计：

$$\hat{P} = \arg\max_P P_k(P) = \frac{k}{n}$$

当区域R很小时，可认为密度在R中近似均匀，从而：

$$P = \int_R p(\mathbf{x}')d\mathbf{x}' \approx p(\mathbf{x})V,$$

其中$\mathbf{x}$为R中的一个点，V是区域R所包含的体积。

综上可得：

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

- 收敛条件：$k/(nV)$是$p(\mathbf{x})$的空间平均值，只有当V趋近于零时，才得到准确的$p(\mathbf{x})$。

非参数方法——Parzen窗法

- 通常来讲，训练样本的个数总是有限的
  - 不能任意减小V
  - 要允许k/n的一定变动
- 解决办法：为估计点$\mathbf{x}$处的概率，构造包含$\mathbf{x}$的区域序列：$R_1, R_2, \cdots$，下标对应逐步增大的样本容量
  - 设V_n为区域R_n的体积，k_n为落在区间R_n中的样本个数，$p_n(\mathbf{x})$表示对$p(\mathbf{x})$的第n次估计：

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

- $p_n(\mathbf{x})$收敛到$p(\mathbf{x})$的三个条件：

1) $\lim_{n\to\infty} V_n = 0$

2) $\lim_{n\to\infty} k_n = \infty$

3) $\lim_{n\to\infty} k_n/n = 0$

非参数方法——Parzen窗法

- 两种获得区域序列的途径：
  ① 从某个初始区域开始按类似$V_n = 1/\sqrt{n}$的规律缩小其体积，这就是“Parzen窗估计法”；

  ② 指定k_n为n的某个函数，例如$k_n = \sqrt{n}$； V_n从$\mathbf{x}$开始增长直至包围$\mathbf{x}$的k_n近邻，这是“k_n最近邻估计法”。

非参数方法——Parzen窗法

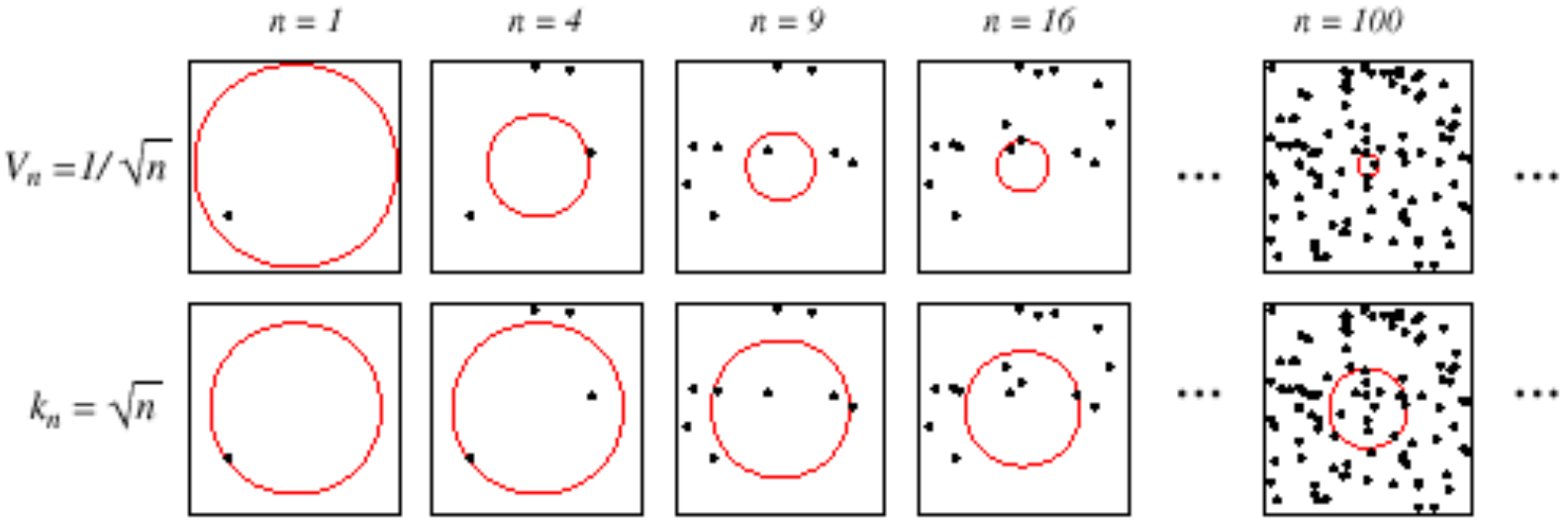


FIGURE 4.2. There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the bottom row, is to decrease the volume in a data-dependent way, for instance letting the volume enclose some number $k_n = \sqrt{n}$ of sample points. The sequences in both cases represent random variables that generally converge and allow the true density at the test point to be calculated. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——Parzen窗法

- 暂时假设区域R_n是一个d维超立方体

$$V_n = h_n^d \text{，其中} h_n \text{为} R_n \text{的边长}$$

- 设$\varphi(\mathbf{u})$为一窗函数：$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \le \frac{1}{2} \quad j = 1,\ldots,d \\ 0 & \text{其他} \end{cases}$

  ⇒则如果$\mathbf{x}_i$落在以$\mathbf{x}$为中心、体积为V_n的超立方体内，$\varphi\left((\mathbf{x}-\mathbf{x}_i)/h_n\right)=1$，否则该式等于0。

  ⇒ 落在以$\mathbf{x}$为中心、体积为V_n的超立方体内样本数为

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

- 可以得到如下估计：

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

上式为函数$\varphi(\cdot)$在各$\mathbf{x}_i$上取值的平均。而$\varphi(\cdot)$可以为更一般的函数。

非参数方法——Parzen窗法

- Parzen窗法应用示例一：$p(x) \sim N(0,1)$

  令

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

  $h_n = h_1/\sqrt{n}$ $(n > 1)$， h_1由人为事先设定

  则

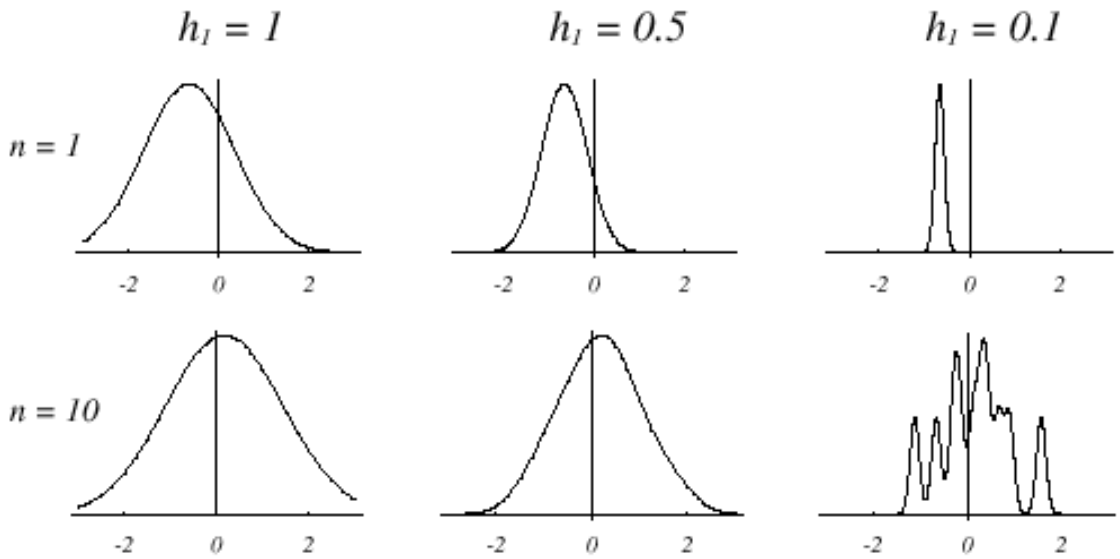
$$p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

  实际上是以样本x_i为中心的n个正态概率密度函数均值。

非参数方法——Parzen窗法

- 数值结果

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



当$n = 1$且$h_1 = 1$时，

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

当$n = 10$且$h_1 = 0.1$时，可清楚观察各样本贡献。

非参数方法——Parzen窗法

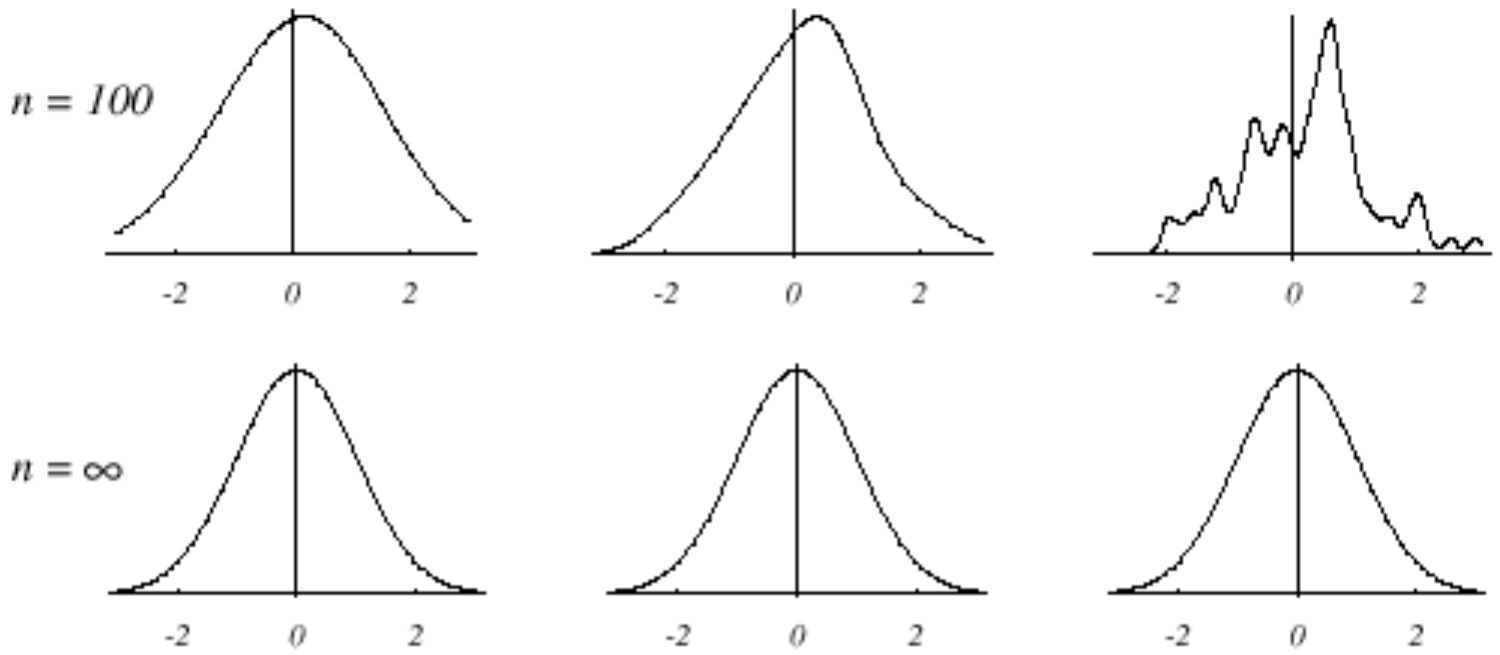


FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——Parzen窗法

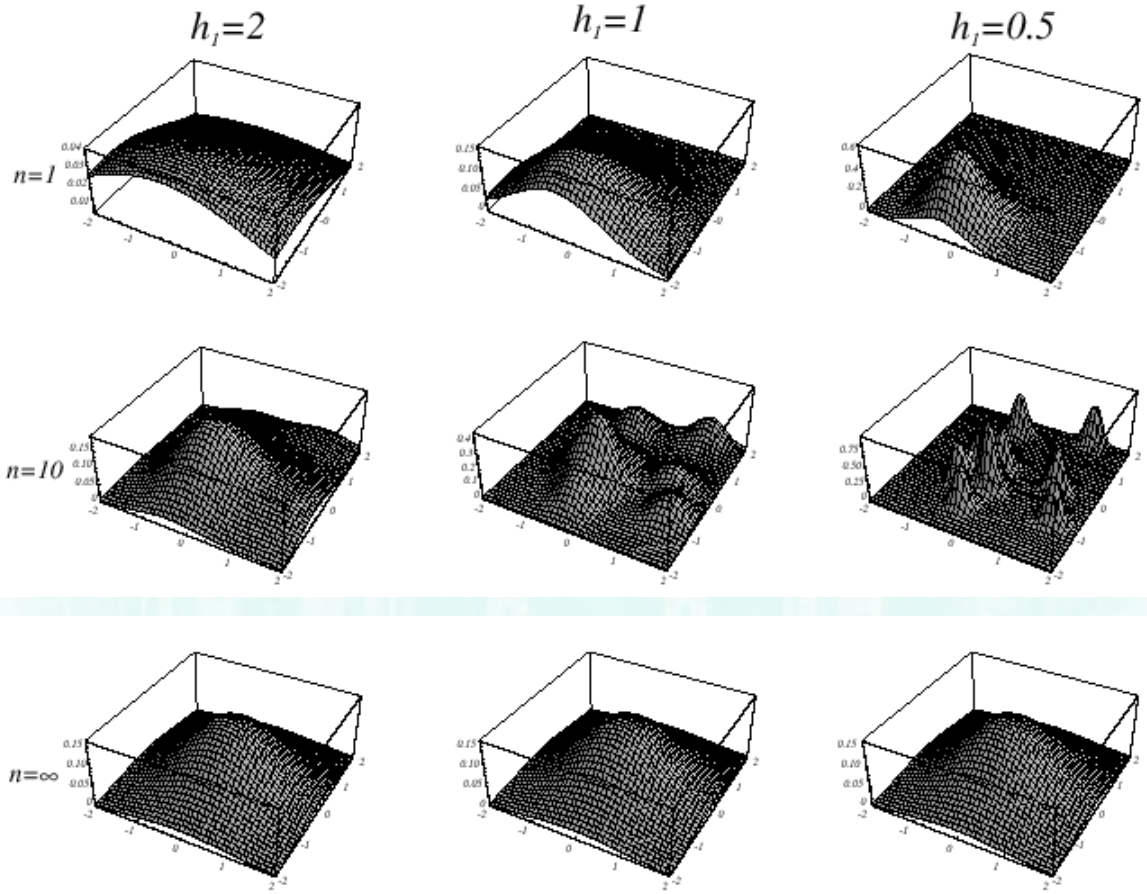
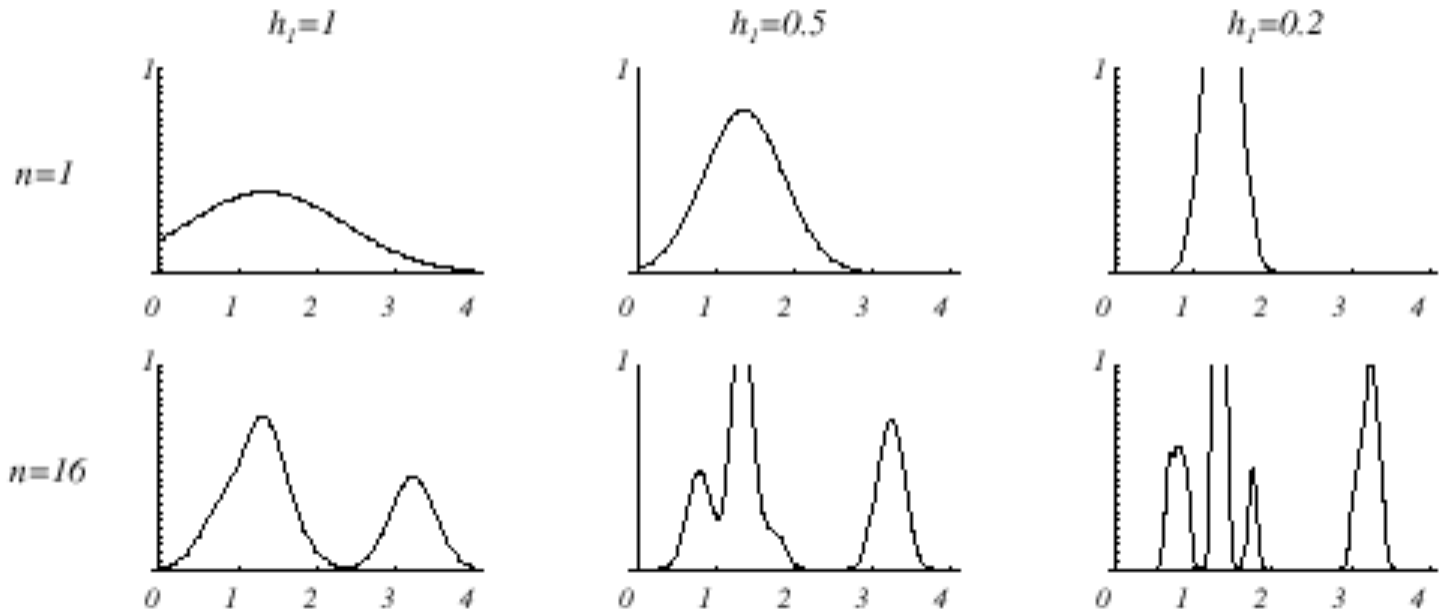


FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——Parzen窗法

- Parzen窗法应用示例二：$p(x)=\lambda_1 U(a,b)+\lambda_2 T(c,d)$，其中$U(a,b)$为$(a,b)$之间的均匀分布，$T(c,d)$为$(c,d)$之间的三角形分布。



非参数方法——Parzen窗法

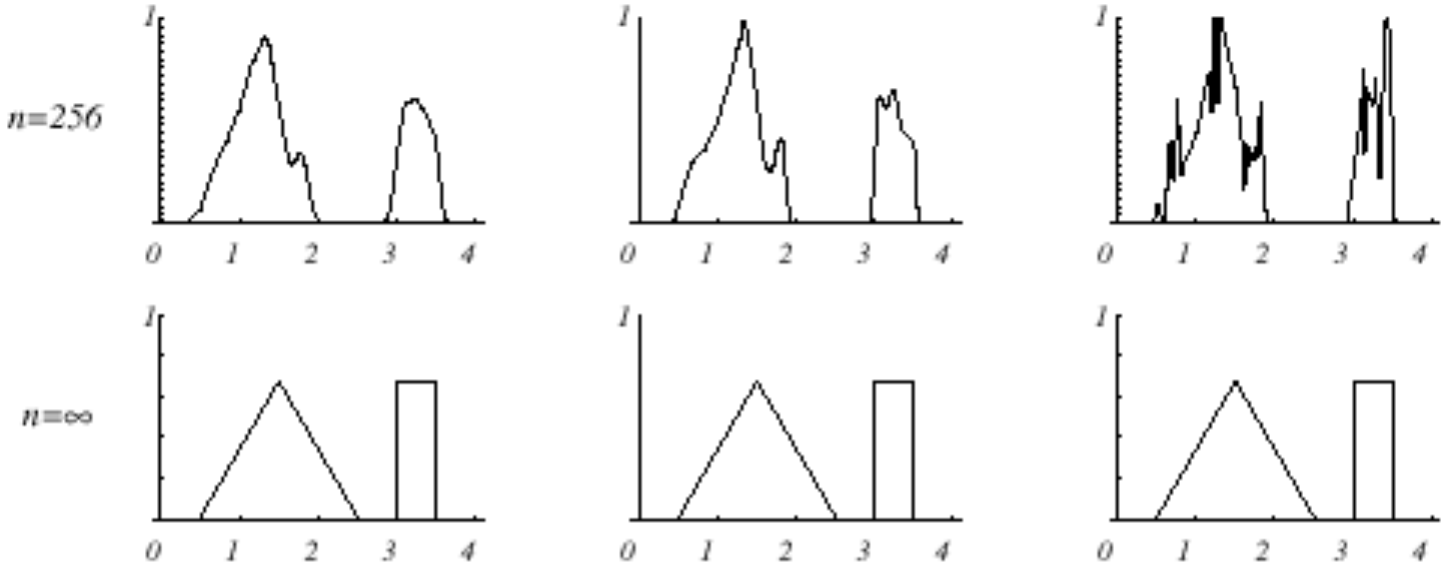


FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——Parzen窗法

- 在Parzen窗法中，对于待分类的特征向量x，我们先估计各类在该点处的概率密度，再利用最大后验概率确定其类别。
- 该方法生成的判别域受到窗函数选择的影响。

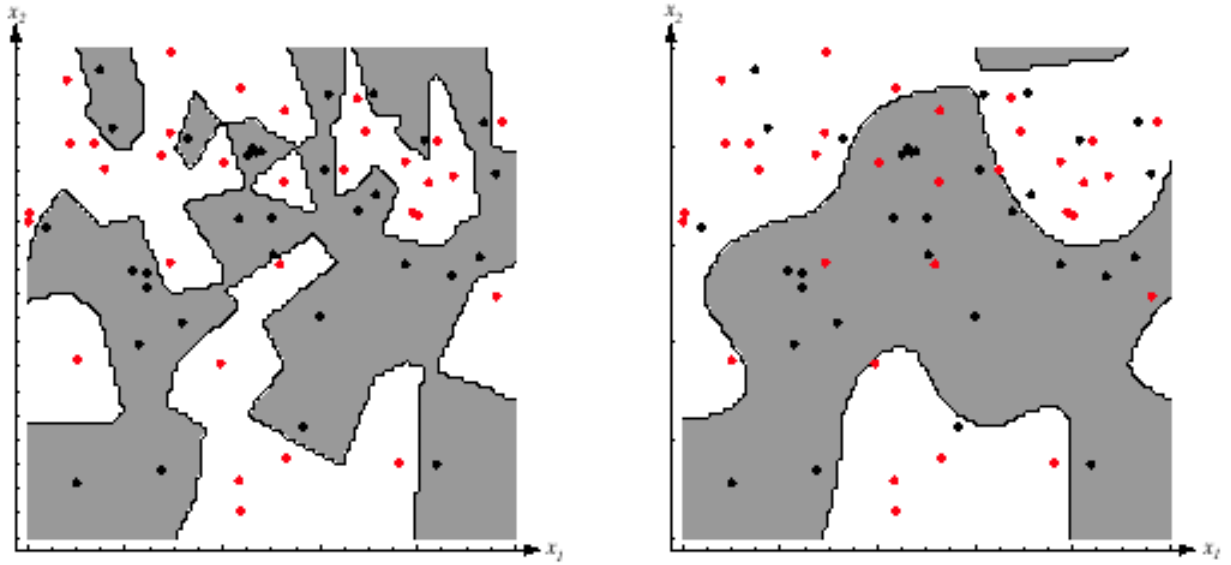


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width *h*. At the left a small *h* leads to boundaries that are more complicated than for large *h* on same data set, shown at the right. Apparently, for these data a small *h* would be appropriate for the upper region, while a large *h* would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——K_n近邻估计法

- 解决未知的“最佳”窗口函数问题：
  - 将V_n设置为训练样本的函数。具体而言，以**x**为中心扩张R_n，直到捕获k_n个样本，其中k_n是n的特定函数；
  - 这些样本被称为点**x**的k_n近邻。

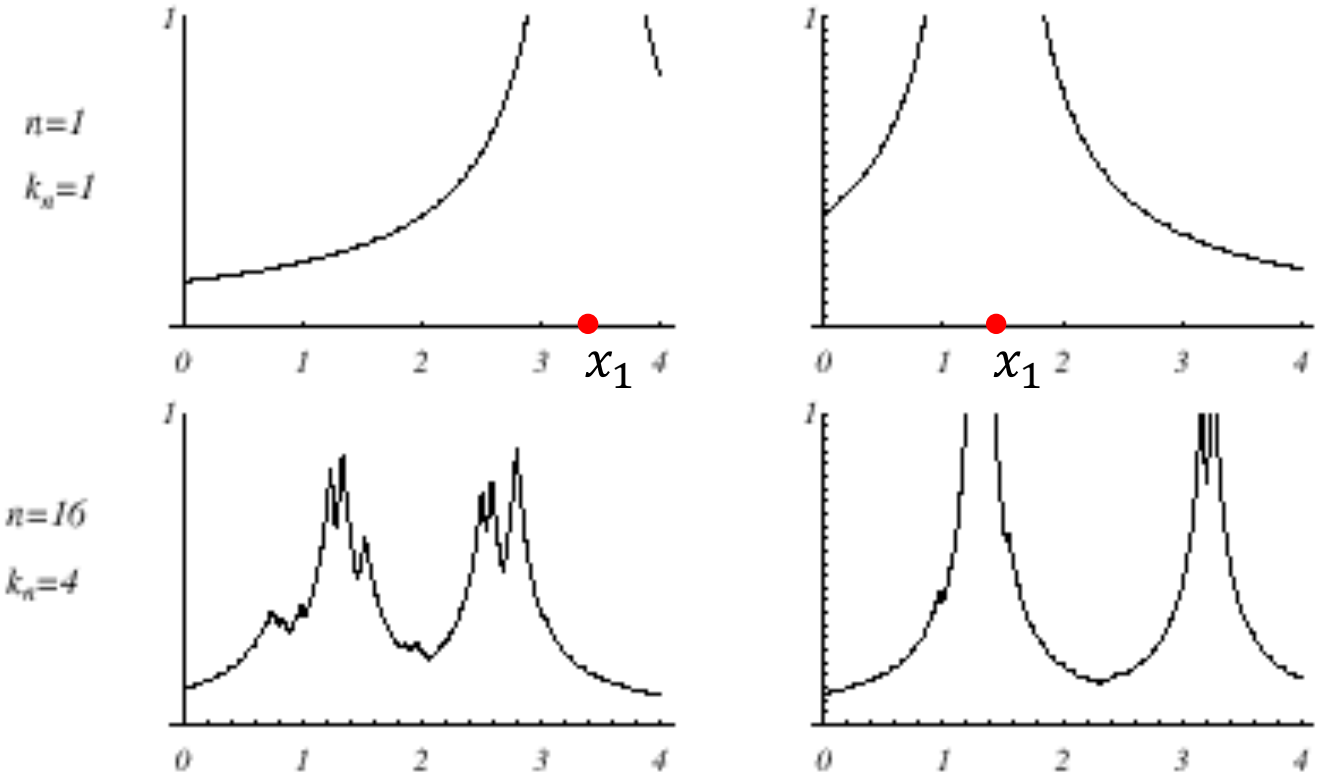
- 考虑两种可能的情况：
  - **x**附近密度高，则V_n小，分辨率高
  - **x**附近密度低，则V_n大，具有平滑噪声的效果

- 我们可以通过设置$k_n = k_1\sqrt{n}$并选择不同的k_1取值而得到一系列的估计结果。

非参数方法——K_n近邻估计法

- 示例：以一维变量为例，当$n = 1$而$k_n = \sqrt{n}$ =1时，密度估计为：

$$p_n(x) = \frac{k_n/n}{V_n} = \frac{1}{V_1} = \frac{1}{2|x - x_1|}$$



非参数方法——K_n近邻估计法

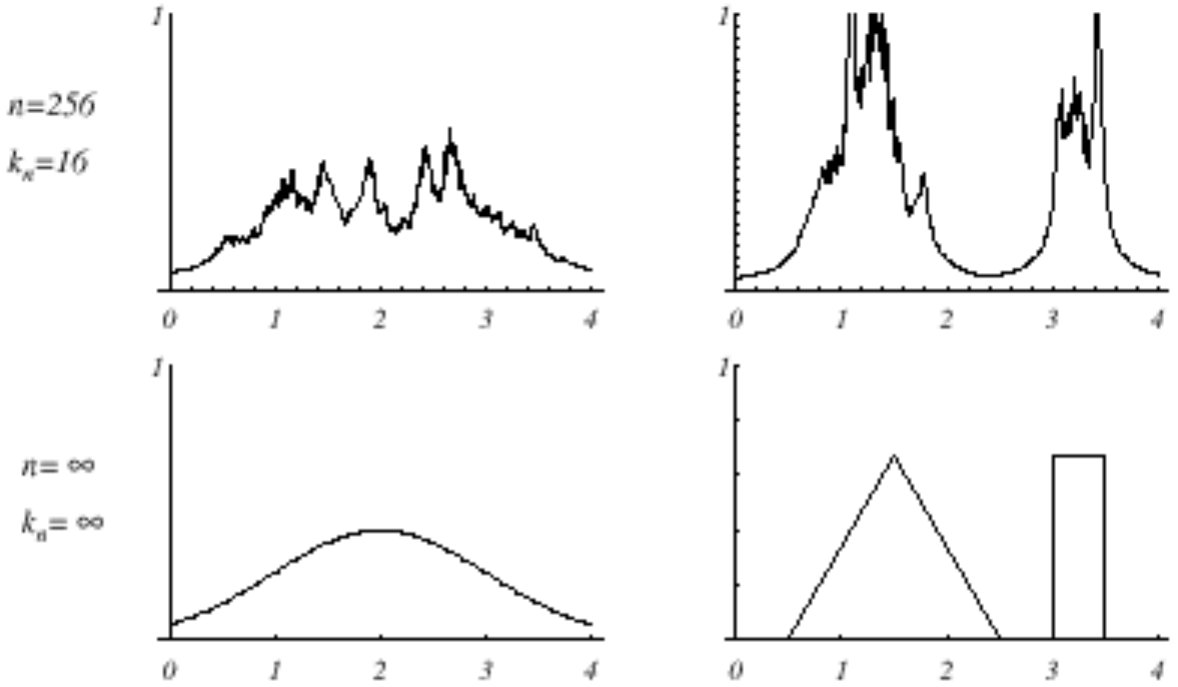


FIGURE 4.12. Several *k*-nearest-neighbor estimates of two unidimensional densities: a Gaussian and a bimodal distribution. Notice how the finite *n* estimates can be quite "spiky." From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——K_n近邻估计法

- 后验概率的估计：从n个带类标样本中估计$P(\omega_i|\mathbf{x})$
  - 考虑一个以$\mathbf{x}$为中心、体积为V的区域，其中包含k个样本。这些样本中有k_i个类别为ω_i。则近似有：

$$p_n(\mathbf{x},\omega_i)=\frac{k_i/n}{V}$$

  对后验概率的估计则为：

$$p_n(\omega_i|\mathbf{x})=\frac{p_n(\mathbf{x},\omega_i)}{\sum_{j=1}^{c}p_n(\mathbf{x},\omega_j)}=\frac{k_i}{k}$$

- 如果k大而V小，则分类性能将接近最佳。

非参数方法——K_n近邻估计法

- 最近邻分类：
  设$D_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$为一组已标记类别的训练样本。如果其中与待识别的测试样本$\mathbf{x}$最近的训练样本为$\mathbf{x}' \in D_n$，则将$\mathbf{x}'$的类标赋给$\mathbf{x}$。

- 最近邻规则导致的错误率大于最小错误率（即贝叶斯错误率）。

- 在有无限训练样本的情况下，最近邻分类的错误率不会超过贝叶斯分类错误率的两倍。

非参数方法——K_n近邻估计法

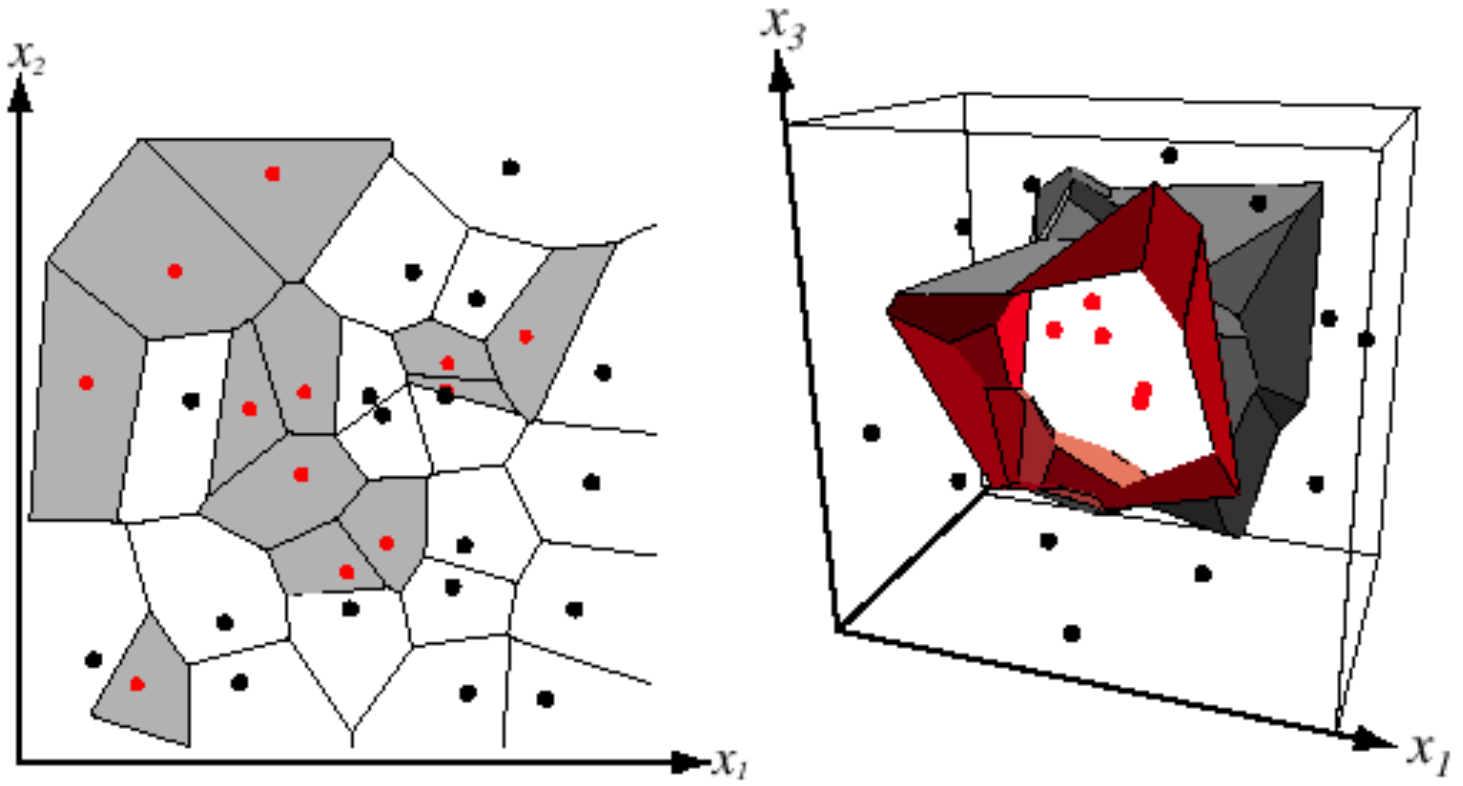
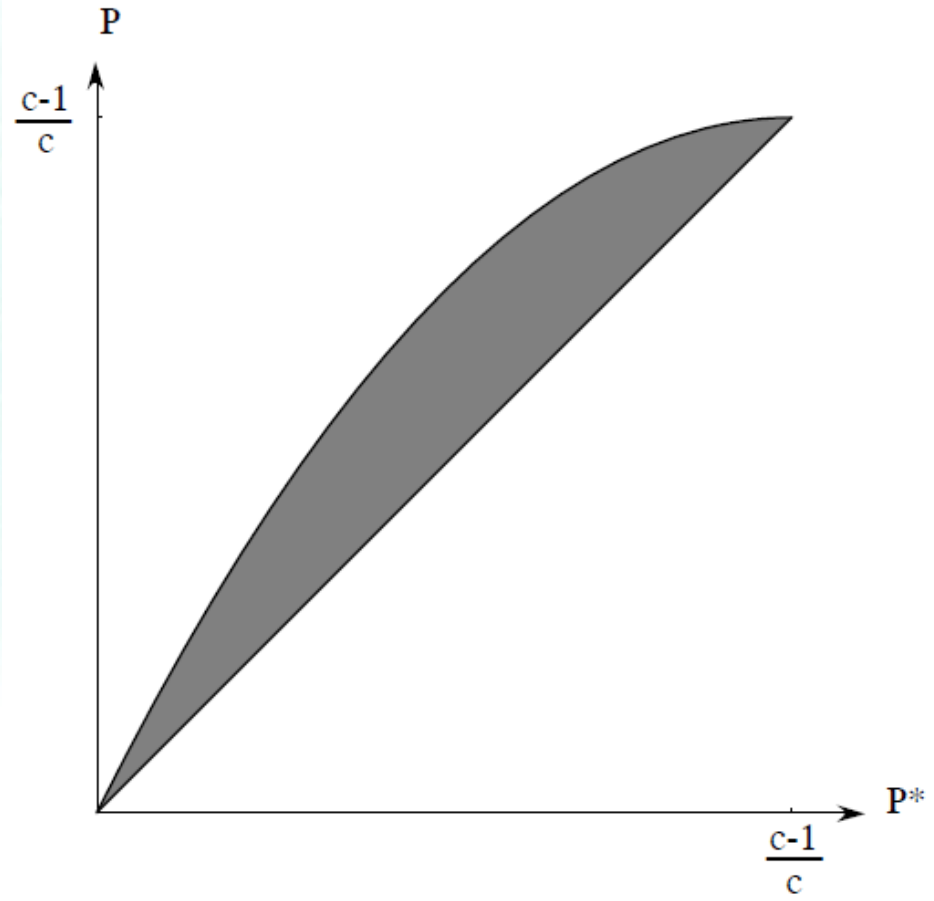


FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——K_n近邻估计法

- 最近邻分类的错误率P：设P^*为贝叶斯错误率，则

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right)$$



非参数方法——K_n近邻估计法

- k-近邻分类规则：将测试样本$\mathbf{x}$分类为与它最接近的k个近邻中出现最多的类别。

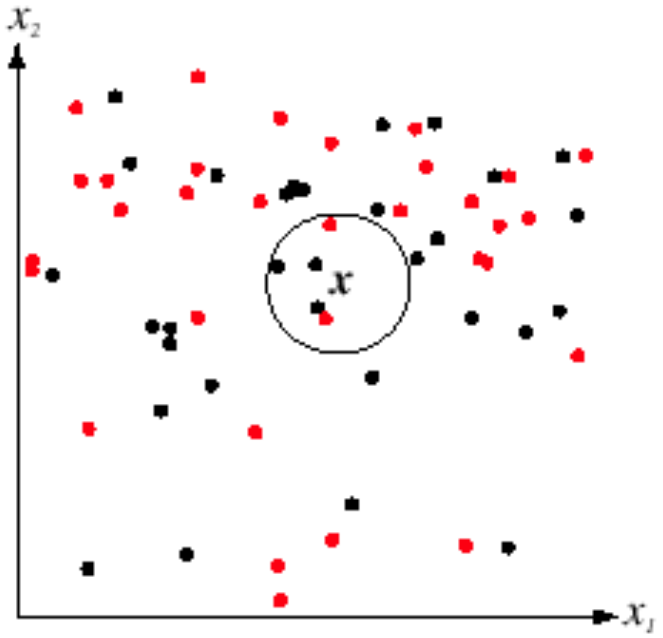
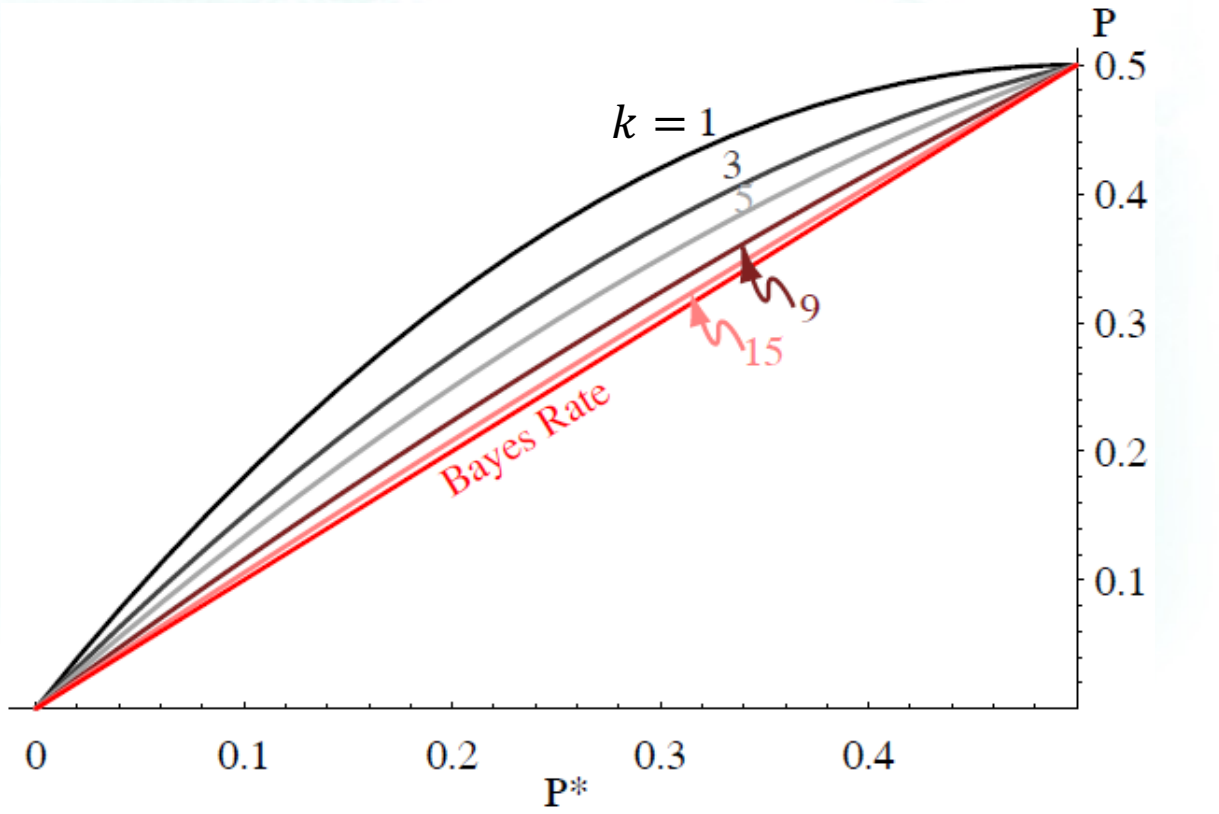


FIGURE 4.15. The k-nearest-neighbor query starts at the test point $\mathbf{x}$ and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point $\mathbf{x}$ would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

非参数方法——K_n近邻估计法

- k-近邻分类的错误率界：当k增加时，错误率上界将逐渐逼近下界——贝叶斯错误率。当 k 趋于无穷大时，上下界重合，此时k-近邻分类规则就成为最优分类规则。



第四部分：因果发现与推断

因果发现与推断

- Better to talk of (in)dependence other than correlation.
- Most statisticians would agree that causality does tell us something about dependence.
- But dependence does tell us something about causality too.

Correlation Is Not Causation	The golden rule of causal analysis: no causal claim can be established purely by a statistical method.

因果发现与推断

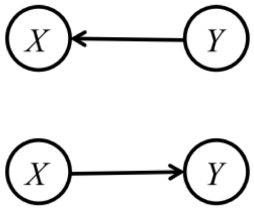
- Reichenbach's *Common Cause Principle* (1956) links causality and (in)dependence.

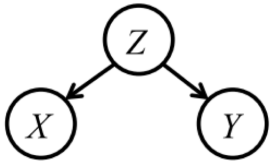
It seems that a dependence between events A and B indicates either that A causes B, or that B causes A, or that A and B have a common cause.

If A and B have a common cause C (only), then conditioning on C would make A and B independent. In this case, C is said to 'screen off' the dependence between A and B.

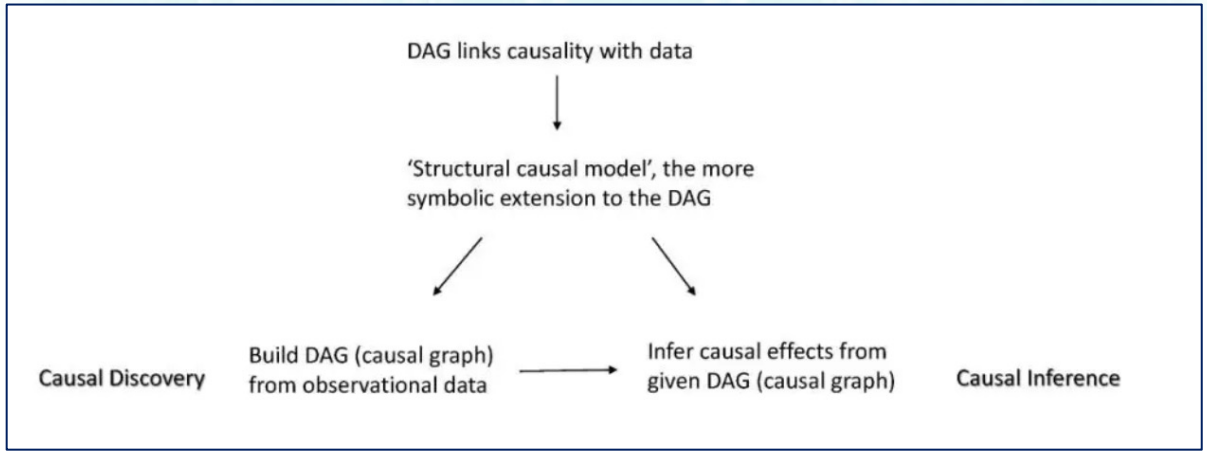
因果发现与推断

- Use the Directed Acyclic Graph (DAG) to represent the cause-effect relations
  - Nodes as variables
  - Edges as direct causal connections





- If a DAG represents the true causal relationship, then the DAG encodes all the conditional independence relations in the true distribution, which can be read off using the *d*-separation criterion.



Causal Graph

- Each causal model $\mathcal{M}$ is associated with a direct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where
  - $\mathcal{V}$ is the set of nodes represent the variables $\boldsymbol{U} \cup \boldsymbol{V}$ in $\mathcal{M}$;
  - $\mathcal{E}$ is the set of edges determined by the structural equations in $\mathcal{M}$: for X_i, there is an edge pointing from each of its parents $\boldsymbol{Pa}_i \cup \boldsymbol{U}_i$ to it.
    - Each direct edge represents the potential direct causal relationship.
    - Absence of direct edge represents zero direct causal relationship.
- Assuming the acyclicity of causality, $\mathcal{G}$ is a directed acyclic graph (DAG).
- Standard terminology
  - parent, child, ancestor, descendent, path, direct path

因果发现与推断

A Causal Model and Its Graph

Observed Variables $\boldsymbol{V} = \{I, H, W, E\}$

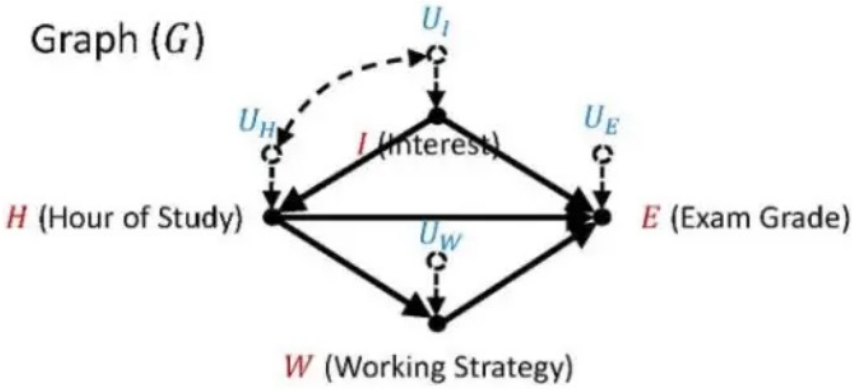
Hidden Variables $\boldsymbol{U} = \{U_I, U_H, U_W, U_E\}$

Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

Assume U_I and U_H are correlated.

Graph (G)



因果发现与推断

A Markovian Model and Its Graph

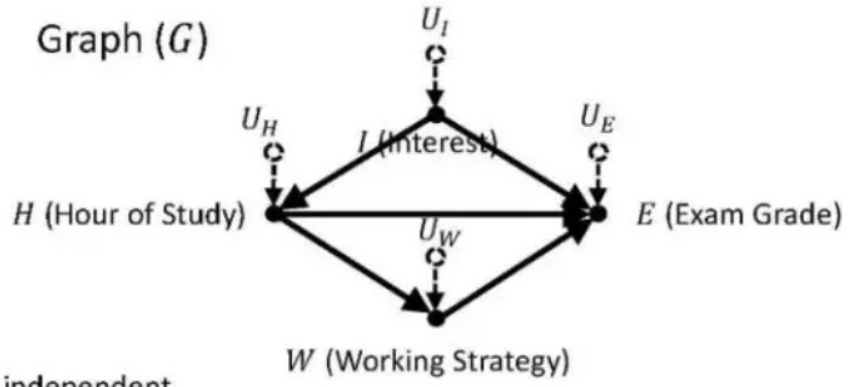
With causal sufficiency assumption

Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

Assume U_I, U_H, U_W, U_E are mutually independent.

Graph (G)



因果充分性是假设在考虑的任何一对变量中都没有未测量的共同原因(没有潜在的混杂因素)。