第三讲 统计决策方法

周文晖

杭州电子科技大学

- **贝叶斯决策理论**
  基本概念，决策准则，概率论基础，…
- **最小错误/最小风险贝叶斯决策**
  最小错误率决策规则、条件风险，期望风险， …
- **概率密度函数估计**
  基本概念，最大似然估计…
- **贝叶斯估计和贝叶斯学习**
  贝叶斯估计，贝叶斯学习 …
- **非参数估计方法**
  基本概念，Parzen窗法 …
- **贝叶斯分类器**
  贝叶斯网络，朴素贝叶斯分类器 …

贝叶斯决策理论

基本概念，决策准则，概率论基础，…

最小错误/最小风险贝叶斯决策

最小错误率决策规则、条件风险，期望风险， …

概率密度函数估计

基本概念，最大似然估计…

贝叶斯估计和贝叶斯学习

贝叶斯估计，贝叶斯学习 …

非参数估计方法

基本概念，Parzen窗法 …

贝叶斯分类器

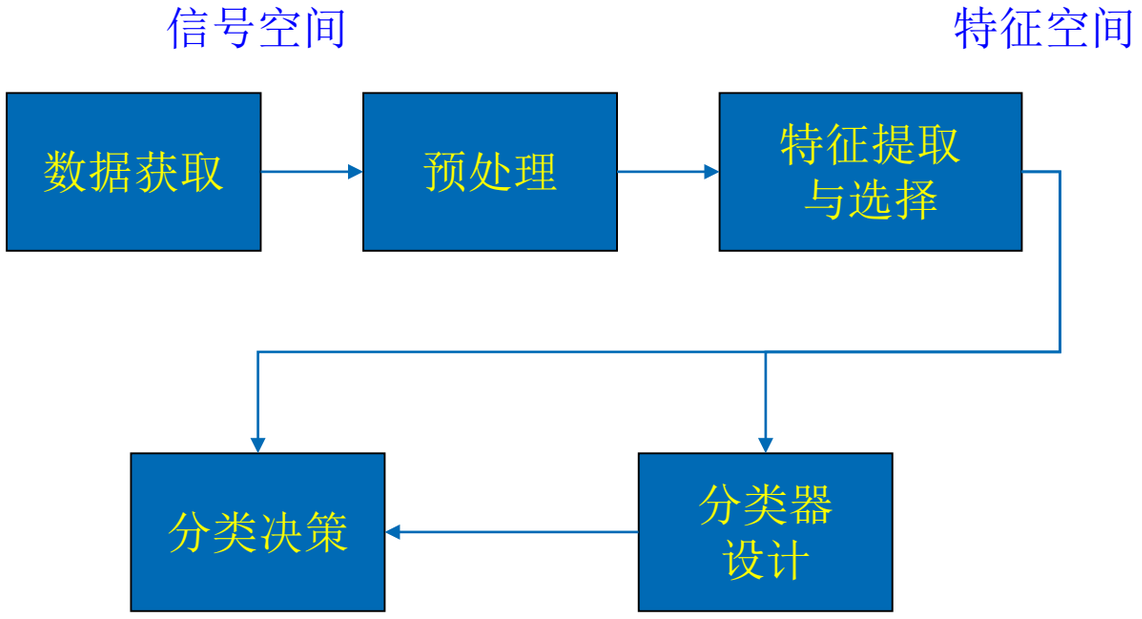
贝叶斯网络，朴素贝叶斯分类器 …

分类决策的系统组成

获取观察数据时，有二种情况：

* **确定性事件**：事物间有确定的因果关系。

* **随机事件**：事物间没有确定的因果关系，观察到的特征具有统计特性，是一个随机向量。只能利用样本集的统计特性进行分类，使分类器发生分类错误的概率最小。



基本概念

模式分类：根据识别对象的观测值确定其类别。

样本与**样本空间**：待研究对象个体用特征向量表示，所有可能的取值范围构成特征空间。

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T \quad \mathbf{x} \in R^n$$

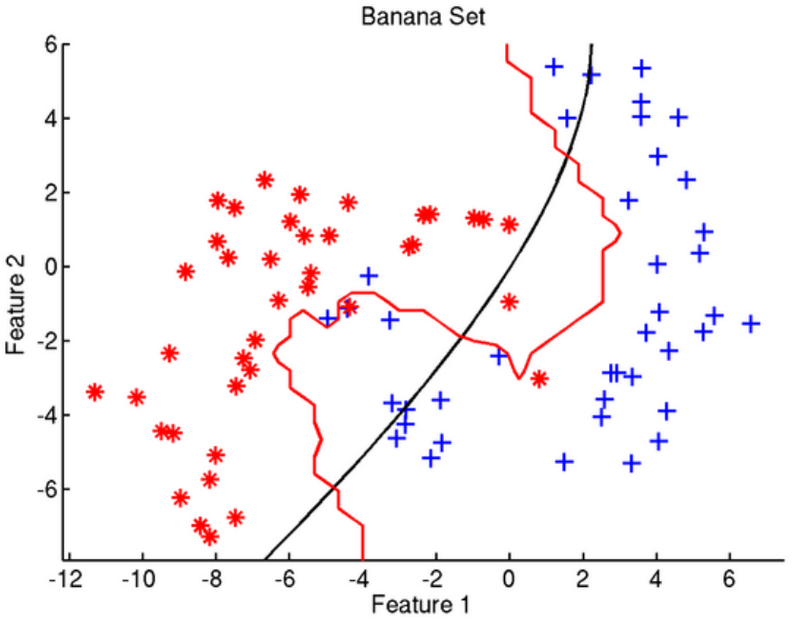
类别与**类别空间**：类别数已知，共c个类别，所有类别状态构成类别空间。

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_i, \ldots, \omega_c\}$$

把x分到哪一类**最合理**？理论基础之一是**统计决策理论**。

决策：是从样本空间S，到决策空间Θ的一个映射，表示为：

D: S → Θ



决策准则

评价决策有多种标准，对于同一个问题，采用不同的标准会得到不同意义下“最优”的决策。

统计模式识别：用概率统计的观点和方法来解决模式识别问题

Bayes决策常用的准则：

- 最小错误率准则
- 最小风险准则
- 在限定一类错误率条件下使另一类错误率为最小的准则
- 最小最大决策准则

概率论基础回顾

概率论是研究随机现象中数量规律的科学。

所谓随机现象是指在相同的条件下重复进行某种实验时，所得实验结果不一定完全相同且不可预知的现象。

本节简单回顾概率论的基本概念和贝叶斯定理。

随机事件

随机实验：

- 可观察结果的人工或自然的过程，其产生的结果可能不止一个，且不能事先确定会产生什么结果。

样本空间：

- 随机实验的全部可能出现的结果集合，通常记作Ω，Ω中的点（即一个可能出现的实验结果）成为样本点，通常记作ω。

随机事件：

- 随机实验的一些可能结果的集合，是样本空间的一个子集。常用大写字母 $A, B, C,\ldots$表示。

事件间的关系和运算

A，B，　A_1，A_2，…A_n为一些事件，它们有下述的运算：

交：

- 记 $C=$“A与B同时发生”，称为事件A与B的交。
- $C=\{\omega \mid \omega\in A$ 且 $\omega\in B\}$，记作 $C=A\cap B$ 或 $C=AB$。
- 类似地用 $\cap A_i=A_1\cap A_2\cap\ldots\cap A_n$ 表示事件“n个事件 $A_1, A_2, \ldots A_n$ 同时发生”。

事件间的关系和运算2

并：

- C = “A与B中至少有一个发生”，称为事件 A 与B的并。
- $C=\{\omega \mid \omega \in A$ 或 $\omega \in B\}$，记作 $C=A \cup B$。
- 类似地用 $\cup A_i=A_1 \cup A_2 \cup \ldots \cup A_n$表示事件 “$n$个事件 $A_1, A_2, \ldots A_n$ 中至少有一个发生”。

差：

- C = “A 发生而 B 不发生”，称为事件A与B的差。
- $C=\{\omega \mid \omega \in A$ 但 $\omega \notin B\}$，记作$C=A-B$。

运算性质

事件的运算有以下几种性质：

- 交换率：$A \cup B = B \cup A \qquad AB = BA$
- 结合律：$(A \cup B) \cup C = A \cup (B \cup C)$

  $(AB)C = A(BC)$
- 分配律：$(A \cup B)C = (AC) \cup (BC)$

  $(AB) \cup C = (A \cup C)(B \cup C)$
- 摩根率：$\sim (\bigcap_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} \sim A_i \qquad \sim (\bigcup_{i=1}^{n} A_i) = \bigcap_{i=1}^{n} \sim A_i$

概率定义

定义：设Ω为一个随机实验的样本空间，对Ω上的任意事件A，规定一个实数与之对应，记为$P(A)$，满足以下三条基本性质，称为事件A发生的概率：

$$0 \le P(A) \le 1$$

$$P(\Omega) = 1 \qquad P(\phi) = 0$$

若两事件A、B互斥，即，则

$$P(A \cup B) = P(A) + P(B)$$

对于两两互斥的事件A_1，A_2，…有

$$P\left(A_1 + A_2 + \cdots\right) = P\left(A_1\right) + P\left(A_2\right) + \cdots$$

完备事件族/基本事件族

定义：设$\{A_n, n=1, 2, \ldots\}$为一组有限或可列无穷多个事件，两两不相交，且$\sum_n A_n = \Omega$，则称事件族$\{A_n, n=1, 2, \ldots\}$为样本空间Ω的一个**完备事件族**。若对任意事件 B 有 $B \cap A_n = A_n$ 或 φ，$n=1, 2, \ldots$，则称$\{A_n, n=1, 2, \ldots\}$为**基本事件族**。

完备事件族与基本事件族有如下的性质：

定理：若$\{A_n, n=1, 2, \ldots\}$为一完备事件族，则$\sum_n P(A_n) = 1$，且对于一事件B有

$$P(B) = \sum_n P(A_n B)$$

又若$\{A_n, n=1, 2, \ldots\}$为基本事件族，则 $P(B) = \sum_{A_n \subset B} P(A_n)$

统计概率性质

对任意事件A，有 $0 \le P(A) \le 1$

必然事件Ω的概率 $P(\Omega)=1$，不可能事件φ的概率 $P(\varphi)=0$

对任意事件A，有 $P(A)=1-P(\sim A)$.

设事件$A_1, A_2, \ldots A_n$（$k \le n$）是两两互不相容的事件，即有$A_i \cap A_j = \varphi$，则

$$P(\bigcup_{i=1}^{k} A_i) = P(A_1) + P(A_2) + \ldots + P(A_k)$$

设A，B是两事件，则

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

条件概率（乘法公式）

定义：设A，B为事件且$P(A)>0$，称

$$P(B\mid A)=\frac{P(AB)}{P(A)}$$

为事件A已发生的条件下，事件B的**条件概率**，$P(A)$在概率推理中称为**边缘概率**。

简称$P(B|A)$为给定A时B发生的概率。$P(AB)$称为A与B的联合概率，A, B同时发生的概率 。

有联合概率公式：

$$P(AB)=P(B\mid A)P(A)=P(A\mid B)P(B)=P(BA)$$

条件概率性质

$$0 \le P(B \mid A) \le 1$$

$$P(\Omega \mid A) = 1 \qquad P(\phi \mid A) = 0$$

若 $B_1 B_2 = \phi$，则 $P(B_1 + B_2 \mid A) = P(B_1 \mid A) + P(B_2 \mid A)$

乘法公式： $P(AB) = P(A)P(B \mid A)$

$$P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 A_2)\ldots P(A_n \mid A_1 A_2 \ldots A_{n-1})$$

事件的独立性

独立：如果X与Y相互独立，则

$$P(X,Y) = P(X)P(Y)$$

$$P(X|Y) = P(X)$$

如感冒和吸烟都会导致支气管炎，感冒和吸烟两事件相互独立。

条件独立：如果在给定Z的条件下，X与Y相互独立，则

$$P(X|Y, Z) = P(X|Z)$$

实际中，条件独立比完全独立更重要

联合概率

联合概率：$P(X_1, X_2, \ldots, X_N)$

- 如果相互独立：$P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2) \ldots P(X_N)$

写成条件概率：$P(X_1, X_2, \ldots, X_N) = P(X_1 | X_2, \ldots, X_N)\, P(X_2, \ldots, X_N)$

迭代表示：$P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2 | X_1)\, P(X_3 | X_2 X_1) \ldots P(X_N | X_{N-1}, \ldots, X_1)$

$$= P(X_N)\, P(X_{N-1} | X_N)\, P(X_{N-2} | X_{N-1} X_N) \ldots P(X_1 | X_2, \ldots, X_N)$$

实际应用中就是利用条件独立性的性质简化网络复杂性的。

全概率公式

设A_1，A_2，…A_n互不相交，$\sum_i A_i = \Omega$ (完备事件集)，且 $P(A_i) > 0, \quad i = 1, 2, ..., n$

则对于任意事件A有

$$P(A) = \sum_i P(A_i)P(A \mid A_i)$$

如何利用前面所学性质证明？

$$A = A\Omega = A(A_1 \cup A_2 \cup ... \cup A_n) = AA_1 \cup ... \cup AA_n$$

$$\begin{aligned} P(A) &= P(AA_1) + ... + P(AA_2) \\ &= P(A \mid A_1)P(A_1) + ... + P(A \mid A_n)P(A_n) \end{aligned}$$

贝叶斯定理

设A，$B_1, B_2, \ldots, B_n$为一些事件，$P(A) > 0$，$B_1, B_2, \ldots, B_n$互不相交，$P(B_i) > 0$，$i=1, \ldots, n$，且 $\sum_i P(B_i) = 1$，则对于 $k = 1, 2, \ldots, n$，

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_i P(B_i)P(A \mid B_i)}$$

如何利用前面所学性质证明？

$$P(B_k \mid A) = \frac{P(AB_k)}{P(A)} = \frac{P(B_k)P(A \mid B_k)}{\sum_i P(B_i)P(A \mid B_i)}$$

贝叶斯定理说明

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_i P(B_i)P(A \mid B_i)}$$

贝叶斯公式容易由条件概率的定义，乘法公式和全概率公式得到。

在贝叶斯公式中，

- $P(B_i)$, i=1, 2, …, n 称为**先验概率**，是根据以往经验和分析得到的概率
- $P(B_i|A)$, i=1, 2, …, n 称为**后验概率**也是**条件概率**。

B_1，B_2，…，B_n 代表n个互不相容的原因，A 代表某种结果。贝叶斯公式就是已知结果反推该结果是由哪种原因产生的，如 P(感冒|头痛) 。

条件概率的三个重要公式

（1）概率乘法公式：

$$P(AB) = P(A) \cdot P(B \mid A)$$

（2）全概率公式：设事件A_1，A_2，…，A_n两两互斥，且 $\sum_{i=1}^{n} A_i = \Omega$，$P(A_i) > 0$，$i = 1,2,...,n$

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B \mid A_i)$$

（3）贝叶斯公式：

$$P(A_i \mid B) = \frac{P(A_i) P(B \mid A_i)}{\sum_{i=1}^{n} P(A_i) P(B \mid A_i)}$$

贝叶斯理论中的三个概率

$$P(\omega_i \mid X) = \frac{P(\omega_i)P(X \mid \omega_i)}{P(X)}$$

设随机样本向量$\boldsymbol{X}$，相关的三个概率：

（1）先验概率$P(\omega_i)$：根据以前的知识和经验得出的ω_i类样本出现的概率，与现在无关。

（2）后验概率$P(\omega_i|\boldsymbol{X})$：相对于先验概率而言。指收到数据$\boldsymbol{X}$（一批样本）后，根据这批样本提供的信息统计出的$\omega_i$类出现的概率。表示$\boldsymbol{X}$属于$\omega_i$类的概率。

（3）类条件概率$P(\mathrm{X} \mid \omega_i)$：或称$\omega_i$的似然函数，已知属于$\omega_i$类的样本$\boldsymbol{X}$，发生某种事件的概率。例对一批得病患者进行一项化验，结果为阳性的概率为95%，ω_1代表得病人群，则$\boldsymbol{X}$化验为阳性的事件可表示为：

$$P(\boldsymbol{X} = \text{阳} \mid \omega_1) = 0.95$$

贝叶斯理论中的三个概率

分类问题用哪种概率合适？

例如：一个2类问题，ω_1诊断为患有某病，ω_2诊断为无病，

则：$P(\omega_1)$表示某地区的人患有此病的概率
$P(\omega_2)$表示该地区人无此病的概率
} 通过统计资料得到

若用某种方法检测是否患有某病，假设 $\boldsymbol{X}$ 表示“试验反应呈阳性”。则：

$P(\boldsymbol{X}|\omega_2)$ 表示无病的人群做该试验时反应呈阳性 (显示有病)的概率。

$P(\omega_2|\boldsymbol{X})$ 表示试验呈阳性的人中，实际没有病的人的概率。

值低 / 高 ？ √

$P(\boldsymbol{X}|\omega_1)$ 表示患病人群做该试验时反应呈阳性的概率。

$P(\omega_1|\boldsymbol{X})$ 表示试验呈阳性的人中，实际确实有病的人的概率。

值低 / 高 ？ √

- **贝叶斯决策理论**
  基本概念，决策准则，概率论基础，…
- **最小错误/最小风险贝叶斯决策**
  最小错误率决策规则、条件风险，期望风险， …
- **概率密度函数估计**
  基本概念，最大似然估计…
- **贝叶斯估计和贝叶斯学习**
  贝叶斯估计，贝叶斯学习 …
- **非参数估计方法**
  基本概念，Parzen窗法 …
- **贝叶斯分类器**
  贝叶斯网络，朴素贝叶斯分类器 …

最小错误率贝叶斯决策

1. 问题分析

目的：确定 **X** 属于哪一类，看 **X** 来自哪类的概率大。在下列三种概率中：

先验概率 $P(\omega_i)$

类条件概率 $P(\mathbf{X}\,|\omega_i)$

后验概率 $P(\omega_i\,|\,\mathbf{X})$

默认 *P* 表示概率，
*p*为概率密度函数

采用哪种概率进行分类最合理？ **后验概率$P(\omega_i\,|\,\mathbf{X})$**

2. 决策规则

设有M类模式，

若 $P(\omega_i\,|\,\mathbf{X})=\max\left\{P(\omega_j\,|\,\mathbf{X})\right\},\ j=1,2,\ldots,M$ 则 $\mathbf{X}\in\omega_i$类　　记作：$i=\underset{j}{\operatorname{argmax}}\,P(\omega_j\,|\,\mathbf{X})$

最小错误率贝叶斯决策

为什么称为最小错误率？

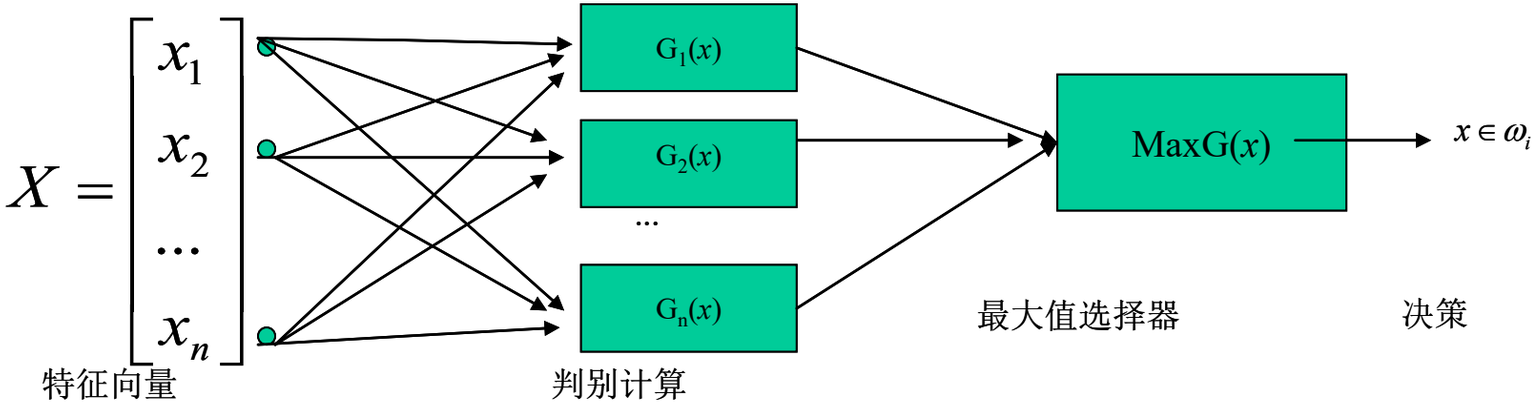
这一决策的错误率是最小的

根据贝叶斯公式，计算 $\boldsymbol{X}$ 属于每一类的后验概率

$$G_i(\mathbf{X}) = P(\omega_i|\mathbf{X}) = \frac{P(\mathbf{X}|\omega_i)P(\omega_i)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|\omega_i)P(\omega_i)}{\sum_{i=1}^{M} P(\mathbf{X}|\omega_i)P(\omega_i)}$$

M：类别数

决策面上进行分类器设计



最小错误率贝叶斯决策证明

对于两类情况：

- 样本(sample) $\mathbf{X} \in R^d$
- 状态(state) 第一类：ω_1， 第二类：ω_2
- 先验概率 (a priori probability or prior) $P(\omega_1)$，$P(\omega_2)$
- 样本分布密度(sample distribution density) $p(\mathbf{X})$（总体概率密度）
- 类条件概率密度(class-conditional probability density)：$p(\mathbf{X}|\omega_1)$，$p(\mathbf{X}|\omega_2)$
- 后验概率(a posteriori probability or posterior)： $P(\omega_1|\mathbf{X})$，$P(\omega_2|\mathbf{X})$

默认 ***P*** 表示概率，
p为概率密度函数

最小错误率贝叶斯决策证明

决策规则

$$\begin{cases} P(\omega_1 \mid \mathbf{X}) > P(\omega_2 \mid \mathbf{X}) & \text{if } \mathbf{X} \text{ is assigned to } \omega_1 \\ P(\omega_1 \mid \mathbf{X}) < P(\omega_2 \mid \mathbf{X}) & \text{if } \mathbf{X} \text{ is assigned to } \omega_2 \end{cases}$$

■ 错误概率 (probability of error) :

$$P(e \mid \mathbf{X}) = \begin{cases} P(\omega_2 \mid \mathbf{X}) = 1 - P(\omega_1 \mid \mathbf{X}) & \text{if } \mathbf{X} \text{ is assigned to } \omega_1 \\ P(\omega_1 \mid \mathbf{X}) = 1 - P(\omega_2 \mid \mathbf{X}) & \text{if } \mathbf{X} \text{ is assigned to } \omega_2 \end{cases}$$

■ 平均错误率 (average probability of error):

$$P(e) = \int P(e \mid \mathbf{X}) p(\mathbf{X}) d\mathbf{X}$$

（平均）错误率是条件错误率的数学期望

要求平均错误率最小

最小错误率贝叶斯决策证明

$$P(e|\mathbf{X})=\begin{cases}P(\omega_2|\mathbf{X}) & \text{if } \mathbf{X} \text{ is assigned to } \omega_1\\ P(\omega_1|\mathbf{X}) & \text{if } \mathbf{X} \text{ is assigned to } \omega_2\end{cases}$$

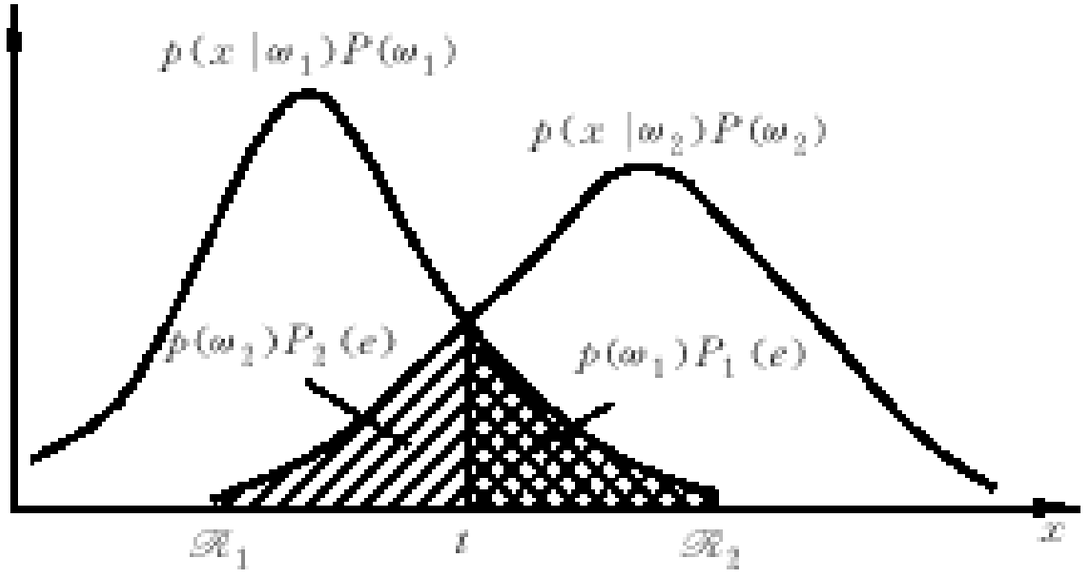
$$P(e)=\int P(e|\mathbf{X})p(\mathbf{X})d\mathbf{X}$$

$$=\int_{-\infty}^{t}P(\omega_2|\mathbf{X})p(\mathbf{X})d\mathbf{X}+\int_{t}^{+\infty}P(\omega_1|\mathbf{X})p(\mathbf{X})d\mathbf{X}$$

$$=\int_{-\infty}^{t}p(\mathbf{X}|\omega_2)P(\omega_2)d\mathbf{X}+\int_{t}^{+\infty}p(\mathbf{X}|\omega_1)P(\omega_1)d\mathbf{X}$$

$$=P(\omega_2)\int_{-\infty}^{t}p(\mathbf{X}|\omega_2)d\mathbf{X}+P(\omega_1)\int_{t}^{+\infty}p(\mathbf{X}|\omega_1)d\mathbf{X}$$

贝叶斯公式 $P(\omega_i|\boldsymbol{X})=\dfrac{P(\boldsymbol{X}|\omega_i)P(\omega_i)}{P(\boldsymbol{X})}$



当 t 取值交点时，$P(e)$最小

- 该决策使得**平均错误率**$P(e)$最小。

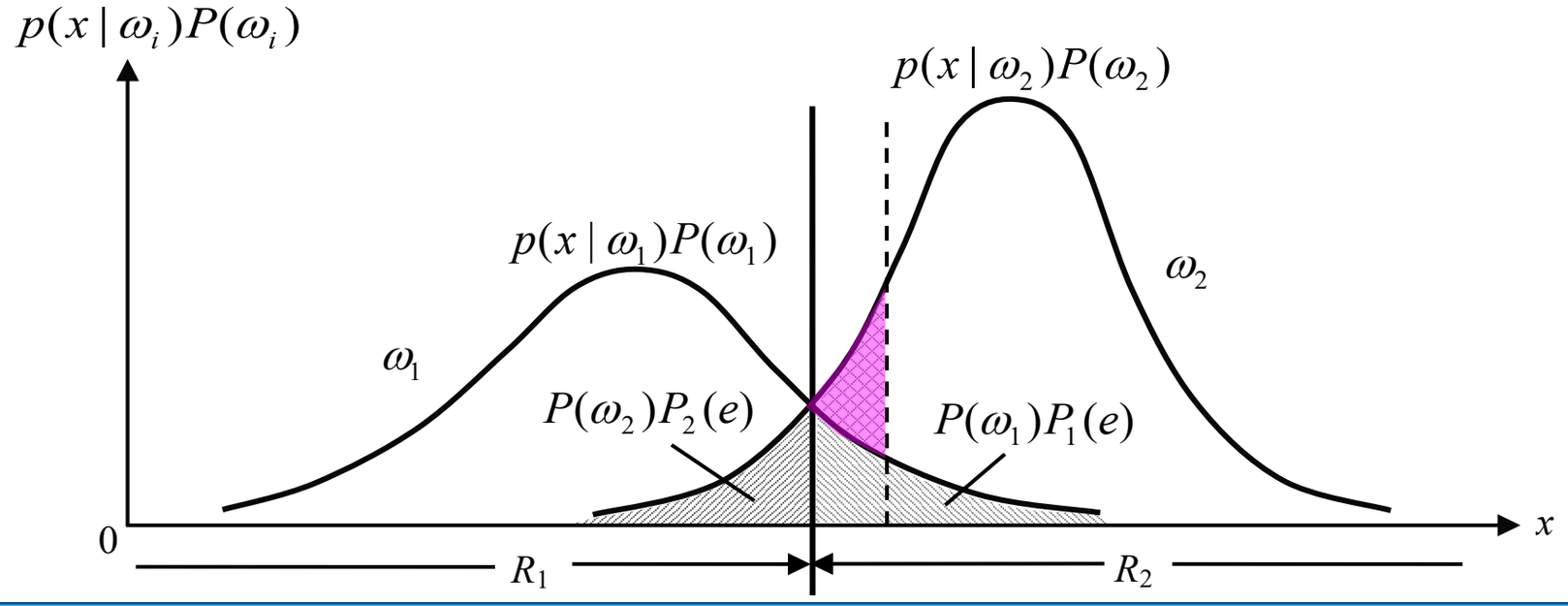
Bayes决策理论是最优的。

最小错误率贝叶斯决策证明

在**最小错误率**贝叶斯决策中，判别界面位于两曲线的交点处，即：

$$\frac{\partial P(e)}{\partial \mathbf{X}}=0 \Rightarrow P(\omega_2)p(\mathbf{X}|\omega_2)-P(\omega_1)p(\mathbf{X}|\omega_1)=0 \Rightarrow \ p\left(\boldsymbol{X}|\omega_1\right)P(\omega_1)=p\left(\boldsymbol{X}|\omega_2\right)P(\omega_2)$$

可以看出这个错误率是所有**错误率**中**最小的**（图中三角形的面积减小到0），但总错误概率不可能为零。



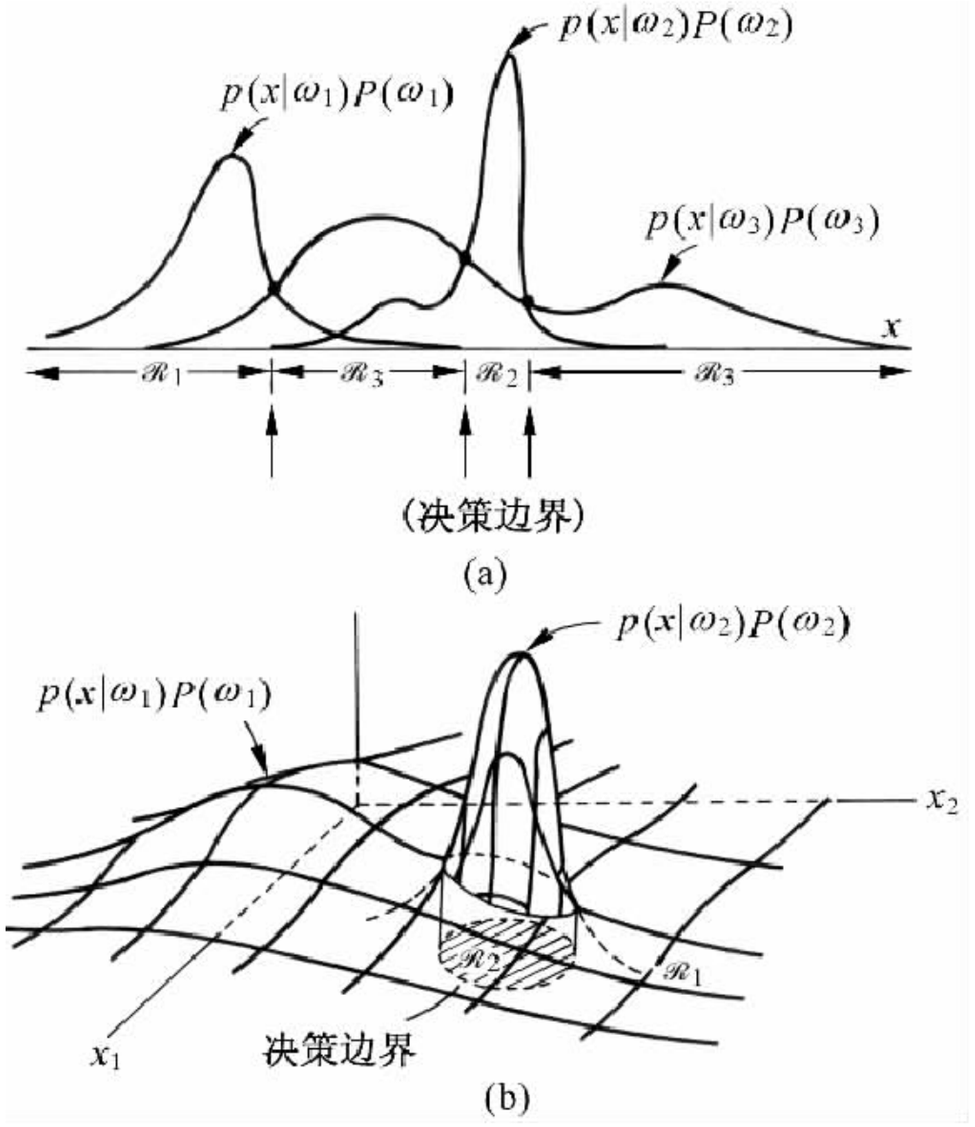
对于多分类情况：

（1） if $P(\omega_i \mid x) = \max\limits_{j=1,\ldots,c} P(\omega_j \mid x)$ then $x \in \omega_i$

（2） if $p(x \mid \omega_i)P(\omega_i) = \max\limits_{j=1,\ldots,c} p(x \mid \omega_j)P(\omega_j)$ then $x \in \omega_i$

错误率：

$$P(e) = 1 - P(c) = 1 - \sum_{j=1}^{c} \int_{\Re_j} p(x \mid \omega_j)dx$$



几种等价形式

$$P(\omega_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\omega_i)P(\omega_i)}{P(\boldsymbol{X})}$$

后验概率 $P(\omega_i|\boldsymbol{X})$ 提供有效的分类信息；

先验概率$P(\omega_i)$和类条件概率 $P(\boldsymbol{X}|\omega_i)$ 可从统计资料中获得；

$P(\boldsymbol{X})$ 可看作是与分类无关的项，可忽略；

分类规则可表示为： $i = \underset{j}{\operatorname{argmax}} P(\omega_j | X) \sim \underset{j}{\operatorname{argmax}} P(X | \omega_j)P(\omega_j)$

几种等价形式

$$\begin{cases} P(\omega_1 \mid \boldsymbol{X}) > P(\omega_2 \mid \boldsymbol{X}) \Rightarrow P(\boldsymbol{X} \mid \omega_1)P(\omega_1) > P(\boldsymbol{X} \mid \omega_2)P(\omega_2) \Rightarrow \dfrac{P(\boldsymbol{X} \mid \omega_1)}{P(\boldsymbol{X} \mid \omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)} \\ P(\omega_1 \mid \boldsymbol{X}) < P(\omega_2 \mid \boldsymbol{X}) \Rightarrow P(\boldsymbol{X} \mid \omega_1)P(\omega_1) < P(\boldsymbol{X} \mid \omega_2)P(\omega_2) \Rightarrow \dfrac{P(\boldsymbol{X} \mid \omega_1)}{P(\boldsymbol{X} \mid \omega_2)} < \dfrac{P(\omega_2)}{P(\omega_1)} \end{cases} \quad \text{则} \quad \boldsymbol{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$$l(\boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid \omega_1)}{P(\boldsymbol{X} \mid \omega_2)} \gtrless \frac{P(\omega_2)}{P(\omega_1)} \text{，则} \quad \boldsymbol{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

统计学中称 $l(\boldsymbol{X})$ 为似然比，$P(\omega_2)/P(\omega_1)$ 为似然比阈值。

取自然对数，有：

$$h(\boldsymbol{X}) = \ln l(\boldsymbol{X}) = \ln P(\boldsymbol{X} \mid \omega_1) - \ln P(\boldsymbol{X} \mid \omega_2) \gtrless \ln \frac{P(\omega_2)}{P(\omega_1)} \text{，则 } \boldsymbol{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

例 假定在细胞识别中，病变细胞的先验概率和正常细胞的先验概率分别为 $P(\omega_1)=0.05$，$P(\omega_2)=0.95$。现有一待识别细胞，其观察值为 $\boldsymbol{X}$，从类条件概率密度发布曲线上查得：$P(\mathbf{X}\mid\omega_1)=0.5$，$P(\mathbf{X}\mid\omega_2)=0.2$。试对细胞 $\boldsymbol{X}$ 进行分类。

解：[方法1] 通过后验概率计算。

$$P(\omega_1\mid\boldsymbol{X})=\frac{P(\boldsymbol{X}\mid\omega_1)P(\omega_1)}{\sum_{i=1}^{2}P(\boldsymbol{X}\mid\omega_i)P(\omega_i)}=\frac{0.5\times0.05}{0.05\times0.5+0.95\times0.2}\approx0.16$$

$$P(\omega_2\mid\boldsymbol{X})=\frac{0.2\times0.95}{0.05\times0.5+0.95\times0.2}\approx0.884$$

$$\because P(\omega_2\mid\boldsymbol{X})>P(\omega_1\mid\boldsymbol{X}) \qquad \therefore \boldsymbol{X}\in\omega_2$$

例 假定在细胞识别中，病变细胞的先验概率和正常细胞的先验概率分别为 $P(\omega_1)=0.05$，$P(\omega_2)=0.95$。现有一待识别细胞，其观察值为 $\boldsymbol{X}$，从类条件概率密度发布曲线上查得：$P(\mathbf{X}\mid\omega_1)=0.5$，$P(\mathbf{X}\mid\omega_2)=0.2$。试对细胞 $\boldsymbol{X}$ 进行分类。

解： [方法2]：利用先验概率和类概率密度计算。

$$P(\boldsymbol{X}\mid\omega_1)P(\omega_1)=0.5\times0.05=0.025$$

$$P(\boldsymbol{X}\mid\omega_2)P(\omega_2)=0.2\times0.95=0.19$$

$$\because P(\boldsymbol{X}\mid\omega_2)P(\omega_2)>P(\boldsymbol{X}\mid\omega_1)P(\omega_1)$$

$\therefore \boldsymbol{X}\in\omega_2$，是正常细胞。

最小错误率贝叶斯决策小结

- 贝叶斯决策理论要求两个前提：
  - 分类类别数目已知
  - 一个是类条件概率密度和先验概率已知

- 基于贝叶斯决策的分类器设计方法是在已知类条件概率密度的情况下讨论的，贝叶斯判别函数中的类条件概率密度是利用样本估计的。

基于最小风险的Bayes决策

思想：最小错误率贝叶斯决策只考虑了错误率，没有考虑到不同的错误带来的损失和后果严重程度是不同的。

将错误率最小的思想进一步扩展，提出更广泛的概念——风险。

例如，在癌细胞识别中，错检和漏检所带来的后果严重程度是不同的；

错检：给病人带来精神上的负担

漏检；导致病人错失良机

在这种情况下，引入风险的概念，以求风险最小的决策更为合理。

决策 \ 状态	阳性	阴性
阳性	真阳性（TP）	假阳性（FP）
阴性	假阴性（FN）	真阴性（TN）

基于最小风险的Bayes决策

思想：考虑到对某一类的错判要比对另一类的错判更为关键，把最小错误率的贝叶斯判决做一些修改，提出了“条件平均风险”的概念。

最小风险贝叶斯决策基本思想：

以各种错误分类所造成的平均风险最小为规则，进行分类决策。

风险的概念常与损失相联系，损失是指当真值和决策结果的不一致会带来后果。而损失函数的期望值便称为风险函数。

基于最小风险的Bayes决策

损失函数：$\lambda\left(\alpha_i,\omega_j\right)\quad i=1,2,...,a;\ \ j=1,2,...,m$ 表示当处于状态 ω_j 时，采取决策为 α_i 所带来的损失。

在决策论中，常以决策表表示各种情况下的决策损失：

损失 状态 / 决策	自然状态 ω_1	ω_2	...	ω_j	...	ω_c
α_1	$\lambda(\alpha_1,\omega_1)$	$\lambda(\alpha_1,\omega_2)$	...	$\lambda(\alpha_1,\omega_j)$	...	$\lambda(\alpha_1,\omega_c)$
α_2	$\lambda(\alpha_2,\omega_1)$	$\lambda(\alpha_2,\omega_2)$	...	$\lambda(\alpha_2,\omega_j)$	...	$\lambda(\alpha_2,\omega_c)$
⋮	⋮	⋮	...	⋮		⋮
α_i	$\lambda(\alpha_i,\omega_1)$	$\lambda(\alpha_i,\omega_2)$	...	$\lambda(\alpha_i,\omega_j)$	...	$\lambda(\alpha_i,\omega_c)$
⋮	⋮	⋮	...	⋮		⋮
α_a	$\lambda(\alpha_a,\omega_1)$	$\lambda(\alpha_a,\omega_2)$	...	$\lambda(\alpha_a,\omega_j)$	...	$\lambda(\alpha_a,\omega_c)$

基于最小风险的Bayes决策

当引入“损失”的概念后，就不能只根据后验概率的大小来做决策，而必须考虑所采取的决策是否损失最小。

对于给定的 $\boldsymbol{x}$，如果采取决策 α_i，λ 在 m 个 $\lambda\left(\alpha_i,\omega_j\right),\ j=1,2,...,m$ 当中任取一个，其相应概率为 $P\left(\omega_j\mid\boldsymbol{x}\right)$。因此在采取决策 α_i 情况下的条件期望损失 $R\left(\alpha_i\mid\boldsymbol{x}\right)$ 为：

$$R\left(\alpha_i\mid\boldsymbol{x}\right)=E\left[\lambda\left(\alpha_i,\omega_j\right)\right]=\sum_{j=1}^{m}\lambda\left(\alpha_i,\omega_j\right)P\left(\omega_j\mid\boldsymbol{x}\right)$$

在决策论中又把采取决策 α_i 的条件期望损失 $R\left(\alpha_i\mid\boldsymbol{x}\right)$ 称为条件平均风险。

对于给定的 x，决策规则为：

如果 $R(\alpha_k\mid\boldsymbol{x})=\min\limits_{i=1,\cdots,a}R(\alpha_i\mid\boldsymbol{x})$ ，则 $\alpha=\alpha_k$

基于最小风险的Bayes决策

条件风险 $R(\alpha_i \mid \boldsymbol{x})$ 反映了对某一 $\boldsymbol{x}$ 的取值采取决策 α_i 所带来的风险。

不同的观察值 $\boldsymbol{x}$，采取哪种决策是由 $\boldsymbol{x}$ 的取值决定的，所以决策 α_i 可看作是 $\boldsymbol{x}$ 的函数

对于一系列 $\boldsymbol{x}$ 及其决策 $\alpha_i(\boldsymbol{x})$，期望风险 $\boldsymbol{R}$ 为条件风险对观测值 $\boldsymbol{x}$ 的数学期望：

$$R(\alpha) = E\left[R(\alpha_i(\boldsymbol{x}) \mid \boldsymbol{x})\right] = \int R(\alpha_i(\boldsymbol{x}) \mid \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

最小风险决策： $\min \ R(\alpha)$ ---- 期望风险最小化

期望风险 $\boldsymbol{R}$ 反映对整个特征空间所有 $\boldsymbol{x}$ 的取值都采取相应的决策 $\alpha(\boldsymbol{x})$ 所带来的**平均风险**；

如果在采取每一个决策或行动时，都使其风险最小，则对所有的 $\boldsymbol{x}$ 做出决策时，其期望风险也必然最小，这样的决策就是**最小风险贝叶斯决策**。

基于最小风险的Bayes决策

条件平均风险与平均风险的区别

- 条件平均风险：对某个样本而言。
- 平均风险：对模式总体而言。

计算：可采取以下步骤（对于给定的样本 $\boldsymbol{x}$):

（1）计算后验概率：

$$P(\omega_j \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \omega_j)P(\omega_j)}{\sum_{i=1}^{c} p(\boldsymbol{x} \mid \omega_i)P(\omega_i)}, \quad j = 1, \cdots, m$$

（2）计算风险：

$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{j=1}^{m} \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid \boldsymbol{x}), \ i = 1, \cdots, a$$

（3）决策：

$$\alpha = \arg \min_{i=1,\cdots,a} R(\alpha_i \mid \boldsymbol{x})$$

基于最小风险的Bayes决策

两类情况：（损失函数表）

状态 / 决策	ω_1	ω_2
α_1	λ_{11}	λ_{12}
α_2	λ_{21}	λ_{22}

若决策α_1 将 $\boldsymbol{x}$ 判为 ω_1类，决策α_2 将 $\boldsymbol{x}$ 判为 ω_2类。

决策α_1：

$$R(\alpha_1|\boldsymbol{x}) = \lambda_{11}P(\omega_1|\boldsymbol{x}) + \lambda_{12}P(\omega_2|\boldsymbol{x}) \sim \lambda_{11}P(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{12}P(\boldsymbol{x}|\omega_2)P(\omega_2)$$

决策α_2：

$$R(\alpha_2|\boldsymbol{x}) = \lambda_{21}P(\omega_1|\boldsymbol{x}) + \lambda_{22}P(\omega_2|\boldsymbol{x}) \sim \lambda_{21}P(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{22}P(\boldsymbol{x}|\omega_2)P(\omega_2)$$

决策规则：

若$R(\alpha_1 \mid \boldsymbol{x}) < R(\alpha_2 \mid \boldsymbol{x})$　　则 $\boldsymbol{x} \in \omega_1$

若$R(\alpha_1 \mid \boldsymbol{x}) > R(\alpha_2 \mid \boldsymbol{x})$　　则 $\boldsymbol{x} \in \omega_2$

基于最小风险的Bayes决策

两类情况：

$$R(\alpha_1|\boldsymbol{x}) < R(\alpha_2|\boldsymbol{x}) \qquad \text{则}\ \boldsymbol{x} \in \omega_1$$

$$\lambda_{11}P(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{12}P(\boldsymbol{x}|\omega_2)P(\omega_2) < \lambda_{21}P(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{22}P(\boldsymbol{x}|\omega_2)P(\omega_2)$$

$$(\lambda_{11}-\lambda_{21})P(\boldsymbol{x}|\omega_1)P(\omega_1) < (\lambda_{22}-\lambda_{12})P(\boldsymbol{x}|\omega_2)P(\omega_2)$$

$$\frac{p(\boldsymbol{x}\,|\,\omega_1)}{p(\boldsymbol{x}\,|\,\omega_2)} > \frac{(\lambda_{22}-\lambda_{12})P(\omega_2)}{(\lambda_{11}-\lambda_{21})P(\omega_1)}$$

$l_{12}(\boldsymbol{x}) = \dfrac{p(\boldsymbol{x}\,|\,\omega_1)}{p(\boldsymbol{x}\,|\,\omega_2)}$ 称为似然比；　　$\theta_{12} = \dfrac{(\lambda_{22}-\lambda_{12})P(\omega_2)}{(\lambda_{11}-\lambda_{21})P(\omega_1)}$ 称为阈值。

0-1损失最小风险贝叶斯决策

损失函数为特殊情况： $\lambda_{ii}=0 \quad \lambda_{ij}=1,\ i\neq j$

决策α_1： $R(\alpha_1|\boldsymbol{x})=P(\omega_2|\boldsymbol{x}) \sim P(\boldsymbol{x}|\omega_2)P(\omega_2)$

决策α_2： $R(\alpha_2|\boldsymbol{x})=P(\omega_1|\boldsymbol{x}) \sim P(\boldsymbol{x}|\omega_1)P(\omega_1)$

决策规则：

若$R(\alpha_1|\boldsymbol{x}) < R(\alpha_2|\boldsymbol{x})$ 则 $\boldsymbol{x}\in\omega_1$

若$R(\alpha_1|\boldsymbol{x}) > R(\alpha_2|\boldsymbol{x})$ 则 $\boldsymbol{x}\in\omega_2$

错了损失为1，不错损失为0

决策＼状态	ω_1	ω_2
α_1	0	1
α_2	1	0

基于最小错误率的决策可看做是基于最小风险的决策的一个特例

例　在细胞识别中，病变细胞和正常细胞的先验概率分别为

$$P(\omega_1)=0.05,\quad P(\omega_2)=0.95$$

现有一待识别细胞，观察值为 $\boldsymbol{x}$，从类概率密度分布曲线上查得

$$P(\boldsymbol{x}\mid\omega_1)=0.5,\quad P(\boldsymbol{x}\mid\omega_2)=0.2$$

损失函数分别为λ_{11}=0，λ_{12}=1，λ_{21}=10，λ_{22}=0。按最小风险贝叶斯决策分类。

解：计算 θ_{12} 和 $l_{12}(\boldsymbol{x})$ 得：

$$\theta_{12}=\frac{(\lambda_{12}-\lambda_{22})P(\omega_2)}{(\lambda_{21}-\lambda_{11})P(\omega_1)}=\frac{(1-0)\times 0.95}{(10-0)\times 0.05}=1.9$$

$$l_{12}(\boldsymbol{x})=\frac{P(\boldsymbol{x}\mid\omega_1)}{P(\boldsymbol{x}\mid\omega_2)}=\frac{0.5}{0.2}=2.5$$

$\because l_{12}(\boldsymbol{x})>\theta_{12}$　　$\therefore \boldsymbol{x}\in\omega_1$ 为病变细胞。

例: 两类细胞识别问题：正常类(ω_1)和异常类(ω_2)

根据已有知识和经验，两类的先验概率为：

正常(ω_1)： $P(\omega_1)=0.9$； 异常(ω_2)： $P(\omega_2)=0.1$

对某一样本观察值$\boldsymbol{x}$，通过计算或查表得到：

$P(\boldsymbol{x}|\omega_1)=0.2$， $P(\boldsymbol{x}|\omega_2)=0.4$

按最小风险决策如何对细胞 $\boldsymbol{x}$ 进行分类？

决策 \ 状态	ω_1	ω_2
α_1	0	6
α_2	1	0

$\lambda_{11}=0$， $\lambda_{12}=6$， $\lambda_{21}=1$， $\lambda_{22}=0$

$$P(\omega_1|\boldsymbol{x})=\frac{P(\omega_1)p(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x})}=\frac{0.9\times0.2}{0.9\times0.2+0.1\times0.4}=\frac{0.18}{0.22}=0.818$$

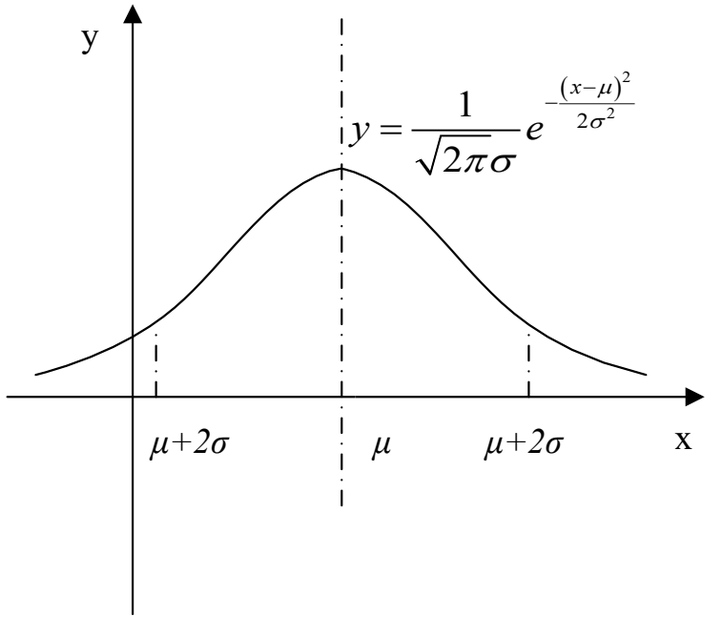
$$P(\omega_2|\boldsymbol{x})=\frac{P(\omega_2)p(\boldsymbol{x}|\omega_2)}{P(\boldsymbol{x})}=\frac{0.1\times0.4}{0.9\times0.2+0.1\times0.4}=\frac{0.04}{0.22}=0.182$$

$$R(\omega_1|\boldsymbol{x})=\sum_{j=1}^{2}\lambda_{1j}P(\omega_j|\boldsymbol{x})=\lambda_{12}P(\omega_2|\boldsymbol{x})=1.092$$

$$R(\omega_2|\boldsymbol{x})=\sum_{j=1}^{2}\lambda_{2j}P(\omega_j|\boldsymbol{x})=\lambda_{21}P(\omega_1|\boldsymbol{x})=0.818$$

$$j=\arg\min_i R(\omega_i|\boldsymbol{x})=2 \qquad \boldsymbol{x}\in\omega_2$$

正态分布时的统计决策



许多实际的数据集：均值附近分布较多的样本；距均值点越远，样本分布越少。此时正态分布（高斯分布）是一种合理的近似。

正态分布概率模型的优点：

* 物理上的合理性。
* 数学上的简单性。

单变量(一维)正态分布及其两个重要参数：

均值（中心）

方差（分散度）

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

$$\mu=E\{x\}=\int p(x)dx$$

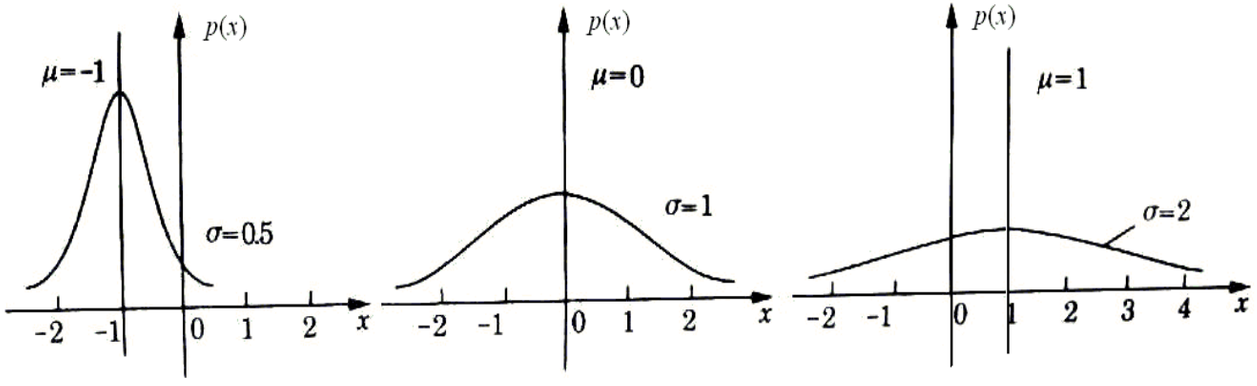
$$\sigma^2=\int(x-\mu)^2p(x)dx=E\left\{(x-\mu)^2\right\}$$

一维正态分布

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

曲线如图示：

① $\mu=-1$，$\sigma=0.5$；　② $\mu=0$，$\sigma=1$；　③ $\mu=1$，$\sigma=2$.



一维正态曲线的性质：

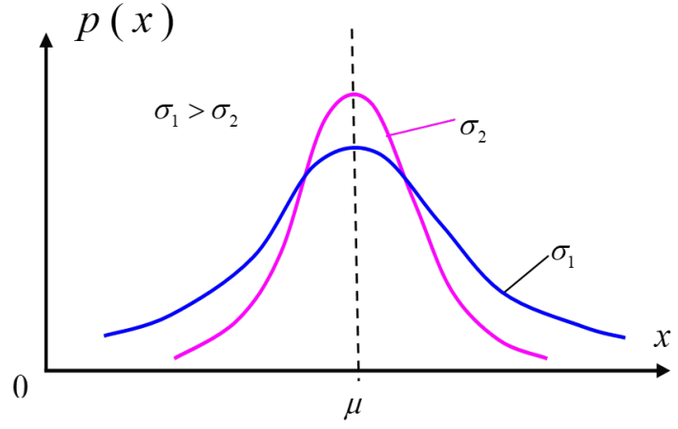
（1）曲线在 x 轴的上方，与x轴不相交。

（2）曲线关于直线 $x=\mu$对称。

（3）当 $x=\mu$时，曲线位于最高点。

（4）当$x<\mu$时，曲线上升；当$x>\mu$时，曲线下降.并且当曲线向左、右两边无限延伸时，以x轴为渐近线，向它无限靠近。

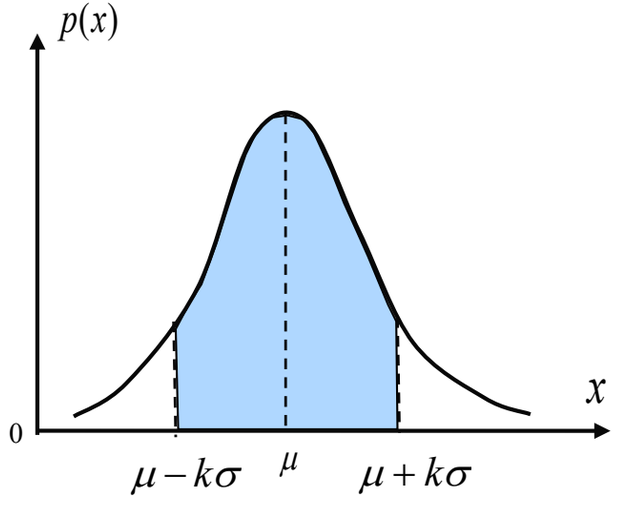
（5）μ一定时，曲线的形状由σ确定。σ越大，曲线越“矮胖”，表示总体的分布越分散；σ越小。曲线越“瘦高”。表示总体的分布越集中。



一维正态分布

3σ规则

$$P\{\mu-k\sigma\le x\le\mu+k\sigma\}=\begin{cases}0.683, & \text{当}k=1\text{时}\\ 0.954, & \text{当}k=2\text{时}\\ 0.997, & \text{当}k=3\text{时}\end{cases}$$



即：绝大部分样本都落在了均值μ附近±3σ的范围内，因此正态密度曲线完全可由均值和方差来确定，常简记为：

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

多变量（*n*维）正态分布

实际应用中，可以同时观测多个值，用向量 $\mathbf{x}=(\mathrm{x}_1,\mathbf{x}_2,\mathbf{x}_3,\ldots,\mathbf{x}_n)^T$ 表示。

多变量正态分布的密度函数：

$$p(\mathbf{x})=\frac{1}{(2\pi)^{n/2}\,|\sum|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\sum{}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\},\quad \mathbf{x}\in R^n$$

均值向量 $\boldsymbol{\mu}=E[\mathbf{x}]=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_n)^T$ 其中$\boldsymbol{\mu}_i=E(\mathbf{x}_i)$

协方差矩阵 $\Sigma=E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]=(\sigma_{ij}^2)_{n*n}=\begin{bmatrix}\sigma_{11}^2 & \cdots & \sigma_{1n}^2\\ \vdots & \ddots & \vdots\\ \sigma_{n1}^2 & \cdots & \sigma_{nn}^2\end{bmatrix}$ 对角线元素为方差

多维正态密度函数完全由它的均值 $\boldsymbol{\mu}$和协方差矩阵$\boldsymbol{\Sigma}$所确定，简记为 $N(\mu,\Sigma)$

正态分布的性质

※ 参数$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$完全决定分布

均值向量$\boldsymbol{\mu}$由n个参数组成，协方差矩阵$\boldsymbol{\Sigma}$由于其对称性故其独立元素为$n(n+1)/2$个；

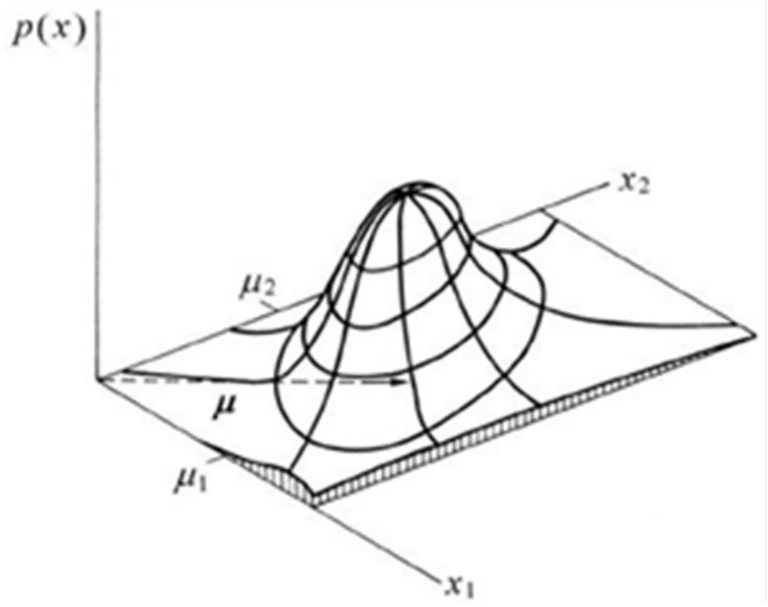
※ 等概率密度轨迹为超椭球面

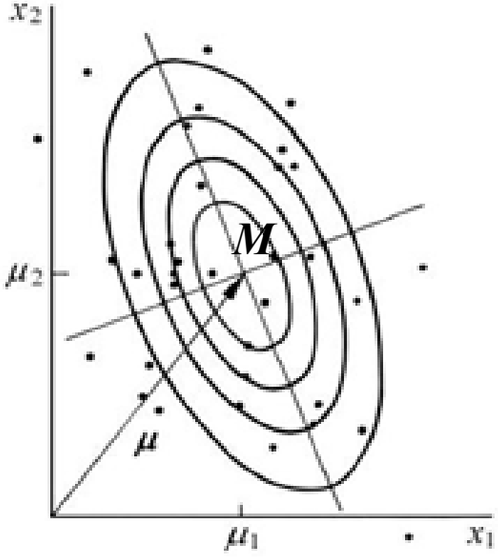
以二维正态密度函数为例：

等高线（等密度线）投影到x_1ox_2面上为椭圆，从原点O到点M 的向量为均值$\boldsymbol{\mu}$ 。

椭圆的位置：由均值向量$\boldsymbol{\mu}$决定；

椭圆的形状：由协方差矩阵$\boldsymbol{\Sigma}$决定。





正态分布的最小错误率贝叶斯决策规则

前面介绍的Bayes方法事先必须求出$p(\boldsymbol{X}|\omega_i)$, $P(\omega_i)$ 。

而当 $p(\boldsymbol{X}|\omega_i)$呈正态分布时，只需要知道 $\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$ 即可。

1）多类情况

具有n种模式类别的多变量正态密度函数为：

$$p(\mathbf{x})=\frac{1}{(2\pi)^{n/2}\,|\sum|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\sum{}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\},\quad \mathbf{x}\in R^n$$

每一类模式的分布密度都完全被其均值向量$\boldsymbol{\mu}_i$和协方差矩阵 $\boldsymbol{\Sigma}_i$所规定，其定义为：

$\sum_i = E[(\mathbf{x}_i-\boldsymbol{\mu}_i)(\mathbf{x}_i-\boldsymbol{\mu}_i)^T]$　　协方差矩阵 $\boldsymbol{\Sigma}_i$：反映样本分布区域的形状；

$\boldsymbol{\mu}_i = E(\mathbf{x}_i)$　　均值向量$\boldsymbol{\mu}_i$：表明了区域中心的位置。

正态分布的最小错误率Bayes决策的判别函数

最小错误率Bayes决策中，ω_i类的判别函数为 $p(\boldsymbol{X}\mid\omega_i)P(\omega_i)$

$$p(\boldsymbol{X}\mid\omega_i)=\frac{1}{(2\pi)^{n/2}\mid\boldsymbol{\Sigma}_i\mid^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right\}$$

为了方便计算，对正态密度函数取对数：

$$\ln[p(\boldsymbol{X}\mid\omega_i)P(\omega_i)]=\ln\left[p(\boldsymbol{X}\mid\omega_i)\right]+\ln\left[P(\omega_i)\right]\ =\ln P(\omega_i)-\frac{n}{2}\ln 2\pi-\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right|-\frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right\}$$

对数是单调递增函数，取对数后仍有相对应的分类性能。

去掉与i无关的项，得判别函数：

$$d_i(\boldsymbol{X})=\ln P(\omega_i)-\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right|-\frac{1}{2}\left\{(\boldsymbol{X}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right\}\quad i=1,2,\ldots,n$$

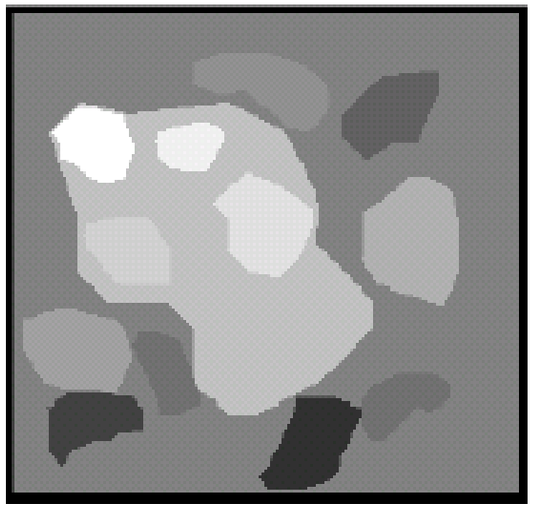
—— 正态分布的最小错误率Bayes决策的判别函数。

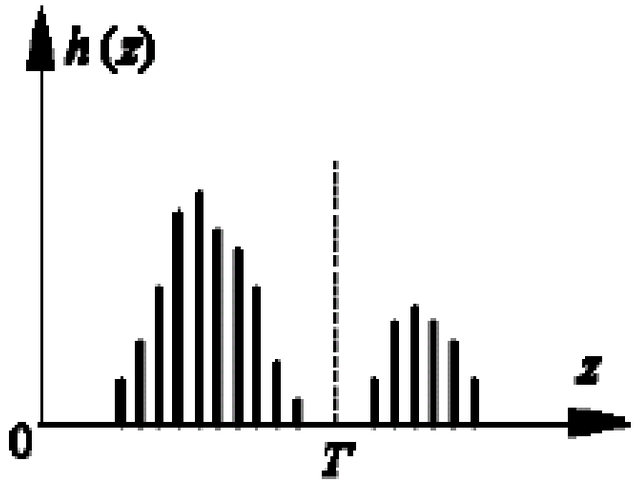
判决规则同前：　若 $d_j(\boldsymbol{X})>d_i(\boldsymbol{X}),\quad i=1,2,\ldots,M,\ i\neq j$　　则 $\boldsymbol{X}\in\omega_j$

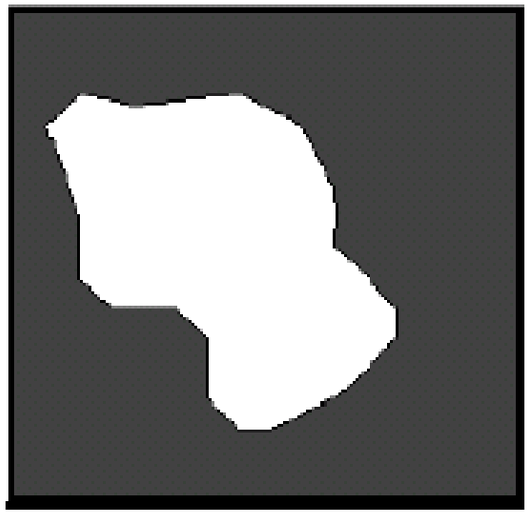
最小错误率贝叶斯决策应用——直方图阈值分割

60年代中期，Prewitt提出了直方图双峰法

- 如果灰度级直方图呈明显的双峰状，则选取两峰之间的谷底所对应的灰度级作为阈值。







问题：如何自动获取最优阈值？

最小错误率贝叶斯决策应用——最优阈值

思路：使图像中目标物和背景分割错误最小的阈值。

设一幅图像只由目标和背景组成，已知像素灰度概率密度分别为 $p_{obj}(z)$ 和 $p_{bg}(z)$，目标像素占整幅图像像素比为 α 。

对于某个阈值 z_k，有：$C_z = \begin{cases} \text{目标} & z \geq z_k \\ \text{背景} & z < z_k \end{cases}$

分析：类条件概率 $p_{obj}(z) \rightarrow p(z|obj)$ $p_{bg}(z) \rightarrow p(z|bg)$

先验概率 $p(obj) = \alpha$ $p(bg) = 1 - \alpha$

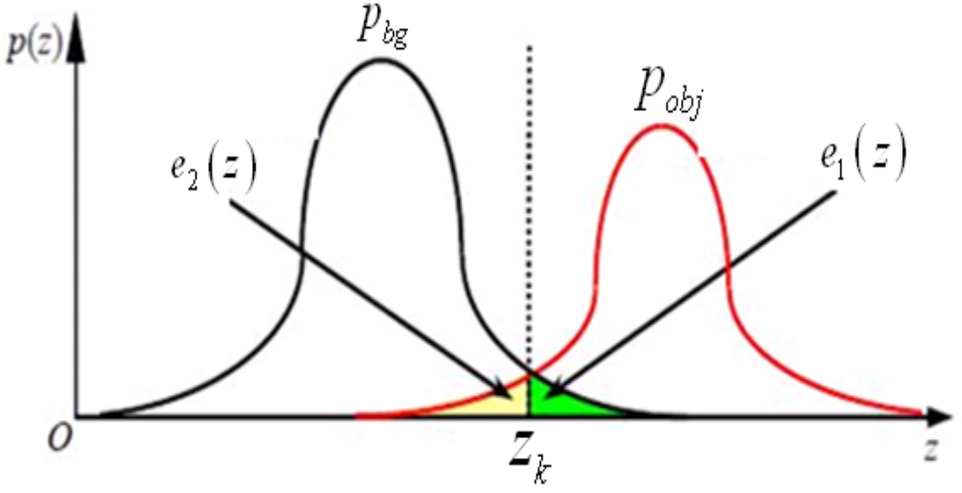
最小错误率贝叶斯决策应用——最优阈值

背景错归为目标的概率：$e_1(z_k)=\int_{z_k}^{+\infty} p_{bg}(z)\,dz$

目标错归为背景的概率：$e_2(z_k)=\int_{-\infty}^{z_k} p_{obj}(z)\,dz$

总的错分概率 $e(z_k)=(1-\alpha)\cdot e_1(z_k)+\alpha\cdot e_2(z_k)$

最佳阈值的目标：寻找阈值 z_k，使得 $e(z_k)$ 最小。



问题：如何找到 $e(z_k)$ 最小？

最小错误率贝叶斯决策应用——最优阈值

最佳阈值的目标： $z_k = \arg\min e(z_k)$

求导，并令 $\dfrac{\partial e(z_k)}{\partial z_k} = 0$

$$e_1(z_k) = \int_{z_k}^{+\infty} p_{bg}(z)dz \Rightarrow \frac{\partial e_1(z_k)}{\partial z_k} = -p_{bg}(z_k)$$

$$e_2(z_k) = \int_{-\infty}^{z_k} p_{obj}(z)dz \Rightarrow \frac{\partial e_2(z_k)}{\partial z_k} = p_{obj}(z_k)$$

$$\alpha \cdot p_{obj}(z_k) = (1-\alpha) \cdot p_{bg}(z_k)$$

假设目标和背景的概率密度函数为高斯模型

$$p_{obj}(z) = \frac{1}{\sqrt{2\pi}\sigma_{obj}} e^{-\frac{(z-\mu_{obj})^2}{2\sigma_{obj}^2}} \qquad p_{bg}(z) = \frac{1}{\sqrt{2\pi}\sigma_{bg}} e^{-\frac{(z-\mu_{bg})^2}{2\sigma_{bg}^2}}$$

最小错误率贝叶斯决策应用——最优阈值

$$\alpha \cdot p_{obj}(z_k) = (1-\alpha) \cdot p_{bg}(z_k)$$

两边取对数，并化简

$$Az_k^2 + Bz_k + C = 0$$

其中

$$\begin{cases} A = \sigma_{obj}^2 - \sigma_{bg}^2 \\ B = 2\left(\mu_{obj}\sigma_{obj}^2 - \mu_{bg}\sigma_{bg}^2\right) \\ C = \mu_{bg}^2\sigma_{obj}^2 - \mu_{obj}^2\sigma_{bg}^2 + 2\sigma_{obj}^2\sigma_{bg}^2 \ln\left(\dfrac{\alpha\sigma_{bg}}{(1-\alpha)\sigma_{obj}}\right) \end{cases}$$

最小错误率贝叶斯决策应用——最优阈值

方程 $Az_k^2 + Bz_k + C = 0$ 在一般情况下存在两个解。

当且仅当 $\sigma_{obj}=\sigma_{bg} = \sigma$ 时，有唯一解：

$$z_k = \frac{\mu_{obj} + \mu_{bg}}{2} + \frac{\sigma^2}{\mu_{obj} - \mu_{bg}} \ln\left(\frac{\alpha}{1-\alpha}\right)$$

若 $\alpha = \frac{1}{2}$，则 $z_k = \frac{\mu_{obj} + \mu_{bg}}{2}$

均值迭代阈值分割方法

最小错误率贝叶斯决策应用——均值迭代阈值分割方法

Step1. 选择一个初始的估计阈值 T

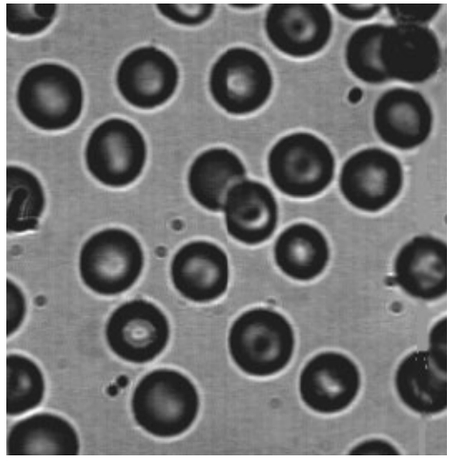
（可以用图像的平均灰度值作为初始阈值）

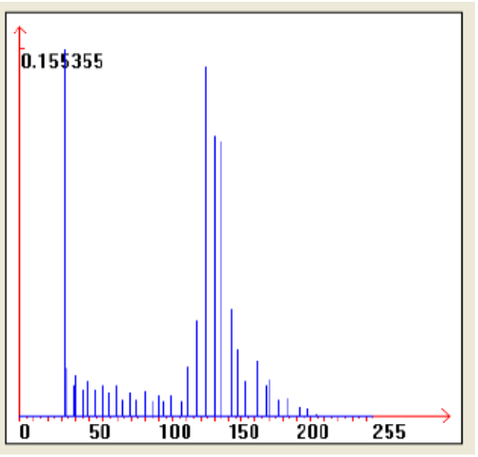
Step2. 用该阈值把图像分割成两个部分 R_1 和 R_2

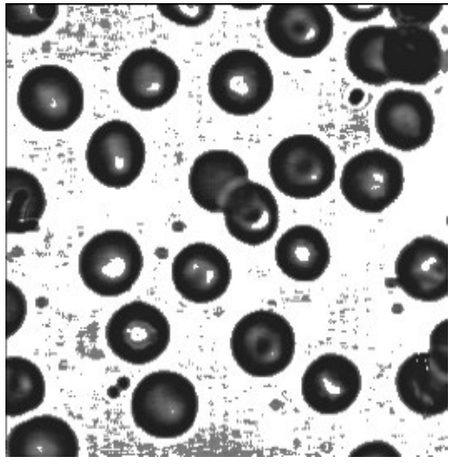
Step3. 分别计算 R_1 和 R_2 的灰度均值 μ_1和 μ_2

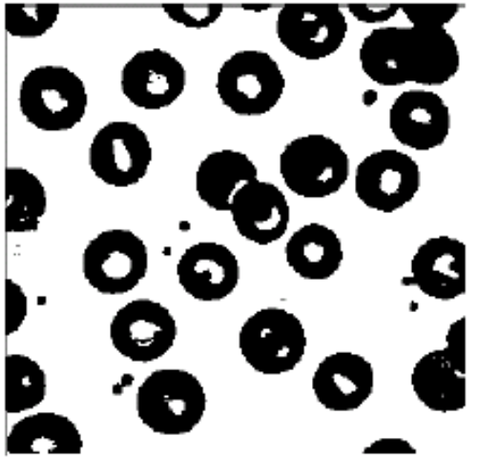
Step4. 选择一个新的阈值 T＝(μ_1＋μ_2)/2

Step5. 重复2-4直至后续迭代中平均灰度值 μ_1和 μ_2保持不变









- **贝叶斯决策理论**
  基本概念，决策准则，概率论基础，…
- **最小错误/最小风险贝叶斯决策**
  最小错误率决策规则、条件风险，期望风险， …
- **概率密度函数估计**
  基本概念，最大似然估计…
- **贝叶斯估计和贝叶斯学习**
  贝叶斯估计，贝叶斯学习 …
- **非参数估计方法**
  基本概念，Parzen窗法 …
- **贝叶斯分类器**
  贝叶斯网络，朴素贝叶斯分类器 …

概率密度函数估计

贝叶斯决策需要已知两种知识：

(1) 各类的先验概率$P(\omega_i)$

(2) 各类的条件概率密度函数 $p(\boldsymbol{x}|\omega_i)$

$$P(\omega_i \mid X) = \frac{P(\omega_i)P(X \mid \omega_i)}{P(X)}$$

对未知样本分类（设计分类器）

实际问题：已知一定数目的样本，对未知样本分类（设计分类器） 怎么办？

一种很自然的想法：

a) 首先根据样本估计 $P(\omega_i)$ 和 $p(\boldsymbol{x} \mid \omega_i)$，记 $\hat{P}(\omega_i)$和 $\hat{p}(\boldsymbol{x} \mid \omega_i)$；

b) 然后用估计的概率密度设计贝叶斯分类器。

——（基于样本的）两步贝叶斯决策

概率密度函数估计

希望：当样本数 $N \to \infty$ 时，得到的分类器收敛于理论上的最优解。

$$\hat{p}(\boldsymbol{x} \mid \omega_i) \xrightarrow{N \to \infty} p(\boldsymbol{x} \mid \omega_i)$$

$$\hat{P}(\omega_i) \xrightarrow{N \to \infty} P(\omega_i)$$

重要前提：

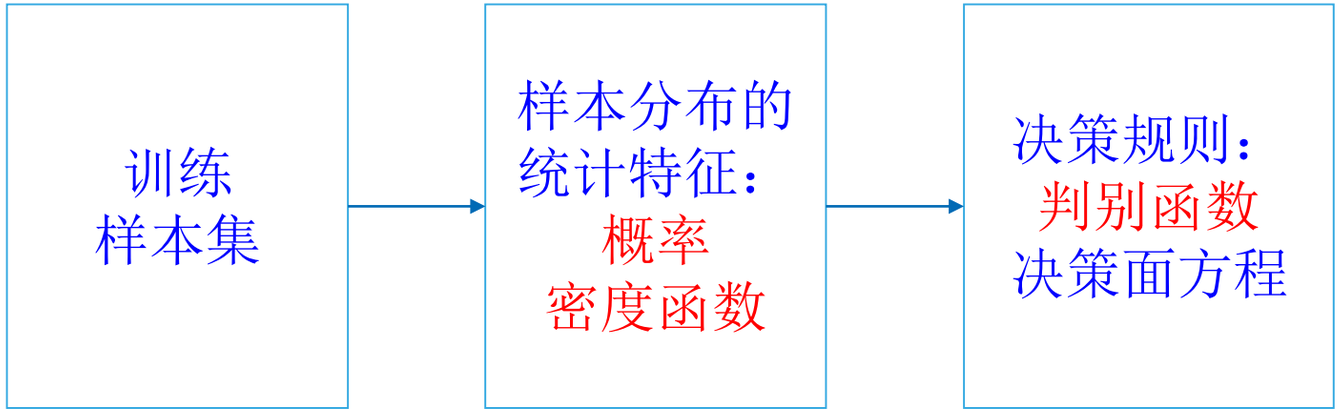
- 训练样本的分布能代表样本的真实分布，所谓i.i.d条件 **independent and identically distributed (i.i.d.)**
- 有充分的训练样本

面临的问题：

- 如何利用样本集进行估计
- 估计量的评价
- 利用样本集估计错误率

概率密度函数估计

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{\sum_j p(\mathbf{x} \mid \omega_j)P(\omega_j)}$$



最一般情况下适用的“最优”分类器：错误率最小，对分类器设计在理论上有指导意义。

获取统计分布及其参数很困难，实际问题中并不一定具备获取准确统计分布的条件。

概率密度函数估计

本讲讨论内容：如何利用样本集估计先验概率和概率密度函数？

- 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
    - 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
    - 神经网络方法

基本概念

参数估计 (parametric estimation)：

- 已知概率密度函数的形式，只是其中几个参数未知，目标是根据样本估计这些参数的值。

非参数估计 (nonparametric methods)：

- 概率密度函数的形式未知，直接估计概率密度函数的方法。

几个名词：

统计量(statistics)：样本的某种函数，用来作为对某参数的估计

参数空间(parametric space)：待估计参数的取值空间 $\theta \in \Theta$

点估计和区间估计

点估计的估计量(estimation)：$\hat{\theta}(x_1, x_2, \cdots, x_N)$

估计量的评价标准

估计量的评价标准：无偏性，有效性，一致性

- **无偏性**：
  - 估计量抽样分布的数学期望等于总体参数的真值，$E(\blacksquare)=\theta$
- **有效性**：
  - 指估计量与总体参数的离散程度，离散程度用方差来衡量。$D(\blacksquare)$小，估计更有效
- **一致性**：
  - 样本数目越大，估计量就越来越接近总体参数的真实值。样本数趋于无穷时，依概率趋于θ：

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

最大似然估计（Maximum Likelihood）

设：ω_i类的类概率密度函数 $p(\boldsymbol{x}|\omega_i)$ 具有某种确定的函数形式；

估计的参数 $\boldsymbol{\theta}$ 是该函数的一个确定的未知参数或参数集（不是随机量）

最大似然估计的思想：根据一组属于类ω_i的已知样本，反推最具有可能（最大概率）的参数 $\boldsymbol{\theta}$

已知样本都是从密度为 $p(\boldsymbol{x}|\omega_i)$ 的总体中独立抽取出来的（即i.i.d）。

样本集中的样本只包含本类的分布信息。

参数 $\boldsymbol{\theta}$ 通常是向量，比如一维正态分布 $N(\mu_i,\sigma_1^2)$，未知参数可能是 $\theta_i=\begin{bmatrix}\mu_i\\ \sigma_i^2\end{bmatrix}$ → **考虑之前是如何做的？有什么理论依据？**

此时 $p(\boldsymbol{x}|\omega_i)$ 可写成 $p(\boldsymbol{x}|\omega_i,\theta_i)$ 或 $p(\boldsymbol{x}|\theta_i)$ 。

最大似然估计把 $\boldsymbol{\theta}$ 当作确定的未知量进行估计。

最大似然估计（Maximum Likelihood）

考虑一类已知分类结果的样本，记已知样本为 $\boldsymbol{\mathcal{X}} = \{x_1, x_2, ..., x_N\}$

似然函数（likelihood function）

$$l(\theta) = p(\boldsymbol{\mathcal{X}} \mid \theta) = p(x_1, x_2, ..., x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$ 当满足i.i.d

——在参数 θ 下观测到样本集 $\boldsymbol{\mathcal{X}}$ 的概率（联合分布）密度

基本思想：

如果在参数 $\theta = \hat{\theta}$ 下，$l(\theta)$ 最大，则 $\hat{\theta}$ 应是“最可能”的参数值，它是样本集的函数，记作 $\hat{\theta} = f(x_1, x_2, ..., x_N) = f(\boldsymbol{\mathcal{X}})$ 。称作最大似然估计量。

$$\hat{\theta}_{ML} = \arg\max_{\theta} l(\theta)$$

最大似然估计（Maximum Likelihood）

$$l(\theta)=p(\mathcal{X}\mid\theta)=p(x_1,x_2,...,x_N\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)\qquad\Rightarrow\hat{\theta}_{ML}=\arg\max_{\theta}l(\theta)$$

若未知参数只有一个，即 θ 为标量， 在似然函数满足连续、可微的正则条件下，

最大似然估计量就是微分方程的解：

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

若未知参数不只一个，即 $\boldsymbol{\theta}$ 为具有S个分量的向量 $\boldsymbol{\theta}=[\theta_1,\theta_2,...,\theta_S]$，则最大似然估计量为：

$$\nabla_{\boldsymbol{\theta}}\,l(\theta)=0\qquad\nabla_{\boldsymbol{\theta}}=\left[\frac{\partial}{\partial\theta_1}\quad\cdots\quad\frac{\partial}{\partial\theta_S}\right]^T$$

最大似然估计（Maximum Likelihood）

为了便于分析，采用似然函数的对数更容易，即：

$$H(\boldsymbol{\theta}) = \ln l(\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} p(x_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(x_i \mid \boldsymbol{\theta})$$

当满足i.i.d

由于对数函数为单调递增函数，因此使对数似然函数最大的 $\boldsymbol{\theta}$ 值也必然使似然函数最大，即

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln p(x_i \mid \boldsymbol{\theta})$$

最大似然估计量为：

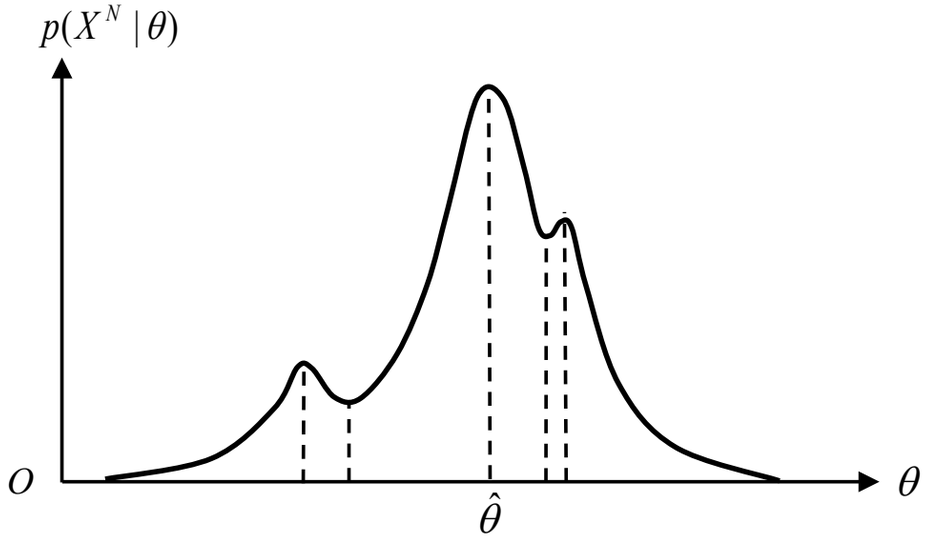
$$\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) = 0 \quad \Rightarrow \quad \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) = \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(x_i \mid \boldsymbol{\theta}) = 0 \quad \Rightarrow \quad \begin{cases} \sum_{i=1}^{N} \dfrac{\partial}{\partial \theta_1} \ln p(x_i \mid \boldsymbol{\theta}) = 0 \\ \sum_{i=1}^{N} \dfrac{\partial}{\partial \theta_2} \ln p(x_i \mid \boldsymbol{\theta}) = 0 \\ \vdots \\ \sum_{i=1}^{N} \dfrac{\partial}{\partial \theta_S} \ln p(x_i \mid \boldsymbol{\theta}) = 0 \end{cases}$$

最大似然估计（Maximum Likelihood）

讨论：

上述 $\nabla_{\boldsymbol{\theta}} l(\theta)=0$ 或 $\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta})=0$ 可能不是唯一解，如右图。

只有 $\hat{\theta}$ 才使得似然函数最大。



$\boldsymbol{\theta}$为一维时的最大似然估计示意图

最大似然估计（Maximum Likelihood）

均匀分布下的最大似然估计示例

随机变量**x**服从均匀分布，但参数 θ_1和θ_2未知，

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln p(x_i \mid \boldsymbol{\theta})$$

$$p(x \mid \boldsymbol{\theta}) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{其他} \end{cases} \quad \Rightarrow \quad \ln p(x \mid \boldsymbol{\theta}) = -\ln(\theta_2 - \theta_1)$$

样本集 $\mathcal{X} = \{x_1, x_2, \cdots, x_N\}$

似然函数 $l(x) = p(\mathcal{X} \mid \theta) = \prod_{k=1}^{N} p(x_k \mid \theta)$

最大似然估计（Maximum Likelihood）

均匀分布下的最大似然估计示例

对数似然函数 $H(\boldsymbol{\theta}) = \ln l(x) = \sum_{k=1}^{N} \ln P(x_k \mid \boldsymbol{\theta})$

最大似然估计量 $\hat{\theta}$ 满足：$\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) = \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(x_k \mid \boldsymbol{\theta}) = 0$

$$\begin{cases} \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_1} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = 0 \\ \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_2} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = 0 \end{cases}$$

$$\ln p(x \mid \boldsymbol{\theta}) = -\ln\left(\theta_2 - \theta_1\right) \quad \Rightarrow \quad \begin{cases} \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_1} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = \sum_{k=1}^{N} \dfrac{1}{\theta_2 - \theta_1} = \dfrac{N}{\theta_2 - \theta_1} \\ \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_2} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = -\sum_{k=1}^{N} \dfrac{1}{\theta_2 - \theta_1} = -\dfrac{N}{\theta_2 - \theta_1} \end{cases}$$

最大似然估计（Maximum Likelihood）

均匀分布下的最大似然估计示例

$$\begin{cases} \dfrac{N}{\theta_2-\theta_1}=0 \\ -\dfrac{N}{\theta_2-\theta_1}=0 \end{cases} \Rightarrow \quad \frac{1}{\theta_2-\theta_1}=0$$

显然只有当参数 θ_1和θ_2有一个为无穷大，才有可能有解，但这个解无意义。

解决方法就是寻找最接近与0的结果，即寻找$\theta_2-\theta_1$的最大值。

即 $\hat{\theta}_1=\min(\mathcal{X})=\min(x_1,x_2,\cdots,x_N)$ 取样本集中的最小值

$\hat{\theta}_2=\max(\mathcal{X})=\max(x_1,x_2,\cdots,x_N)$ 取样本集中的最大值

最大似然估计（Maximum Likelihood）

正态分布下的最大似然估计示例

以单变量正态分布为例

$$\theta = [\theta_1, \theta_2]^T \qquad \theta_1 = \mu \qquad \theta_2 = \sigma^2$$

$$p(x \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

$$\Rightarrow \quad \ln p(x \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

$$= -\frac{1}{2}\ln\left(2\pi\theta_2\right) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln p(x_i \mid \boldsymbol{\theta})$$

样本集 $\mathcal{X} = \{x_1, x_2, ..., x_N\}$

似然函数 $l(x) = p(\mathcal{X} \mid \theta) = \prod_{k=1}^{N} p(x_k \mid \theta)$

最大似然估计（Maximum Likelihood）

正态分布下的最大似然估计示例

对数似然函数 $H(\boldsymbol{\theta}) = \ln l(x) = \sum_{k=1}^{N} \ln P(x_k \mid \boldsymbol{\theta})$

最大似然估计量 $\hat{\theta}$ 满足： $\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) = \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(x_k \mid \boldsymbol{\theta}) = 0 \quad \Rightarrow \quad \begin{cases} \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_1} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = 0 \\ \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_2} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = 0 \end{cases}$

$$\ln p(x \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln\left(2\pi\theta_2\right) - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \quad \Rightarrow \quad \begin{cases} \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_1} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = \sum_{k=1}^{N} \dfrac{x_k - \theta_1}{\theta_2} \\ \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_2} \ln p\left(x_k \mid \boldsymbol{\theta}\right) = \sum_{k=1}^{N} \left(-\dfrac{1}{2\theta_2} + \dfrac{1}{2}\left(\dfrac{x_k - \theta_1}{\theta_2}\right)^2 \right) \end{cases}$$

最大似然估计（Maximum Likelihood）

正态分布下的最大似然估计示例

$$\begin{cases}\sum_{k=1}^{N}\dfrac{x_k-\theta_1}{\theta_2}=0\\ \sum_{k=1}^{N}\left(-\dfrac{1}{2\theta_2}+\dfrac{1}{2}\left(\dfrac{x_k-\theta_1}{\theta_2}\right)^2\right)=0\end{cases}\Rightarrow\begin{cases}\sum_{k=1}^{N}x_k-N\theta_1=0\\ -\dfrac{N}{2\theta_2}+\dfrac{1}{2\theta_2^2}\sum_{k=1}^{N}(x_k-\theta_1)^2=0\end{cases}\Rightarrow\begin{cases}\theta_1=\dfrac{1}{N}\sum_{k=1}^{N}x_k\\ \theta_2=\dfrac{1}{N}\sum_{k=1}^{N}(x_k-\theta_1)^2\end{cases}$$

由以上方程组解得均值和方差的估计量为

$$\hat{\mu}=\hat{\theta}_1=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k \qquad \hat{\sigma}^2=\hat{\theta}_2=\frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k-\hat{\mu})^2$$

最大似然估计（Maximum Likelihood）

正态分布下的最大似然估计示例

类似地，多维正态分布情况：

$$\hat{\boldsymbol{M}}_i = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k$$

$$\hat{\boldsymbol{C}}_i = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k - \hat{\boldsymbol{M}}_i)(\boldsymbol{x}_k - \hat{\boldsymbol{M}}_i)^{\mathrm{T}}$$

- **贝叶斯决策理论**
  基本概念，决策准则，概率论基础，…
- **最小错误/最小风险贝叶斯决策**
  最小错误率决策规则、条件风险，期望风险， …
- **概率密度函数估计**
  基本概念，最大似然估计…
- **贝叶斯估计和贝叶斯学习**
  贝叶斯估计，贝叶斯学习 …
- **非参数估计方法**
  基本概念，Parzen窗法 …
- **贝叶斯分类器**
  贝叶斯网络，朴素贝叶斯分类器 …

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

最大似然估计思想：参数 $\boldsymbol{\theta}$ 是类概率密度函数 $p(\boldsymbol{x}|\omega_i)$ 的一个确定的但未知的参数或参数集；根据一组属于类 ω_i 的已知样本，通过最大化似然函数来求解参数 $\boldsymbol{\theta}$ 。

贝叶斯估计和贝叶斯学习：将未知参数 $\boldsymbol{\theta}$ 看作为具有某种已知先验分布的随机变量，即参数 $\boldsymbol{\theta}$ 不是一个确定的未知数，而是符合一定的先验分布，如正态分布等；因此，在贝叶斯估计和贝叶斯学习中除了类条件概率密度 $p(\boldsymbol{x}|\omega_i)$ 符合一定的先验分布，参数 $\boldsymbol{\theta}$ 也符合一定的先验分布。然后通过贝叶斯规则将参数的先验分布转化为后验概率进行求解。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯估计和贝叶斯学习将未知参数看作随机参数进行考虑。

贝叶斯估计

1）贝叶斯估计

思路与贝叶斯决策类似，只是离散的决策状态变成了连续的估计。

设有样本集 $\boldsymbol{K}$（而不是一个样本$\boldsymbol{x}$），要求找出估计量 $\hat{\theta}$（而不是选出最佳决策α_k），用来估计$\boldsymbol{K}$ 所属总体分布的某个真实参数θ（而不是真实状态w_k），使带来的贝叶斯风险最小。

决策问题	估计问题
样本 $\boldsymbol{x}$	样本集 $\mathscr{X}$
决策 α_i	估计量 $\hat{\theta}$
真实状态 ω_j	真实参数 θ
状态空间 A 是离散空间	参数空间 Θ 是连续空间
先验概率 $P(\omega_j)$	参数的先验分布 $p(\theta)$

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯估计基本思想

把待估计参数 θ 看作具有先验分布$p(\theta)$的随机变量，其取值与样本集 $\boldsymbol{\mathcal{X}}$ 无关，

根据样本集 $\boldsymbol{\mathcal{X}}=\{x_1, x_2, \ldots, x_N\}$ 估计θ。

损失函数：把 θ 估计为 $\hat{\theta}$ 所造成的损失，记为 $\lambda(\hat{\theta},\theta)$

期望风险：$R=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\boldsymbol{x},\theta)d\theta d\boldsymbol{x}=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\boldsymbol{x})p(\boldsymbol{x})d\theta d\boldsymbol{x}=\int_{E^d}R(\hat{\theta}\,|\,\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$

其中 $\boldsymbol{x}\in E^d$，$\theta\in\Theta$，$R(\hat{\theta}\,|\,\boldsymbol{x})=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\boldsymbol{x})d\theta$ 为条件风险。

最小化期望风险 $\Rightarrow$ 最小化条件风险 （对所有可能的$\boldsymbol{x}$），此时 $\hat{\theta}$ 称为 θ 的贝叶斯估计量。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯估计基本思想

$$\hat{\theta}_{\mathrm{BE}} = \arg\min_{\theta} R(\theta \mid \boldsymbol{x})$$

不同损失函数，有不同的贝叶斯估计量。

若损失函数为 $L2$ norm，即平方误差损失函数 $\lambda(\hat{\theta},\theta) = (\theta - \hat{\theta})^2$

条件风险

$$R\left(\hat{\theta} \mid \boldsymbol{x}\right) = \int_{\Theta} \lambda\left(\hat{\theta},\theta\right) p\left(\theta \mid \boldsymbol{x}\right) d\theta = \int_{\Theta} \left(\theta - \hat{\theta}\right)^2 p\left(\theta \mid \boldsymbol{x}\right) d\theta$$ （证明略）

该项与$\hat{\theta}$ 无关且非负。

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

该项与$\hat{\theta}$ 有关，当该项为0时，条件风险最小。

要使条件风险最小，有$\hat{\theta}$ 为在给定$\boldsymbol{\mathcal{X}}$ 时 θ 的条件期望 $\hat{\theta} = E\left(\theta \mid \boldsymbol{\mathcal{X}}\right) = \int_{\Theta} \theta p\left(\theta \mid \boldsymbol{\mathcal{X}}\right) d\theta$

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

平方误差损失下的贝叶斯估计

步骤：

1、确定θ的先验概率$p(\theta)$

2、由样本集 $\mathcal{X}=\{x_1, x_2, \ldots, x_N\}$，计算样本集的联合分布 $p(\mathcal{X}\mid\theta)=\prod_{i=1}^{N} p(\boldsymbol{x}_i\mid\theta)$

3、利用贝叶斯公式，计算θ的后验概率

$$p(\theta\mid\mathcal{X})=\frac{p(\mathcal{X}\mid\theta)p(\theta)}{p(\mathcal{X})}$$

4、求贝叶斯估计量

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

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯学习基本思想

贝叶斯学习是一种“增量学习（incremental learning）”过程，本质是参数值随着样本增多趋近于真实值的过程。

基本思想：利用θ的先验分布$p(\theta)$及样本提供的信息，求出θ的后验密度$p(\theta|\mathcal{X})$后，不再去估计$\hat{\theta}$，而直接推断总体分布$p(\boldsymbol{x}|\mathcal{X})$。当$\hat{\theta}$为$\theta$的最大似然估计值，即贝叶斯解的结果$p(\boldsymbol{x}|\mathcal{X})$与最大似然估计的结果近似，即：$p(\boldsymbol{x}|\mathcal{X}) \sim p(\boldsymbol{x}|\hat{\theta})$

根据联合密度有：$$p(\boldsymbol{x}|\mathcal{X})=\int p(\boldsymbol{x},\theta|\mathcal{X})d\theta=\int p(\boldsymbol{x}|\theta)p(\theta|\mathcal{X})d\theta$$

其中$p(\boldsymbol{x}|\theta)$已知，则其关键在于求参数θ的后验密度$p(\theta|\mathcal{X})$。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯学习递推

利用θ的先验分布 $p(\theta)$及样本提供的信息求出 θ 的后验分布 $p(\theta|\mathcal{X})$，然后求类概率密度函数 $p(\boldsymbol{x}|\mathcal{X})$

假定样本集$\mathcal{X}^N=\{x_1, x_2, \ldots, x_N\}$是独立抽取的 ω_i类的一组样本，根据贝叶斯公式：

$$p\left(\theta|\mathcal{X}^N\right)=\frac{p\left(\mathcal{X}^N|\theta\right)p\left(\theta\right)}{\int p\left(\mathcal{X}^N|\theta\right)p\left(\theta\right)d\theta}$$

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

其中 $\mathcal{X}^{N-1}=\{x_1, x_2, \ldots, x_{N-1}\}$

有递推后验概率公式：$$p(\theta|\mathcal{X}^N)=\frac{p(x_N|\theta)p(\theta|\mathcal{X}^{N-1})}{\int p(x_N|\theta)p(\theta|\mathcal{X}^{N-1})d\theta}$$

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯学习递推

递推后验概率公式：$p(\theta|\mathcal{X}^N)=\dfrac{p(x_N|\theta)p(\theta|\mathcal{X}^{N-1})}{\int p(x_N|\theta)p(\theta|\mathcal{X}^{N-1})d\theta}$ —— 迭代计算式

设 $p\left(\theta|\mathcal{X}^0\right)=p\left(\theta\right)$

则随着样本数增多，可得后验概率密度函数序列：$p(\theta)$，$p(\theta|\boldsymbol{x}_1)$，$p(\theta|\boldsymbol{x}_1,\boldsymbol{x}_2),...$

—— 参数估计的递推贝叶斯方法

如果此序列收敛于以真实参数值为中心的 δ 函数，则把这一性质称作贝叶斯学习

如果 样本数 $N\to\infty$，则总体分布 $p(\boldsymbol{x}|\mathcal{X})$ 与类概率密度函数确切相等，且参数估计量 $\hat{\theta}$ 就是真实参数 θ：

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

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯学习迭代式

迭代式的使用：

* 根据先验知识得到θ的先验概率密度函数的初始估计 $p(\theta)$ 相当于 $N=0$ 时（ $\mathcal{X}^N=\mathcal{X}^0$ ）密度函数的一个估计

* 用 $\boldsymbol{x}_1$ 对初始的 $p(\theta)$ 进行修改。 $$p(\theta\mid\mathcal{X}^1)=p(\theta\mid\boldsymbol{x}_1)=\frac{p(\boldsymbol{x}_1\mid\theta)p(\theta)}{\int p(\boldsymbol{x}_1\mid\theta)p(\theta)d\theta}$$

* 给出$\boldsymbol{x}_2$，对用$\boldsymbol{x}_1$估计的结果进行修改。

$$p(\theta\mid\mathcal{X}^2)=p(\theta\mid\boldsymbol{x}_1,\boldsymbol{x}_2)=\frac{p(\boldsymbol{x}_2\mid\theta)p(\theta\mid\mathcal{X}^1)}{\int p(\boldsymbol{x}_2\mid\theta)p(\theta\mid\mathcal{X}^1)d\theta}$$

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

贝叶斯学习迭代式

迭代式的使用：

* 逐次给出$\boldsymbol{x}_3$，$\boldsymbol{x}_4$，…，$\boldsymbol{x}_N$，得到 $$p(\theta \mid \mathcal{X}^N) = \frac{p(x_N \mid \theta) p(\theta \mid \mathcal{X}^{N-1})}{\int p(\boldsymbol{x}_N \mid \theta) p(\theta \mid \mathcal{X}^{N-1}) d\theta}$$

$p(\boldsymbol{x} \mid \omega_i)$直接由 $p\left(\theta \mid \mathcal{X}^N\right)$计算得到，写为 $p\left(\boldsymbol{x} \mid \mathcal{X}^N\right)$：

$$p\left(\boldsymbol{x} \mid \mathcal{X}^N\right) = \int p(\boldsymbol{x}, \theta \mid \mathcal{X}^N) d\theta = \int p(\boldsymbol{x} \mid \theta) p(\theta \mid \mathcal{X}^N) d\theta$$

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯估计

单变量正态分布、方差σ^2已知，待估计量为均值μ。

设ω_i类：$p(x|\mu)\sim N(\mu,\sigma^2)$，$\mu$为未知随机参数。

可以合理地假定μ服从正态分布，设μ的先验概率密度$p(\mu)$：

$$p(\mu)\sim N(\mu_0,\sigma_0^2)$$ 其中μ_0和σ_0^2已知

对于平均误差损失函数的贝叶斯估计

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

即：

$$\hat{\mu}=\int \mu p(\mu|X)d\mu$$ 因此转为求μ的后验分布 $p(\mu|X)$

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯估计

设 $\mathcal{X}^N=\{x_1,x_2,...,x_N\}$ 是 ω_i 类的 N 个独立抽取的样本，根据贝叶斯公式，μ的后验概率为：

$$p\left(\mu\mid\mathcal{X}^N\right)=\frac{p(\mathcal{X}^N\mid\mu)p(\mu)}{\int p(\mathcal{X}^N\mid\mu)p(\mu)d\mu}$$

其中 $p(\mathcal{X}^N\mid\mu)=\prod_{k=1}^{N}p(x_k\mid\mu)$，有：

$$p(\mu\mid\mathcal{X}^N)=\alpha\prod_{k=1}^{N}p(x_k\mid\mu)p(\mu)$$

其中 $\alpha=1\Big/\int p(\mathcal{X}^N\mu)p(\mu)d\mu$，为与$\mu$无关的比例因子。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯估计

由于 $p(x\mid\mu)\sim N(\mu,\sigma^2)$ $\quad$ $p(\mu)\sim N(\mu_0,\sigma_0^2)$

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

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

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

式中，α' 和 α'' 为与 μ 无关的项。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯估计

$$p(\mu|\mathcal{X}^N)=\alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}x_k+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

考虑到 $p(\mu|\mathcal{X}^N)$ 仍是一个正态密度，即 $p(\mu|\mathcal{X}^N)\sim N(\mu_N,\sigma_N^2)\Rightarrow p(\mu|\mathcal{X}^N)=\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right\}$

通过待定系数法，可求得：

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

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

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯估计

平均误差损失函数的贝叶斯估计 $\hat{\mu}=\int \mu\, p\left(\mu \mid X^{N}\right) d\mu=\int \mu \frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right] d\mu=\mu_N$

即 $\hat{\mu}=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2} m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$ 式中 $m_N=\frac{1}{N}\sum_{k=1}^{N} x_k$

当先验分布 $p(\mu)$为标准正态分布$N(\mu_0,\sigma_0^2)$=$N(0,1)$，且总体分布的方差σ^2为1，则

$$\hat{\mu}=\frac{1}{N+1}\sum_{k=1}^{N} x_k$$

—— 与最大似然估计形式类似

贝叶斯学习（Bayesian Learning）

正态分布密度函数的贝叶斯学习

贝叶斯学习：求出 μ 的后验分布 $p\left(\mu|\mathcal{X}^N\right)$ 后，直接推断总体分布 $p\left(\boldsymbol{x}|\mathcal{X}^N\right)$

μ 的后验分布计算同前：$p(\mu|\mathcal{X}^N)=\dfrac{1}{\sqrt{2\pi}\sigma_N}\exp\left\{-\dfrac{1}{2}\left(\dfrac{\mu-\mu_N}{\sigma_N}\right)^2\right\}$

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

式中，$m_N=\dfrac{1}{N}\sum_{k=1}^{N}x_k$

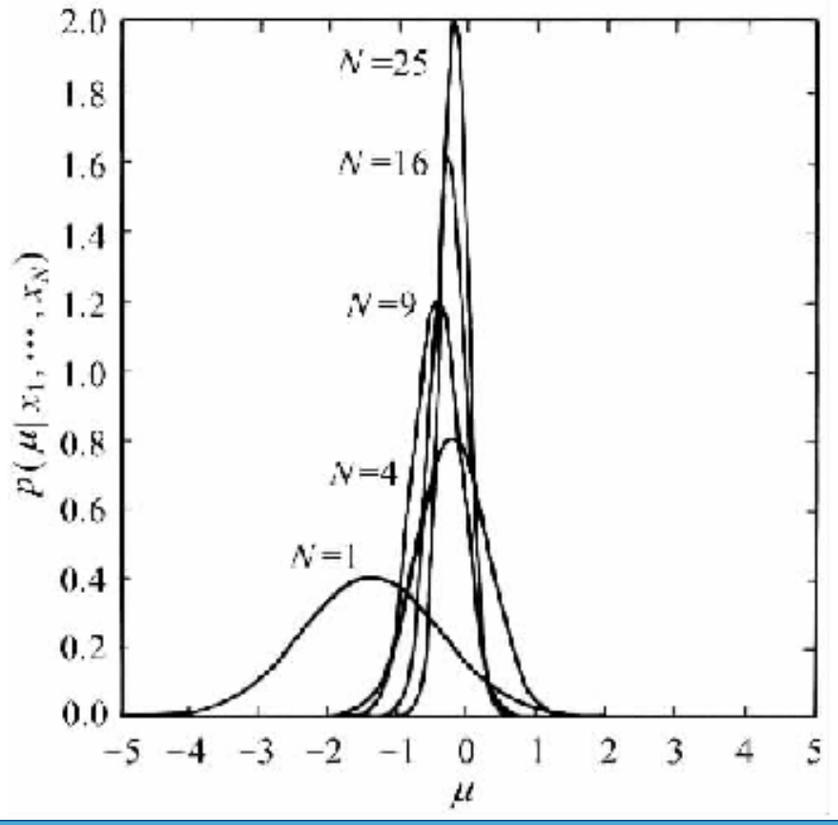
$$\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

μ_N：观察了 N 个样本后对 μ 的最好估计；

σ_N^2：估计的不确定性。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯学习过程示意图



当 N 增大时：
$p(\mu|\mathcal{X}^N)$ 越来越尖峰突起，
σ_N^2 单调减小；

当 N 趋于无穷时：
$p(\mu|\mathcal{X}^N)$ 趋于 δ 函数，
σ_N^2 趋于零。

贝叶斯估计和贝叶斯学习（Bayesian Estimation and Bayesian Learning）

正态分布密度函数的贝叶斯学习

由后验概率密度 $p(\mu|\mathcal{X}^N)$ 计算类概率密度函数 $p\left(\boldsymbol{x}|\mathcal{X}^N\right)$：

$$p\left(\boldsymbol{x}|\mathcal{X}^N\right)=\int p(\boldsymbol{x}|\mu)p(\mu|\mathcal{X}^N)d\mu=\int\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\cdot\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]d\mu$$

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

可见：　$p\left(\boldsymbol{x}|\mathcal{X}^N\right)$ 是正态分布；

均值为 μ_N；

方差为 $(\sigma^2+\sigma_N^2)$。

最大似然估计、贝叶斯估计和贝叶斯学习 小结

最大似然估计思想：参数 $\boldsymbol{\theta}$ 是类概率密度函数 $p(\boldsymbol{x}|\omega_i)$ 的一个确定的但未知的参数或参数集；根据一组属于类 ω_i 的已知样本，通过最大化似然函数来求解参数 $\boldsymbol{\theta}$ 。

贝叶斯估计和贝叶斯学习：将未知参数 $\boldsymbol{\theta}$ 看作为具有某种已知先验分布的随机变量，即参数 $\boldsymbol{\theta}$ 符合一定的先验分布，然后通过贝叶斯规则将参数的先验分布转化为后验概率进行求解。

讨论：

1）最大似然估计和贝叶斯估计在训练样本趋近于无穷时结果是一致的；

2）在实际应用中训练样本有限：

（a）计算复杂度：最大似然估计只需微分运算，而贝叶斯估计则需用到复杂的多重积分；

（b）准确性：训练样本有限情况下贝叶斯估计误差更小。

- **贝叶斯决策理论**
  基本概念，决策准则，概率论基础，…
- **最小错误/最小风险贝叶斯决策**
  最小错误率决策规则、条件风险，期望风险， …
- **概率密度函数估计**
  基本概念，最大似然估计…
- **贝叶斯估计和贝叶斯学习**
  贝叶斯估计，贝叶斯学习 …
- **非参数估计方法**
  基本概念，Parzen窗法 …
- **贝叶斯分类器**
  贝叶斯网络，朴素贝叶斯分类器 …

概率密度函数的非参数估计（nonparametric estimation）

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**。

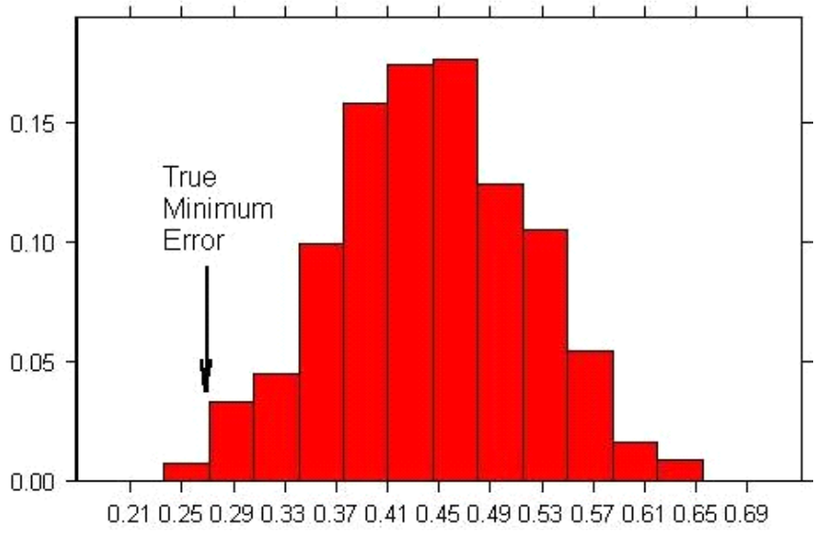
参数估计：需要事先假定一种分布函数，利用样本数据估计其参数。又称作**基于模型的方法。**

直方图方法：非参数概率密度估计的最简单方法

(1) 把 $\boldsymbol{x}$ 的每个分量分成 k 个等间隔小窗（若 $x \in E^d$，则形成 k^d 个小舱）；

(2) 统计落入各小舱内的样本数 q_i；

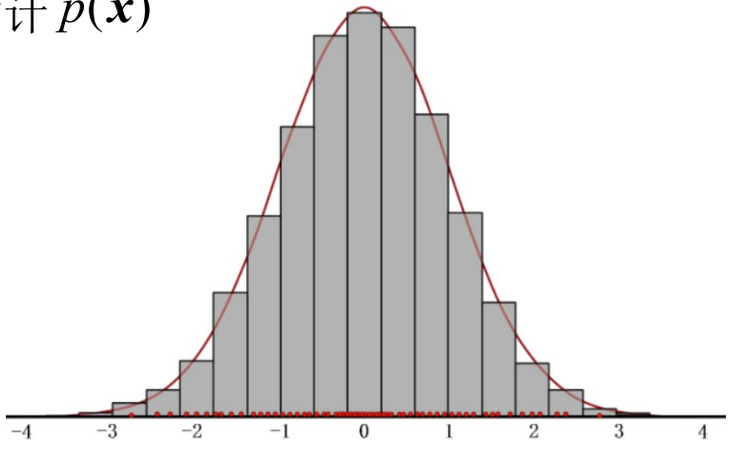
(3) 相应小舱的概率密度为 $q_i/(NV)$，其中N为样本总数，V为小舱体积。



概率密度函数的非参数估计（nonparametric estimation）

基本原理(不严格的说明)

问题：已知样本集 $\mathcal{X}=\{\boldsymbol{x}_1,...,\boldsymbol{x}_N\}$，其中样本均从服从 $p(\boldsymbol{x})$ 的总体中独立抽取，求估计 $\hat{p}(\boldsymbol{x})$



1. 出发点：基于事实

随机向量 $\boldsymbol{x}$ 落入区域 $\boldsymbol{R}$ 的概率P为 $P_R=\int_R p(\boldsymbol{x})d\boldsymbol{x}$ 。　$p(\boldsymbol{x})$：类概率密度函数。

$\mathcal{X}$ 中有k个样本落入区域$\boldsymbol{R}$的概率 $P_k=C_N^k P_R^k\left(1-P_R\right)^{N-k}$ ，N为样本总数。

区域 $\boldsymbol{R}$ 内有k个样本的概率 $\hat{P}_R=\dfrac{k}{N}$

设$p(x)$连续，且$\boldsymbol{R}$足够小， $\boldsymbol{R}$的体积为V，即$p(x)$在$\boldsymbol{R}$中没有变化，x是$\boldsymbol{R}$中的点。有：

$$P_R=\int_R p(\boldsymbol{x})d\boldsymbol{x}=p(\boldsymbol{x})\cdot V \Rightarrow \hat{p}\left(x\right)=\frac{k}{NV}$$ ——$\boldsymbol{x}$点概率密度的估计

要满足上式，必须要体积V趋于零，至使 $p(x)$ 在$\boldsymbol{R}$中没有变化。

概率密度函数的非参数估计（nonparametric estimation）

基本原理(不严格的说明)

存在的两个问题

1）固定体积V，样本数N增多，则k/N在概率上收敛，但只能得到在体积V中的平均估计，估计过于粗糙。

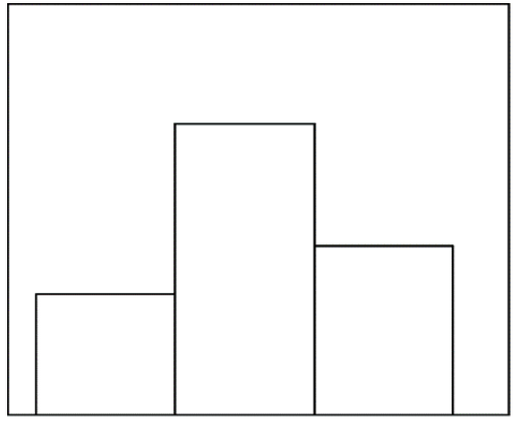
2）样本数N固定，体积V趋于零，可能导致某些区域中无样本，估计没有意义。

必须注意V、k、k/N 随N变化的趋势和极限，保持合理性。

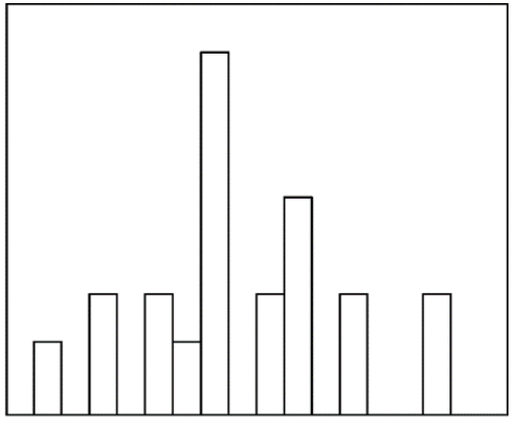
实际上，样本数总是有限的，所以体积 V 也不允许任意小。

因此采用这种估计的话， k/N和 $\hat{p}(x)$ 存在随机性，

即 k/N 和 $\hat{p}(x)$ 有一定的方差。



(a) 小窗过宽



(b) 小窗过窄

概率密度函数的非参数估计（nonparametric estimation）

估计步骤：

* 构造一串包含$\boldsymbol{x}$的区域R_1，R_2，…，R_N，…

* 假定V_N是R_N的体积，k_N是落入R_N内的样本数目，$\hat{p}_N(\boldsymbol{x})$是$p(\boldsymbol{X})$的第$N$次估计，有

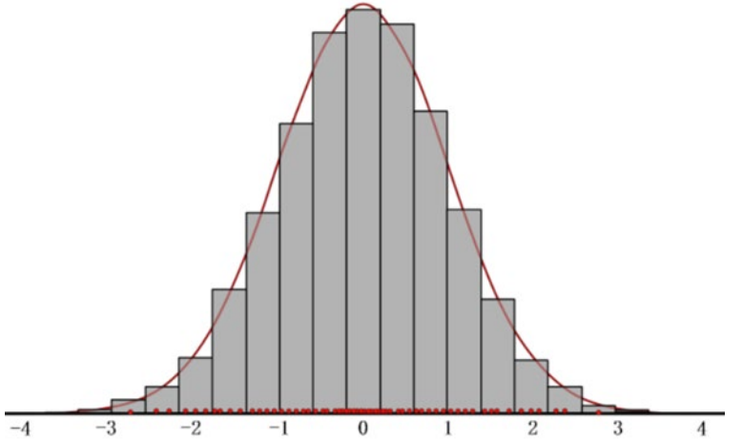
$$\hat{p}_N(\boldsymbol{x}) = \frac{k_N}{NV_N}$$

为保证估计合理性，即 $\hat{p}_N(\boldsymbol{x})$ 收敛于 $p(\boldsymbol{x})$，应满足的三个条件：

1） $\lim_{N\to\infty} V_N = 0$ ——— $\hat{p}_N(\boldsymbol{x})$ 能代表$\boldsymbol{x}$点的密度$p(\boldsymbol{x})$。

2） $\lim_{N\to\infty} k_N = \infty$ ——— 使式右边能收敛于概率$p(\boldsymbol{x})$

3） $\lim_{N\to\infty} k_N / N = 0$ ——— 落入R_N中的样本数始终是总数中的极小部分



概率密度函数的非参数估计（nonparametric estimation）

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

满足上述三个条件的区域序列通常有两种方法：

（a）Parzen窗法

固定V_N（如 $V_N = 1/\sqrt{N}$），同时对 k_N 和k_N/N 加限制以保证收敛。

（b）k_N-近邻法

固定k_N（如 $k_N = \sqrt{N}$ ），V_N 为正好包含k_N个近邻样本。

概率密度函数的非参数估计（nonparametric estimation）

Parzen窗估计的基本概念

设区域R_N：d维超立方体，棱长：h_N，则 $V_N = h_N^d$

定义窗函数$\varphi(u)$：

$$\varphi(u) = \begin{cases} 1, & \text{当} |u_j| \le \frac{1}{2}; \quad j = 1,2,...,d \\ 0, & \text{其它} \end{cases}$$

以原点为中心的超立方体

当$\boldsymbol{x}_i$落入以$\boldsymbol{x}$为中心，体积为V_N的超立方体时：

$$\varphi(u) = \varphi\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h_N}\right) = 1$$

否则 $\varphi(u) = 0$

概率密度函数的非参数估计（nonparametric estimation）

Parzen窗估计的基本概念

落入超立方体内的样本数为 $k_N = \sum_{i=1}^{N} \varphi\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h_N}\right)$

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

—— Parzen窗法基本公式

$\hat{p}_N(\boldsymbol{x})$ 为密度函数应满足的两个条件：

（1）$\varphi(u) \geq 0$ 保证 $\hat{p}_N(\boldsymbol{x}) \geq 0$

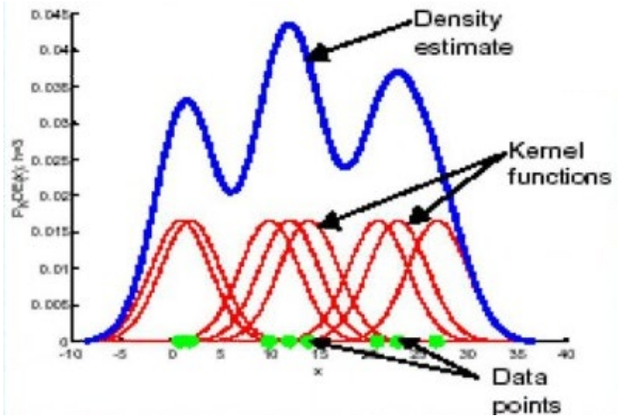
（2）$\int \varphi(u)du = 1$ 保证 $\int \hat{p}_N(\boldsymbol{x})dx = 1$

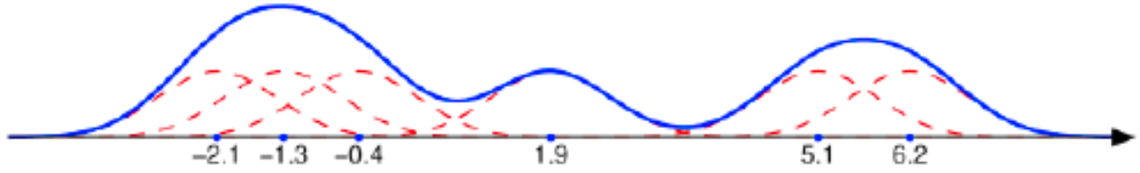
Parzen窗估计的几何意义：

定义核函数

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

$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_i)$





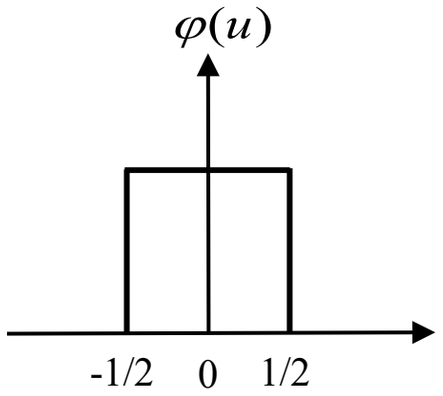
窗函数（核函数）$k(\boldsymbol{x}, \boldsymbol{x}_i)$，反映$x_i$对$p(x)$的贡献，实现小区域选择。

概率密度函数的非参数估计（nonparametric estimation）

Parzen窗的选择

1）方窗函数

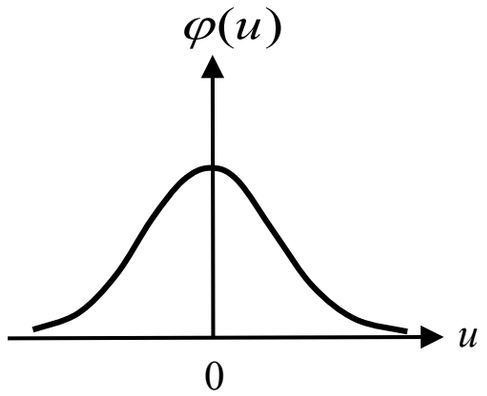
$$\varphi(u)=\begin{cases}1, & |u|\le\frac{1}{2}\\ 0, & \text{其它}\end{cases}$$



（a）

2）正态窗函数

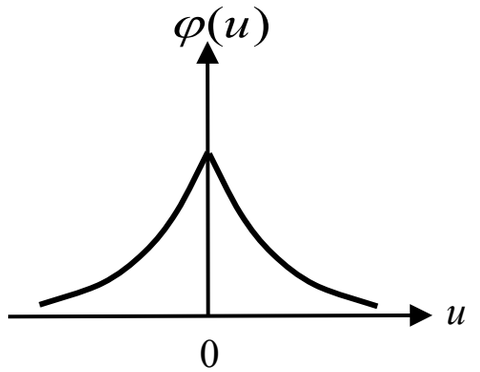
$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}u^2\}$$



（b）

3）指数窗函数

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



（c）

概率密度函数的非参数估计（nonparametric estimation）

Parzen窗宽的影响

例　设待估计的 $p(\boldsymbol{x})$是均值为零，方差为 1 的正态密度函数。随机地抽取含有 1 个、16 个、256 个学习样本 x_i 的样本集。试分别根据这三个样本集用 Parzen 窗法估计 $p(\boldsymbol{x})$，即求估计式 $\hat{p}_N(\boldsymbol{x})$。

解：考虑 $\boldsymbol{x}$ 是一维模式向量的情况。选择正态窗函数

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

并设 $h_N=h_1/\sqrt{N}$，h_1 为可调节的参数。

有 $$\varphi\left(\frac{x-x_i}{h_N}\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-x_i}{h_N}\right)^2\right]$$

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

概率密度函数的非参数估计

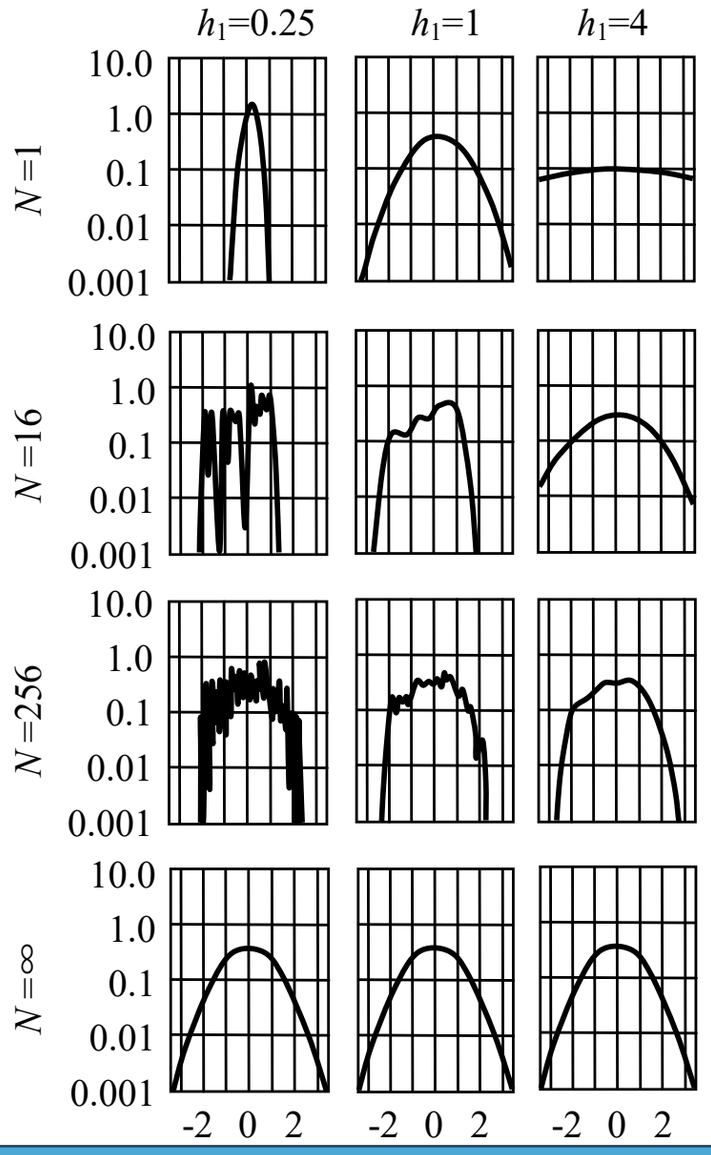
Parzen窗宽的影响

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

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

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

估计结果：



(1) 当N=1，$\hat{p}_N(x)$ 是以单个样本为中心的正态形状。

(2) 当N=16，h_1=0.25 可看出各单个样本的贡献，随着增加，单个样本作用被模糊了。

(3) 随着N增加，估计量 $\hat{p}_N(x)$ 越来越接近真实值。

(4) 当 N 趋于无穷大时，$\hat{p}_N(x)$收敛于平滑的正态曲线。

概率密度函数的非参数估计

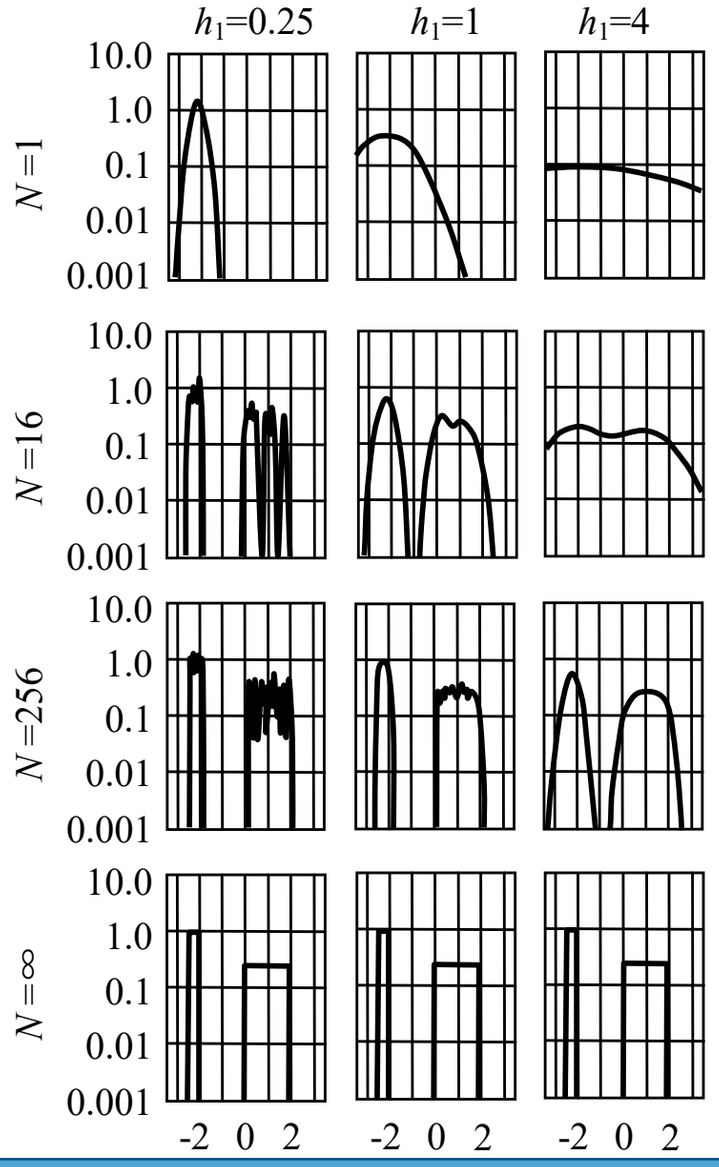
Parzen窗宽的影响

例　仍以一维情况为例，假定待估计的概率密度函数 $p(x)$ 为两个均匀分布密度的混合

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

随机抽取含 1 个、16 个、256 个学习样本的样本集，求 $p(x)$ 的估计 $\hat{p}_N(x)$。

解：估计结果



Parzen窗法特点：

- 具有一般性，适用于单峰、多峰形式。
- 要得到较精确的估计必须抽取大量的样本。
- 需要比参数估计所需样本数多的多的样本量。

概率密度函数的非参数估计（nonparametric estimation）

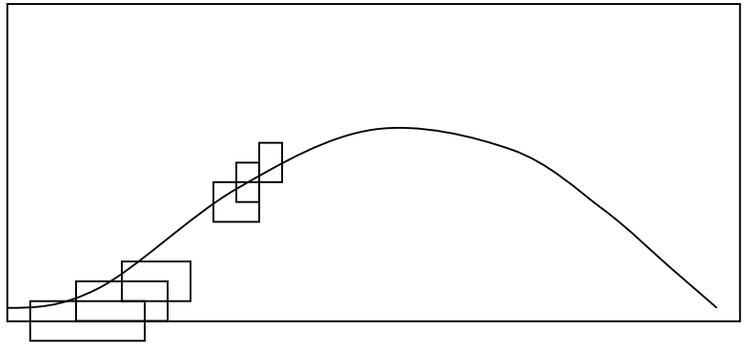
k_N-近邻法

确定k_N：总样本为N时每个小舱内的样本数

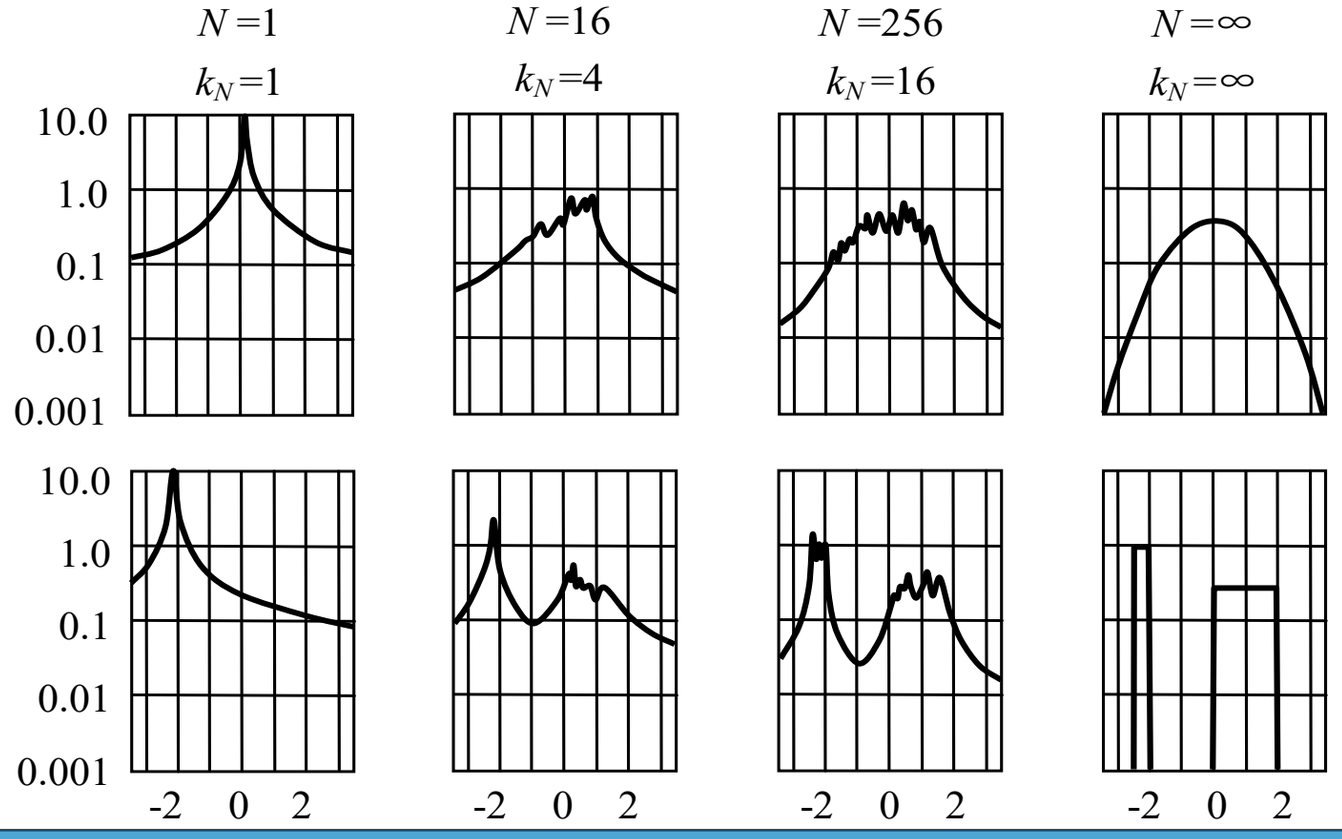
调整包含$\boldsymbol{x}$小舱体积，直到落入k_N个样本

$$\hat{p}_N(\boldsymbol{x}) = \frac{k_N}{NV_N}$$

通过控制小区域内的样本数 k_N来确定小区域大小



上两例中，用k_N-近邻法估计的$p(\boldsymbol{x})$ 的结果：

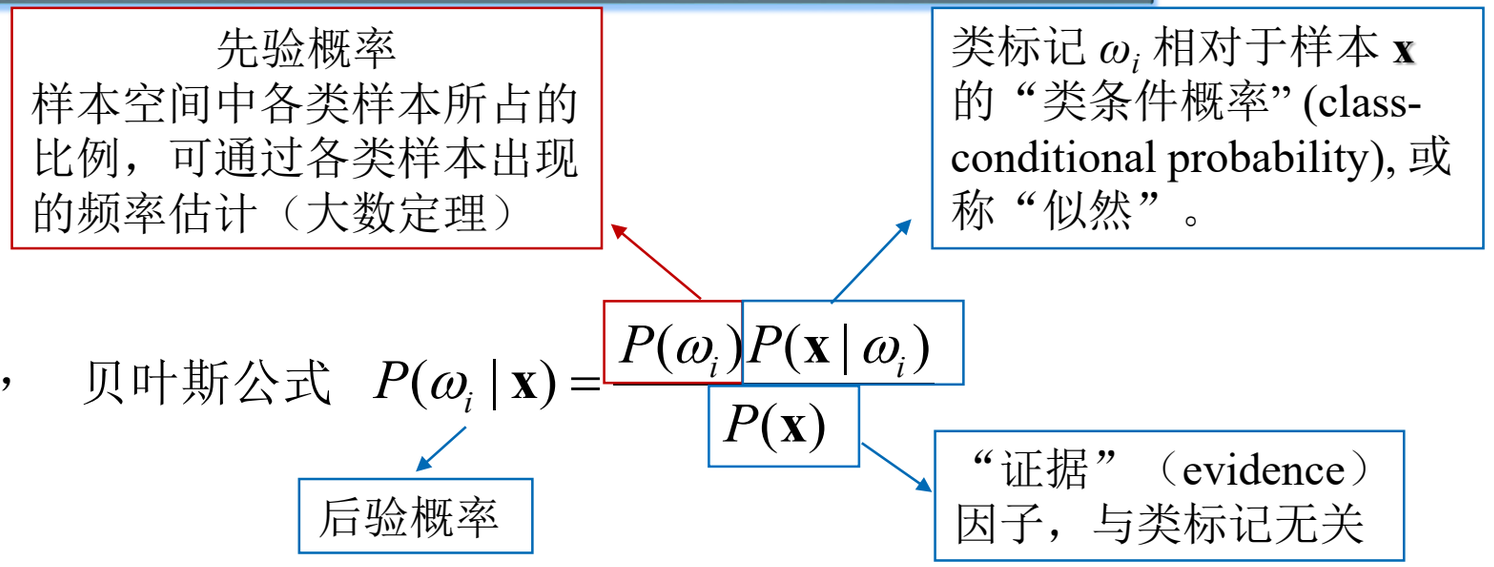


- **贝叶斯决策理论**
  基本概念，决策准则，概率论基础，…
- **最小错误/最小风险贝叶斯决策**
  最小错误率决策规则、条件风险，期望风险， …
- **概率密度函数估计**
  基本概念，最大似然估计…
- **贝叶斯估计和贝叶斯学习**
  贝叶斯估计，贝叶斯学习 …
- **非参数估计方法**
  基本概念，Parzen窗法 …
- **贝叶斯分类器**
  贝叶斯网络，朴素贝叶斯分类器 …

朴素贝叶斯分类器

估计后验概率 $P(\omega_i|\mathbf{x})$ 的主要困难：

类条件概率 $P(\mathbf{x}|\omega_i)$ 是所有属性上的联合概率，难以从有限的训练样本估计获得。



朴素贝叶斯分类器 (Naïve Bayes Classifier)之所以被称为「朴素」，是因为采用了“属性条件独立性假设” (attribute conditional independence assumption)：即每个属性独立地对分类结果发生影响。

优点：简单易懂、学习效率高、在某些领域的分类问题中能与决策树、神经网络相媲美。

缺点：由于该算法以自变量之间的独立（条件特征独立）性和连续变量的正态性假设为前提，就会导致算法精度在某种程度上受影响。

朴素贝叶斯分类器

贝叶斯定理：

$$P(A_k \mid B_1 B_2 \ldots B_n) = \frac{P(A_k) P(B_1 B_2 \ldots B_n \mid A_k)}{P(B_1 B_2 \ldots B_n)}$$

对于 $P(A_k|B_1B_2\ldots B_n)$，分母 $P(B_1B_2\ldots B_n)$ 为固定值。因此只需考虑 $P(A_k|B_1B_2\ldots B_n)$ 的大小；

根据朴素贝叶斯学习的假定，即$B_1, B_2, \ldots, B_n$ 相互独立，有

$$\begin{aligned} P(A_k \mid B_1 B_2 \ldots B_n) &= P(A_k) P(B_1 \mid A_k) P(B_2 \mid A_k) \cdots P(B_n \mid A_k) \\ &= P(A_k) \prod_i P(B_i \mid A_k) \end{aligned}$$

基于属性条件独立性假设，贝叶斯公式可重写为

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\omega_i) P(\mathbf{x} \mid \omega_i)}{P(\mathbf{x})} = \frac{P(\omega_i)}{P(\mathbf{x})} \prod_{j=1}^{N} P(x_j \mid \omega_i)$$

其中N为属性数目，x_j为 $\mathbf{x}$ 在第 j 个属性上的取值

朴素贝叶斯分类器（例）

某个医院早上来了六个门诊的病人，他们的情况如下表所示:

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

朴素贝叶斯分类器（例）

问题：第七个病人，是一个打喷嚏的建筑工人，问患感冒的概率有多大？

$$P(\text{感冒}|\text{打喷嚏，建筑工人})=\frac{P(\text{感冒})P(\text{打喷嚏，建筑工人}|\text{感冒})}{P(\text{打喷嚏，建筑工人})}$$

根据朴素贝叶斯条件独立性的假设可知，"打喷嚏"和"建筑工人"这两个特征是独立的，因此，

$$P(\text{感冒}|\text{打喷嚏，建筑工人})=\frac{P(\text{感冒})P(\text{打喷嚏}|\text{感冒})P(\text{建筑工人}|\text{感冒})}{P(\text{打喷嚏})P(\text{建筑工人})}$$

朴素贝叶斯分类器（例）

根据表格估计概率，

P(感冒)=3/6=0.5

P(打喷嚏|感冒)=2/3=0.66

P(建筑工人|感冒)=1/3=0.33

P(打喷嚏)=3/6=0.5

P(建筑工人)=2/6=0.33

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

$$P(\text{感冒}\mid\text{打喷嚏，建筑工人})=\frac{P(\text{感冒})P(\text{打喷嚏}\mid\text{感冒})P(\text{建筑工人}\mid\text{感冒})}{P(\text{打喷嚏})P(\text{建筑工人})}$$

$$=\frac{0.5\times0.66\times0.33}{0.5\times0.33}=0.66$$

朴素贝叶斯分类器

朴素贝叶斯分类器的训练器的训练过程就是基于训练集 D 估计类先验概率$P(\omega_i)$，并为每个属性估计类条件概率$P(x_j|\omega_i)$ 。

令 D_i 为训练集 D 中第 ω_i 类样本的集合，若有充足且独立同分布样本，则类先验概率为：$P(\omega_i)=\dfrac{|D_i|}{|D|}$

对离散属性而言，令 $D_{i,j}$ 为 D_i 中第 j个属性上取值为 x_j 的样本集合，则类条件概率为：$P(x_j\,|\,\omega_i)=\dfrac{|D_{i,j}|}{|D_i|}$

对连续属性而言可考虑概率密度函数，假定 $P(x_j\,|\,\omega_i)\sim\mathcal{N}\left(\mu_{i,j},\sigma_{i,j}^2\right)$，其中 $\mu_{i,j}$ 和 $\sigma_{i,j}^2$ 分别是第 i 类样本在第 j个属性上取值的均值和方差，则有

$$P(x_j\,|\,\omega_i)=\frac{1}{\sqrt{2\pi}\cdot\sigma_{i,j}}\exp\left(-\frac{\left(x_j-\mu_{i,j}\right)^2}{2\sigma_{i,j}^2}\right)$$

朴素贝叶斯分类器

例 用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

类先验概率：

$$P(\text{好瓜}=\text{是})=\frac{8}{17}\approx 0.471$$

$$P(\text{好瓜}=\text{否})=\frac{9}{17}\approx 0.529$$

为每个离散型属性估计类条件概率：

$$P_{\text{青绿}|\text{是}}=P(\text{色泽}=\text{青绿}\mid\text{好瓜}=\text{是})=\frac{3}{8}=0.375$$

$$P_{\text{青绿}|\text{否}}=P(\text{色泽}=\text{青绿}\mid\text{好瓜}=\text{否})=\frac{3}{9}\approx 0.333$$

$$P_{\text{蜷缩}|\text{是}}=P(\text{根蒂}=\text{蜷缩}\mid\text{好瓜}=\text{是})=\frac{5}{8}=0.375$$

$$P_{\text{蜷缩}|\text{否}}=P(\text{根蒂}=\text{蜷缩}\mid\text{好瓜}=\text{否})=\frac{3}{9}\approx 0.333$$

$$P_{\text{浊响}|\text{是}}=P(\text{敲声}=\text{浊响}\mid\text{好瓜}=\text{是})=\frac{6}{8}=0.750$$

$$P_{\text{浊响}|\text{否}}=P(\text{敲声}=\text{浊响}\mid\text{好瓜}=\text{否})=\frac{4}{9}\approx 0.444$$

$$P_{\text{清晰}|\text{是}}=P(\text{纹理}=\text{清晰}\mid\text{好瓜}=\text{是})=\frac{7}{8}=0.875$$

$$P_{\text{清晰}|\text{否}}=P(\text{纹理}=\text{清晰}\mid\text{好瓜}=\text{否})=\frac{2}{9}\approx 0.222$$

$$P_{\text{凹陷}|\text{是}}=P(\text{脐部}=\text{凹陷}\mid\text{好瓜}=\text{是})=\frac{6}{8}=0.750$$

$$P_{\text{凹陷}|\text{否}}=P(\text{脐部}=\text{凹陷}\mid\text{好瓜}=\text{否})=\frac{2}{9}\approx 0.222$$

$$P_{\text{硬滑}|\text{是}}=P(\text{触感}=\text{硬滑}\mid\text{好瓜}=\text{是})=\frac{6}{8}=0.750$$

$$P_{\text{硬滑}|\text{否}}=P(\text{触感}=\text{硬滑}\mid\text{好瓜}=\text{否})=\frac{6}{9}\approx 0.667$$

朴素贝叶斯分类器 例

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

为每个连续型属性估计类条件概率

$$p_{\text{密度: 0.697}|\text{是}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi}\cdot 0.129}\exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right) \approx 1.959$$

$$p_{\text{密度: 0.697}|\text{否}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.195}\exp\left(-\frac{(0.697-0.496)^2}{2\cdot 0.195^2}\right) \approx 1.203$$

$$p_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi}\cdot 0.101}\exp\left(-\frac{(0.460-0.279)^2}{2\cdot 0.101^2}\right) \approx 0.788$$

$$p_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.108}\exp\left(-\frac{(0.460-0.154)^2}{2\cdot 0.108^2}\right) \approx 0.066$$

$$P(\text{好瓜}=\text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.038,$$

$$P(\text{好瓜}=\text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5}.$$

结论:好瓜

拉普拉斯修正

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

改为“清脆”

若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题。

比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，即便其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然不合理。 $P_{清脆|是}=P(敲声=清脆|好瓜=是)=\frac{0}{8}=0$

为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“拉普拉斯修正”（Laplacian correction）

类先验概率 $P(\omega_i)=\frac{|D_i|}{|D|} \Rightarrow P(\omega_i)=\frac{|D_i|+1}{|D|+N}$

N为训练集D中可能的类别个数

类条件概率 $P(x_j|\omega_i)=\frac{|D_{i,j}|}{|D_i|} \Rightarrow P(x_j|\omega_i)=\frac{|D_{i,j}|+1}{|D_i|+N_j}$

N_j为第j个属性可能的取值数

此时，$P_{清脆|是}=P(敲声=清脆|好瓜=是)=\frac{0+1}{8+3}\approx 0.091$

贝叶斯网络

在日常生活中，人们往往进行常识推理，而这种推理通常是不准确的。包括人工智能中的专家系统，确定性推理系统等。

- 侦探推理
- “天黑请闭眼杀人游戏”等

局限：

- 简单、粗略的归纳和推理
- 缺少准确的量来刻画（简单逻辑运算）
- 缺少准确的统计特性（时间轴）

为了提高推理的准确性，人们引入了概率理论。最早由Judea Pearl于1988年提出了贝叶斯网络(Bayesian Network)。

贝叶斯网络是Bayes方法的扩展，是目前不确定知识表达和推理领域最有效的理论模型之一。

贝叶斯网络（Bayesian network），又称信念网络（belief network）或是有向无环图模型（directed acyclic graphical model），是一种概率图型模型。

贝叶斯网络：从应用实例开始——胸部疾病诊所

假想你是一名新毕业的医生，专攻肺部疾病。你决定建一个胸部疾病诊所，主治肺病及相关疾病。

大学课本已经中告诉你了肺癌、肺结核和支气管炎的发生比率以及这些疾病典型的临床症状、病因等，于是你就可以根据课本里的理论知识建立自己的Bayes网。

根据如下数据信息：

1）全国有30%的人吸烟.

2）每10万人中就有70人患有肺癌.

3）每10万人中就有10人患有肺结核.

4）每10万人中就有800人患有支气管炎.

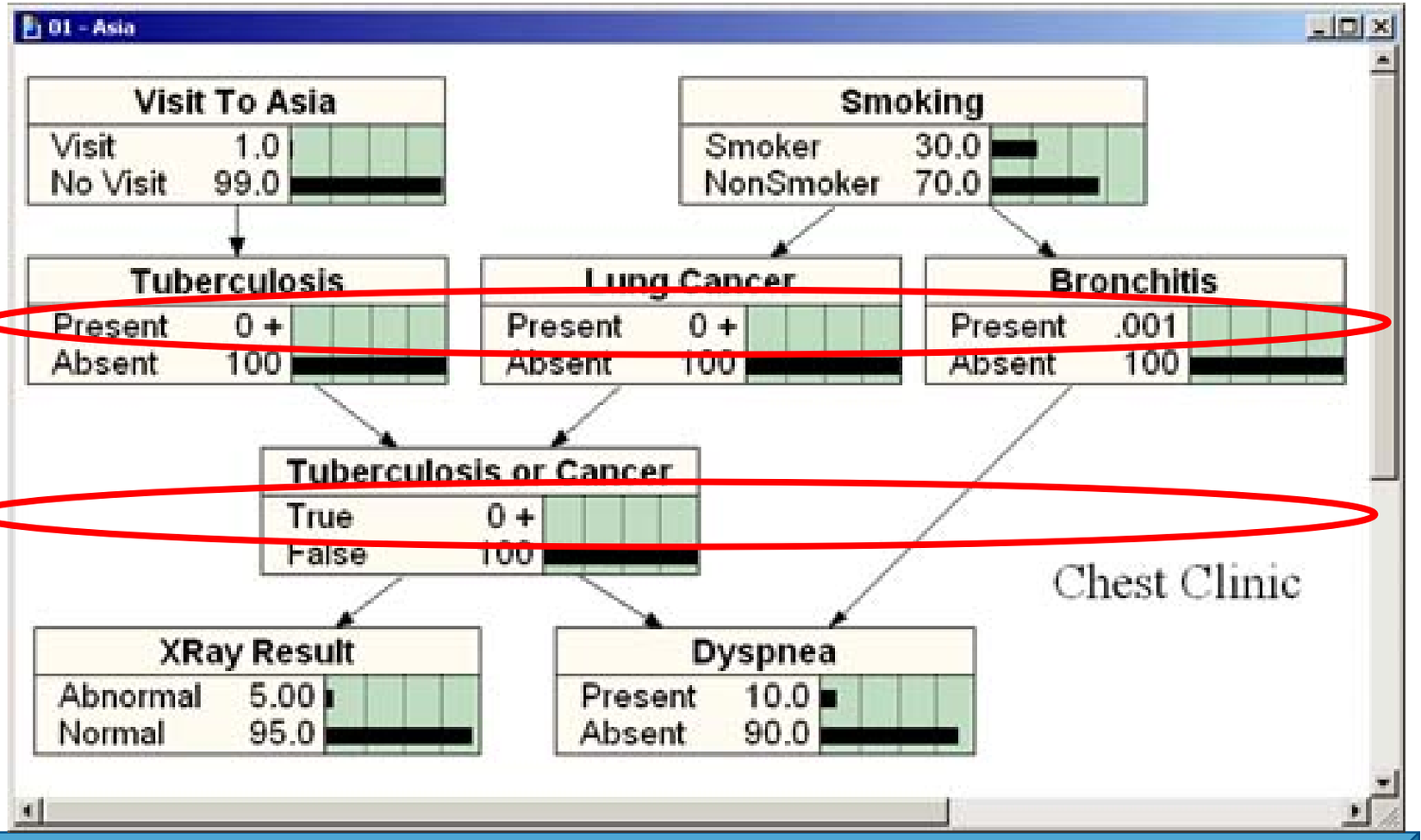
5）10%人存在呼吸困难症状, 大部分由哮喘、支气管炎和其他非肺结核、非肺癌性疾病引起.

根据上面的数据可以建立如下BN模型：

贝叶斯网络：从应用实例开始——胸部疾病诊所

全国有30%的人吸烟.
每10万人中有70人患有肺癌.
每10万人中有10人患有肺结核.
每10万人中有800人患有支气管炎.
10%人存在呼吸困难症状

这样的BN模型意义不大，因为它没有用到来你诊所病人的案例数据，不能反映真实病人的情况。



贝叶斯网络：从应用实例开始——胸部疾病诊所

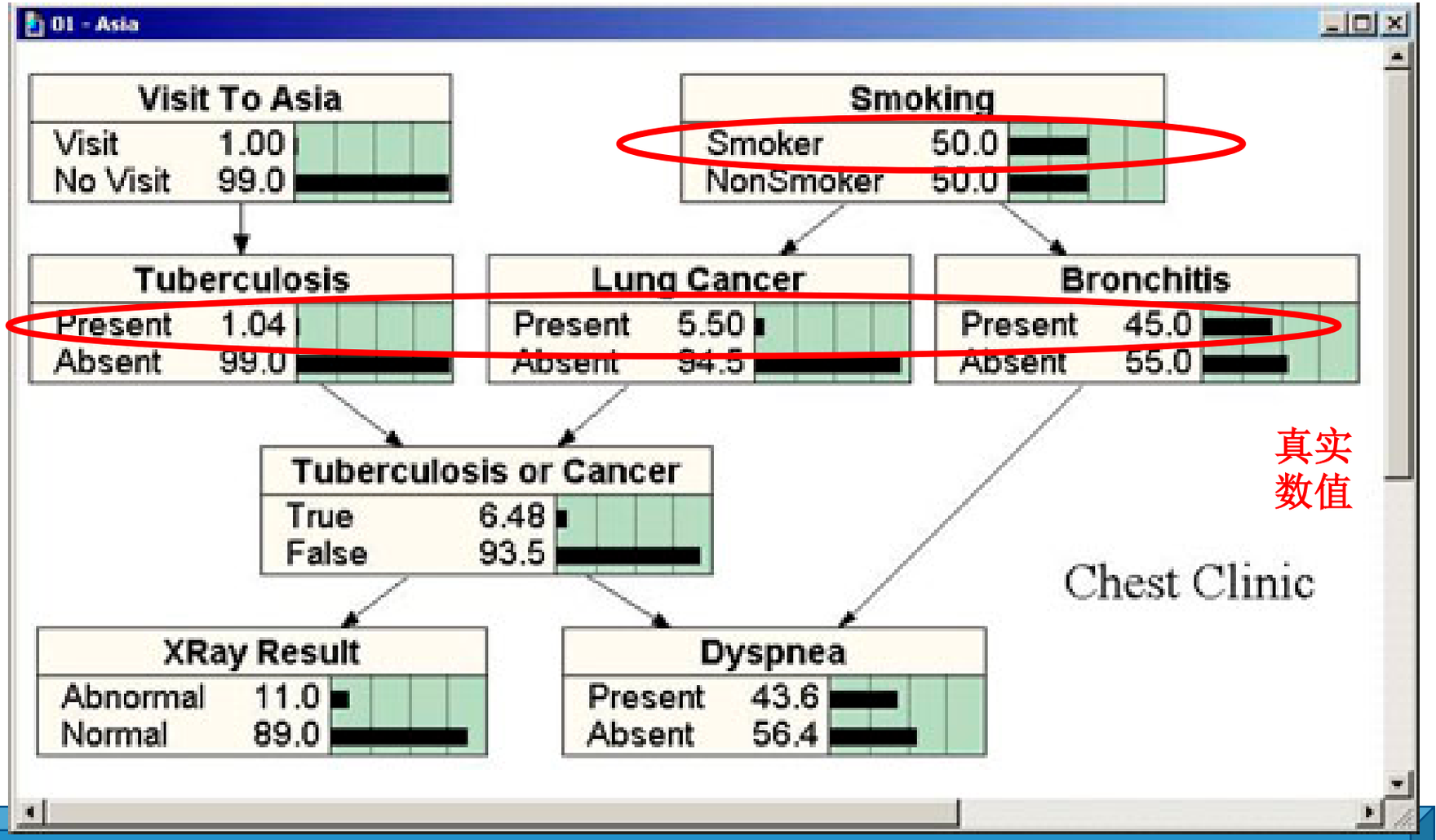
当诊所诊治了数千病人后，实际诊所数据显示：

50%的病人吸烟.

1%患有肺结核.

5.5% 得了肺癌.

45% 患有不同程度支气管炎.



贝叶斯网络：从应用实例开始——胸部疾病诊所

在诊断前没有患者的任何信息。

当我们向患者咨询信息时，BN网中的概率就会自动调整。

贝叶斯网络最强大之处在于从每个阶段结果所获得的概率都是数学与科学的反映。

换句话说，若我们了解患者的足够信息，根据这些信息获得统计知识，网络就会告诉我们合理的推断。

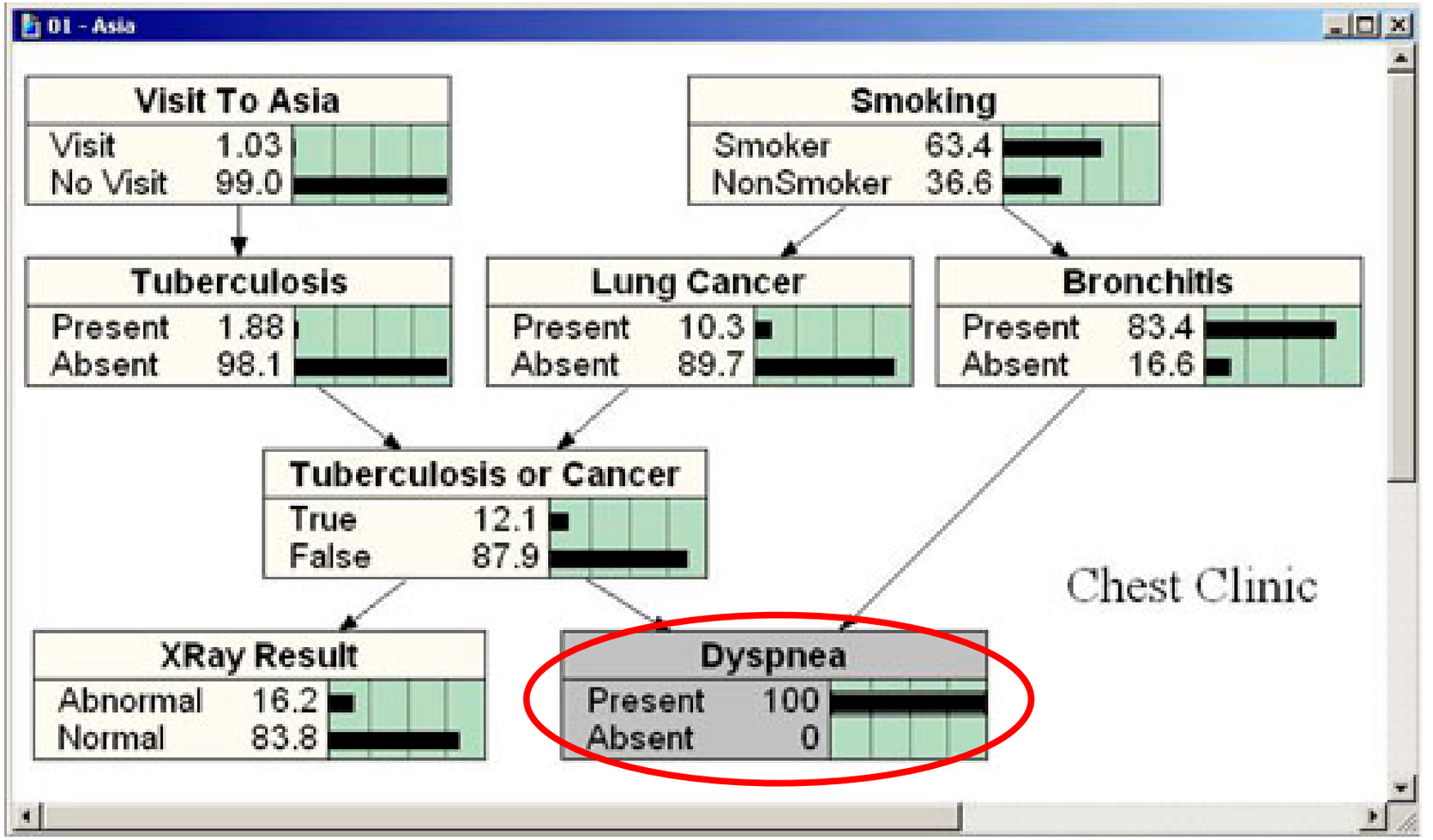
贝叶斯网络：从应用实例开始——胸部疾病诊所

当一个病人进入诊所，我们开始和他谈论：

他告诉我们她呼吸困难。

我们将这个信息输入到网络。我们相信病人的信息，认为其存在100%呼吸困难。

调整网络



贝叶斯网络：从应用实例开始——胸部疾病诊所

病人有呼吸困难症状，三种疾病的概率都增大了，因为这些疾病都有呼吸困难的症状。

明显增大的是支气管炎，从 45% 到 83.4%. 因为支气管炎病比癌症和肺结核更常见。

病人抽烟的几率也会随之增大，从50% 到63.4%.

近期访问过亚洲的几率也会增大: 从1% 到1.03%, 显然是不重要的.

X光照片不正常的几率也会上涨，从11% 到16%.

贝叶斯网络：从应用实例开始——胸部疾病诊所

目前只能说患有支气管炎的可能性很大。

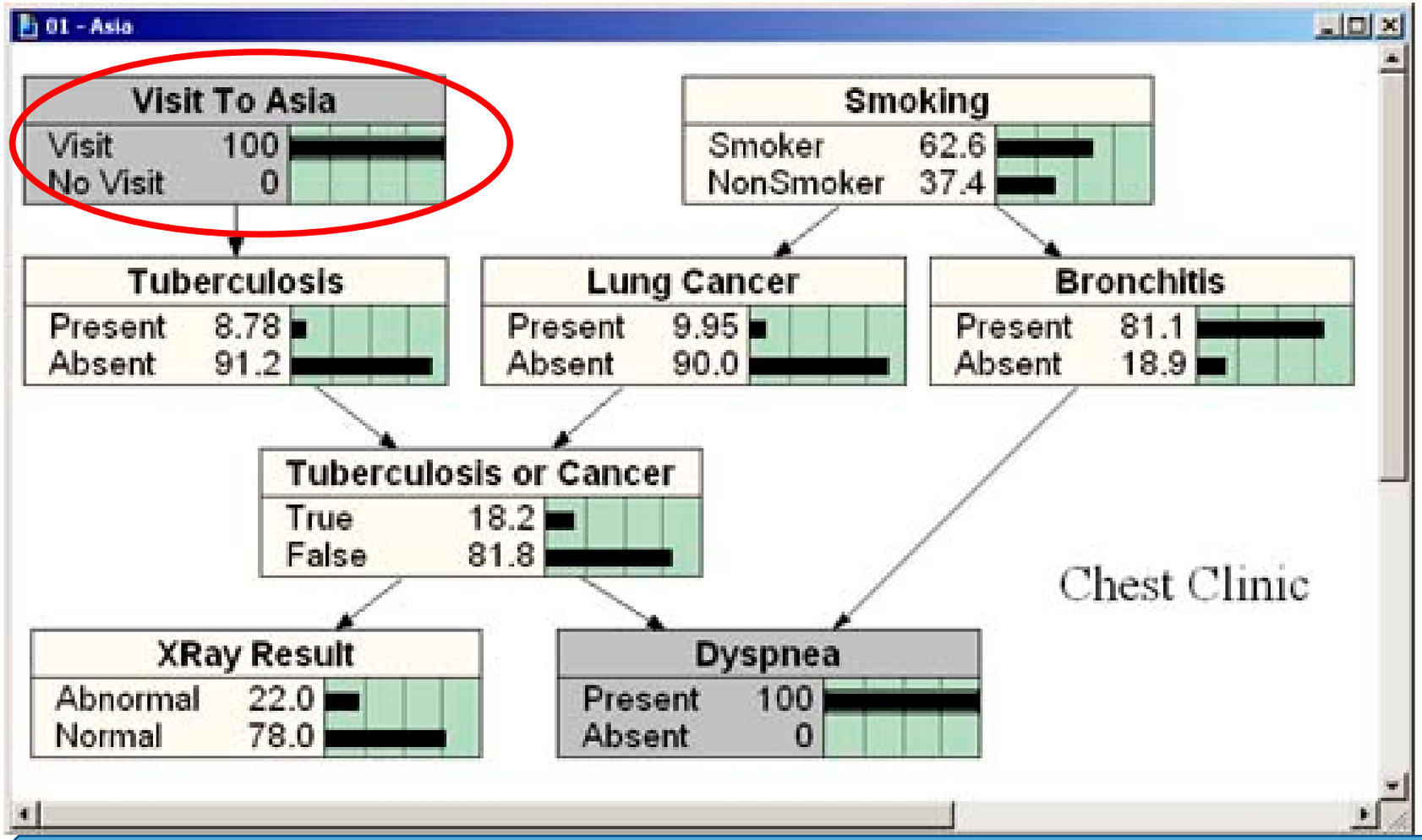
还应获得更多信息来确定我们的判断。

如果我们现在就主观定了病症，她可能得的是支气管炎，那我们就是一个烂医生。

需要更多信息来做最后的决定。

因此，我们按照流程依此问她一些问题，比如她最近是不是去过亚洲国家，吃惊的是她回答了“是”。现在获得的信息就影响了BN模型。

贝叶斯网络：从应用实例开始——胸部疾病诊所



患肺结核的几率显然增大，从2%到9%.

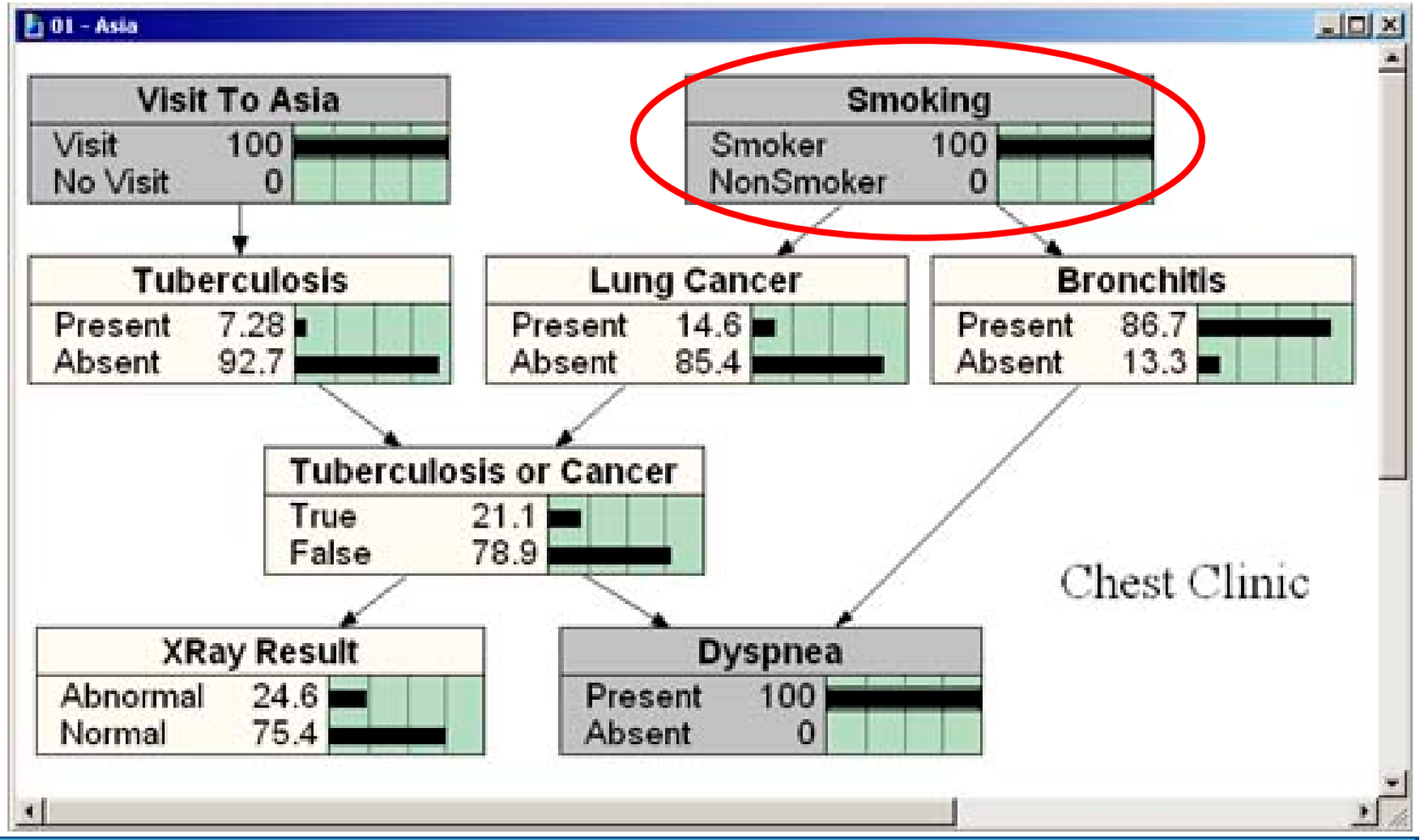
而患有癌症、支气管炎以及该患者是吸烟患者的几率都有所减少。为什么呢？因为此时呼吸困难的原因相对更倾向于肺结核。

贝叶斯网络：从应用实例开始——胸部疾病诊所

继续问患者一些问题，假设患者是个吸烟者，则网络变为：

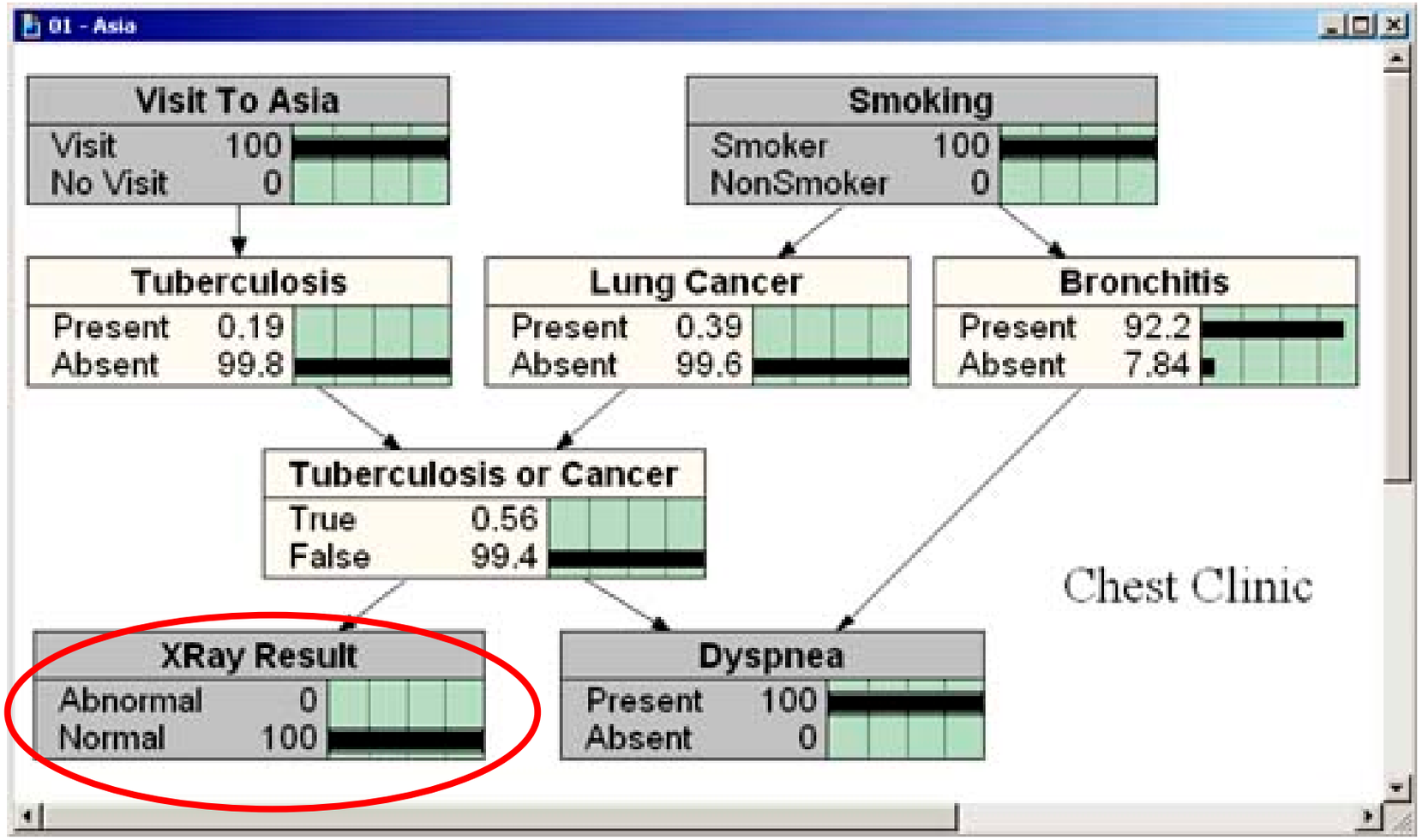
此时注意到最好的假设仍然是认为患者患有支气管炎。

为了确认我们要求她做一个X光透视。



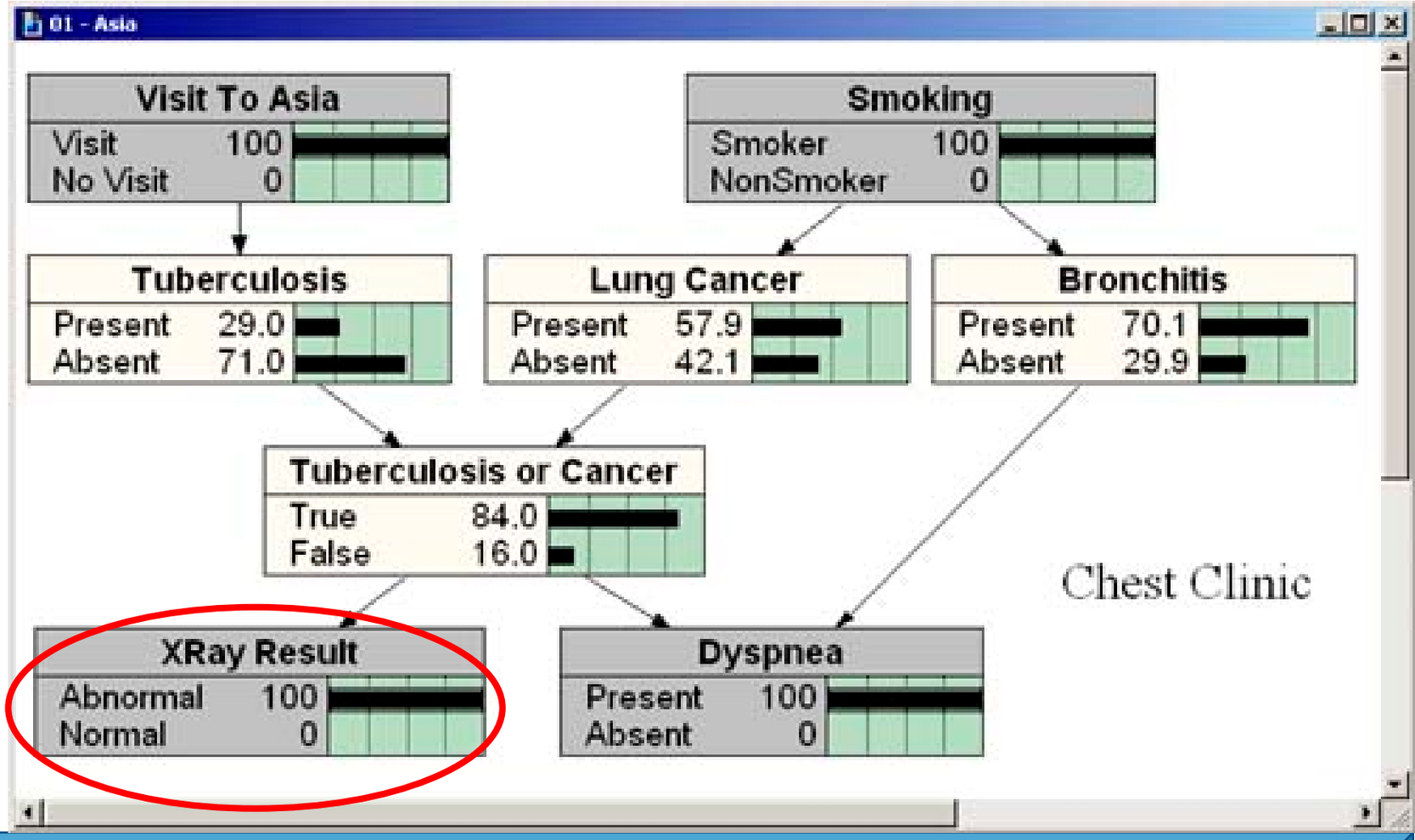
贝叶斯网络：从应用实例开始——胸部疾病诊所

若X光透视结果显示其正常，这就更加肯定我们的推断她患有支气管炎。



贝叶斯网络：从应用实例开始——胸部疾病诊所

若X光显示不正常，则结果将有很大不同。



贝叶斯网络的基本概念

贝叶斯网络：

一系列变量的联合概率分布的图形表示；

使用条件概率表 (Conditional Probability Table, CPT) 来表述属性的联合概率分布；

一个表示变量之间的相互依赖关系的数据结构；

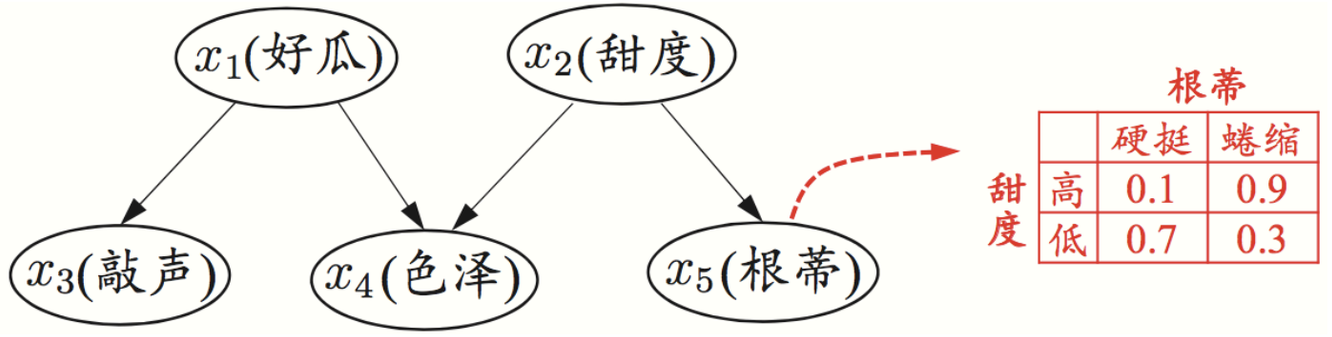
图论与概率论的结合；

借助有向无环图 (Directed Acyclic Graph, DAG) 来刻画属性间的依赖关系；

亦称“信念网”(brief network)，

贝叶斯网络的基本概念

例



根蒂

甜度	硬挺	蜷缩
高	0.1	0.9
低	0.7	0.3

西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

从网络图结构可以看出 -> “色泽”直接依赖于“好瓜”和“甜度”

从条件概率表可以得到 -> “根蒂”对“甜度”的量化依赖关系 $P(\text{根蒂}=\text{硬挺}|\text{甜度}=\text{高})=0.1$

贝叶斯网络（因果关系网络）

假设：

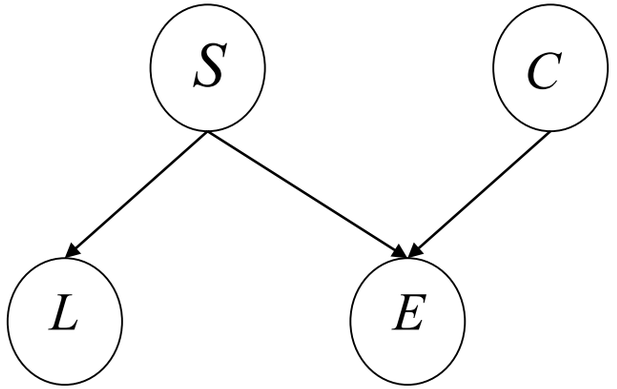
- 命题 S (smoker)：该患者是一个吸烟者
- 命题 C (coal Miner)：该患者是一个煤矿矿井工人
- 命题 L (lung Cancer)：他患了肺癌
- 命题 E (emphysema)：他患了肺气肿
- 由专家给定的假设可知，命题 S 对命题 L 和命题 E 有因果影响，而 C 对 E 也有因果影响。
- 命题之间的关系可以描绘成因果关系网。

因果关系网络 2

每一个节点代表一个证据。

每一条弧代表一条规则 (假设)。

连接结点的弧表达了由规则给出的节点间的直接因果关系。



节点 S，C 是节点 L 和 E 的父节点或称双亲节点。

L，E也称为是 S 和 C 的子节点或称后代节点。

贝叶斯网就是一个在弧的连接关系上加入连接强度的因果关系网络 。

贝叶斯网络举例

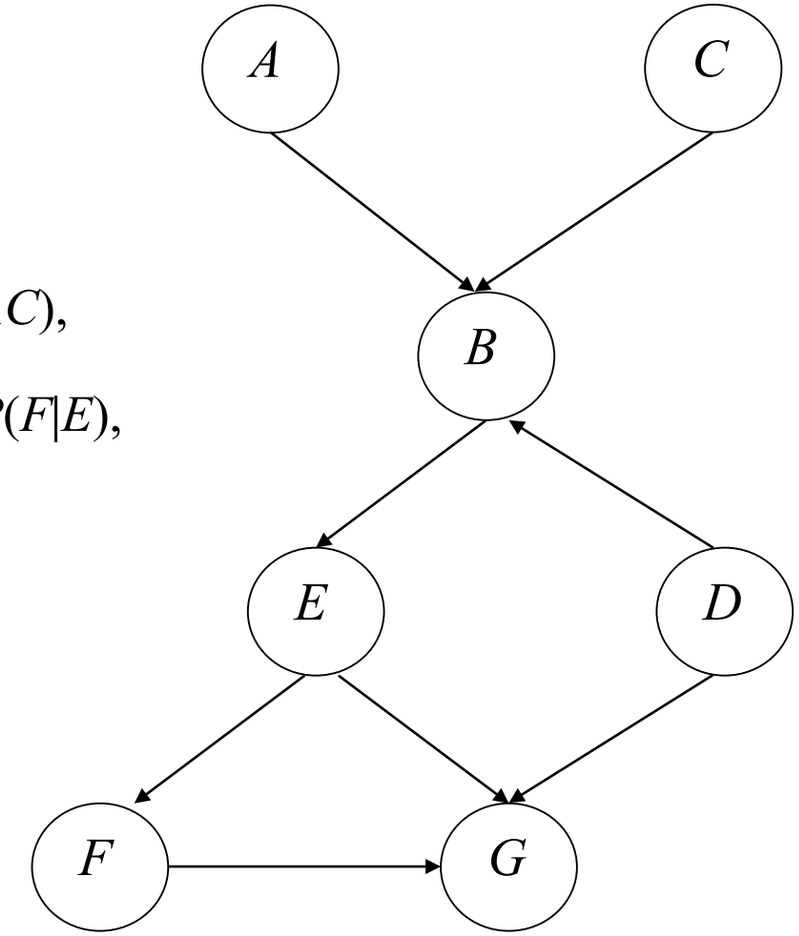
无环图

指定概率值

$P(A)$, $P(B)$, $P(B|AC)$,

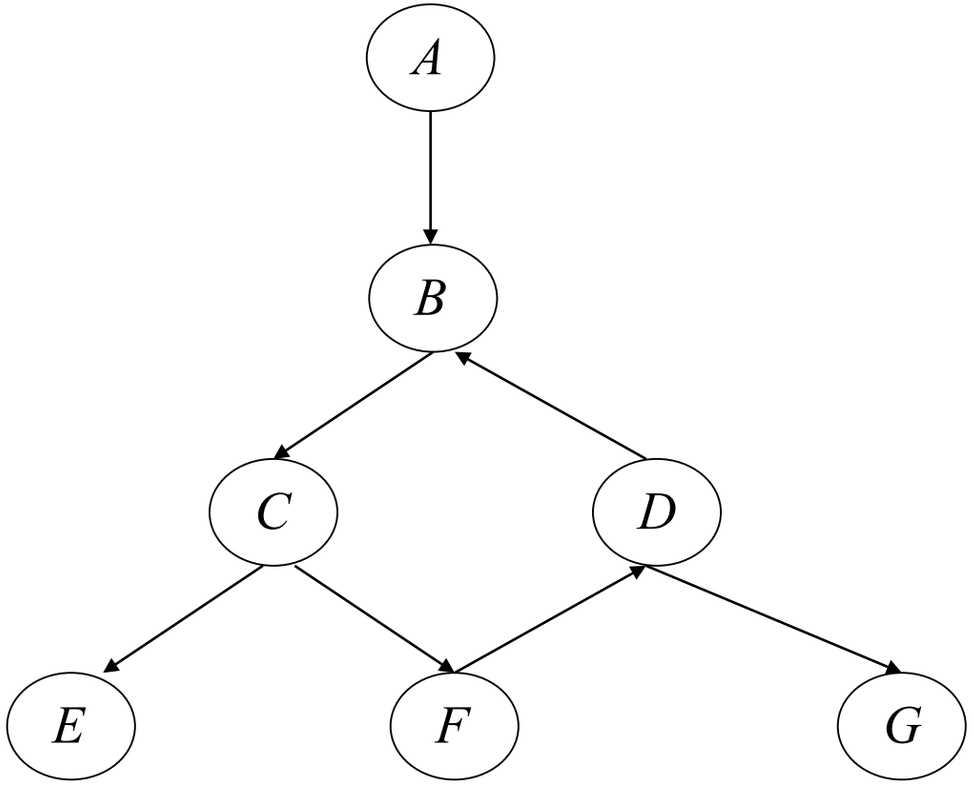
$P(E|B)$, $P(B|D)$, $P(F|E)$,

$P(G|DEF)$



有环图

相互嵌套，无法确定



贝叶斯网络结构

包含两个部分：

- 贝叶斯网络结构图是一个有向无环图（DAG: Directed Acyclic Graph）。
  - 其中图中的每个节点代表相应的变量。
  - 当有向弧由节点A指向节点B时，则称：A是B的**父节点**；B是A的**子节点**。
- 节点和节点之间的条件概率表（Conditional Probability Table, CPT），表示局部条件概率分布：P(node|parents) 。

目的：由证据得出原因发生的概率。即观察到结果概率$P(Y)$，求引起结果中某原因X的$P(X|Y)$。

如何构造贝叶斯网络

选择变量，生成节点；

从左至右（从上到下），排列节点；

填充网络连接弧，表示节点之间的关系；

得到条件概率关系表。

构造贝叶斯网络的思路

有向非循环图是各个节点变量关系传递的合理表达形式。

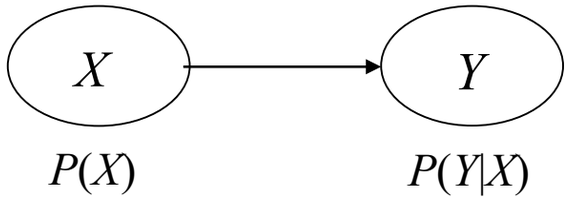
条件概率的引入使得计算较之全连接网络有了大大的简化。

CPT表相对比较容易得到。有时可以用某种概率分布表示。

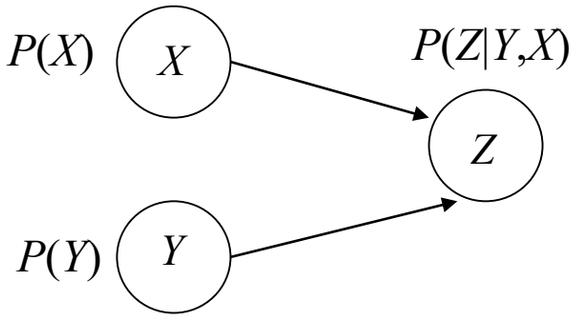
构造贝叶斯网络

简单的联合概率可以直接从网络关系上得到，

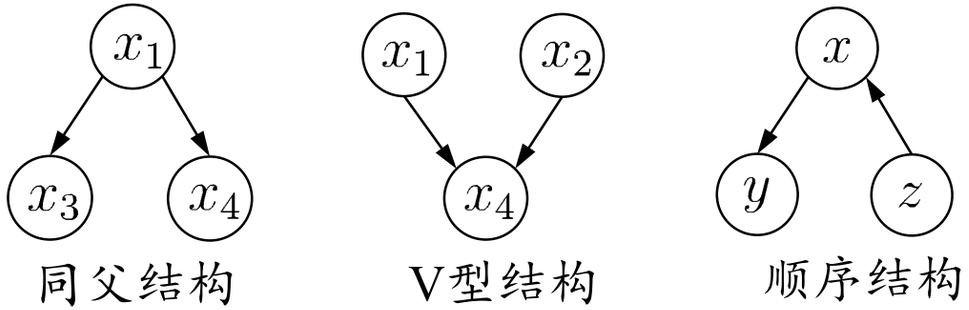
如：$P(X, Y) = P(X)P(Y|X)$



又如：$P(X, Y, Z) = P(X)P(Y)P(Z|X, Y)$



贝叶斯网中三个变量之间的典型依赖关系：



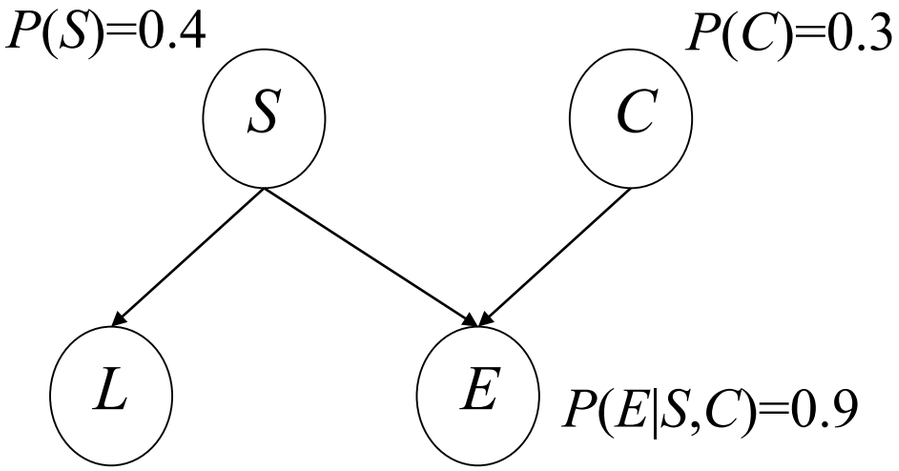
构造贝叶斯网络举例

例1 CPT表为：

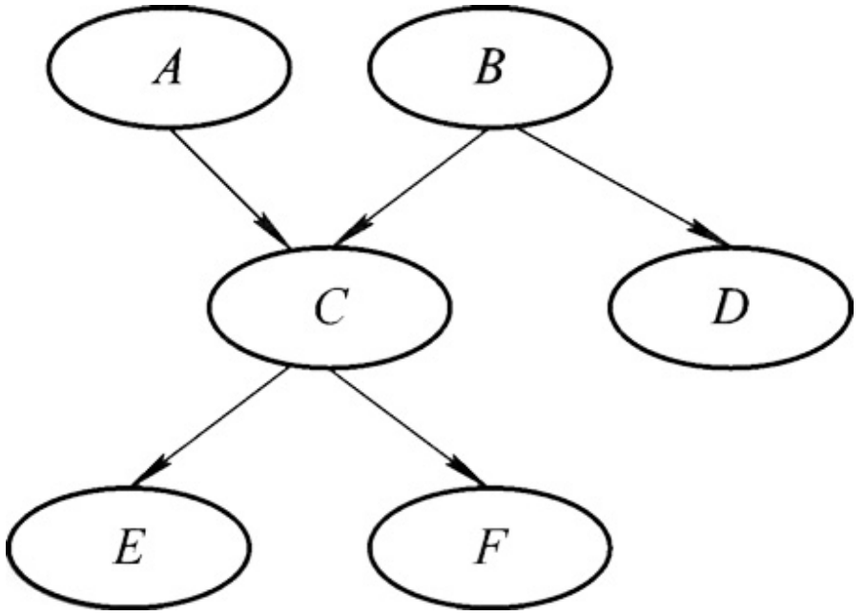
$P(S) = 0.4$； $P(C) = 0.3$； $P(E|S, C) = 0.9$；

$P(E|S, \sim C) = 0.3$； $P(E|\sim S, C) = 0.5$；

$P(E|\sim S, \sim C) = 0.1$ 。



例2 一个贝叶斯网络。其中A, B, C, D, E, F为随机变量; 5条有向边描述了相关节点或变量之间的关系;



各节点的条件概率表

表 1

$P(A)$
0.1

表 2

$P(B)$
0.2

表 3

A	B	$P(C \mid A, B)$
t	t	1
t	f	0.85
f	t	0.60
f	f	0

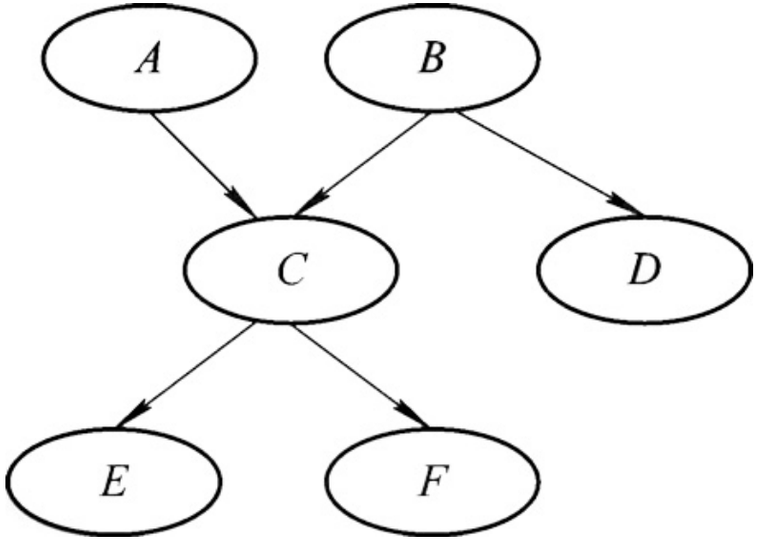


表 4

B	$P(D \mid B)$
t	0.95
f	0

表 5

C	$P(E \mid C)$
t	0.99
f	0.01

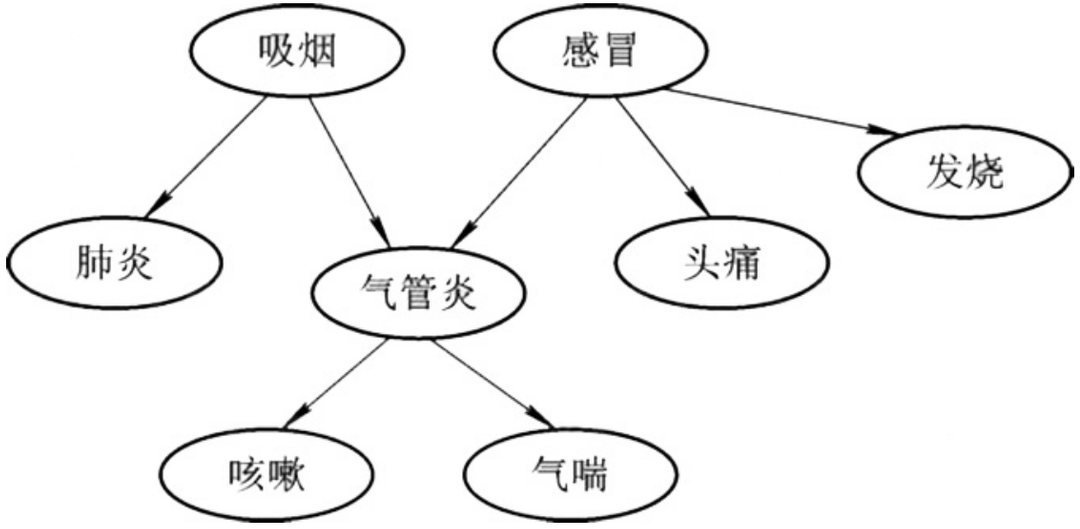
表 6

C	$P(F \mid C)$
t	1
f	0

构造贝叶斯网络举例3

医学得知: 吸烟可能会患气管炎; 感冒也会引起气管发炎, 并还有发烧、头痛等症状; 气管炎又会有咳嗽或气喘等症状。

将这些知识表示为一个贝叶斯网络。



建立条件概率表

P(吸烟)=0.6，　P(不吸烟)=0.4；

P(感冒)=0.8，　P(没感冒)=0.2；

P(气管炎|吸烟,感冒)=0.35，

P(气管炎|不吸烟,感冒)=0.25，

P(气管炎|吸烟,没感冒)=0.011，

P(气管炎|不吸烟,没感冒)=0.002；

P(咳嗽|气管炎)=0.85，

P(咳嗽|非气管炎)=0.15；

P(气喘|气管炎)=0.50，

P(气喘|非气管炎)=0.10。

贝叶斯网络：学习

根据贝叶斯网络(概率网络)结构,利用网络中的一些已知节点概率, 推算出网络中另外一些节点(未知) 的概率。

分类：

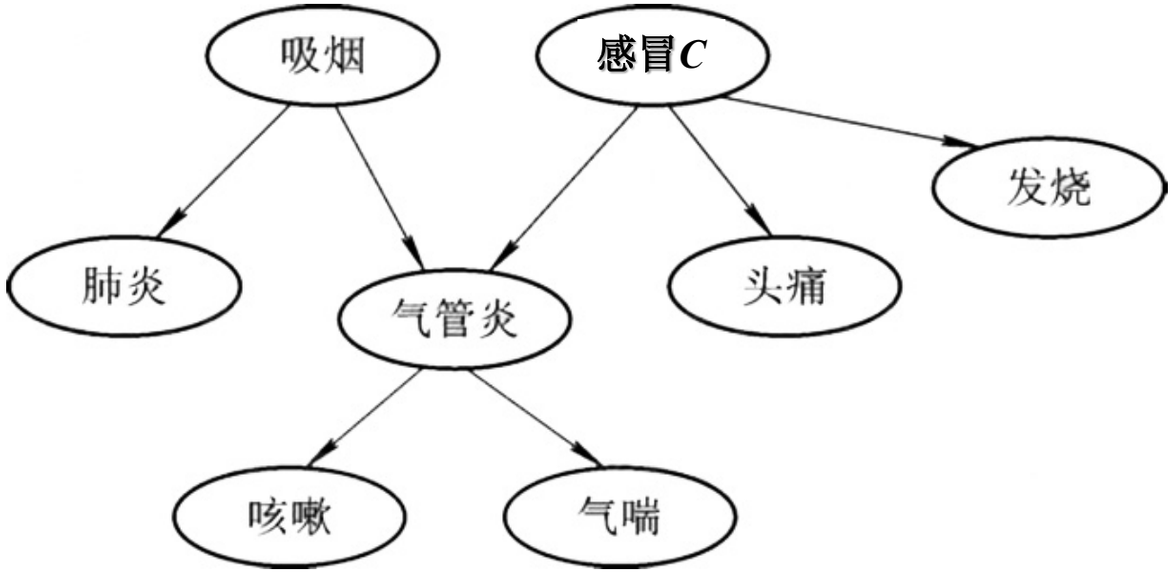
- 因果推理
- 诊断推理
- 混合推理 (略)

因果推理

因果推理：

- 原因到结果的推理。
- 已知网络中的祖先节点，计算后代节点的条件概率。
- 是一种自上而下的推理。

例如：已知某人吸烟(S), 计算他患气管炎(T)的概率$P(T|S)$ 。



$$P(T|S) = P(T,C|S) + P(T,\sim C|S)$$

因吸烟得气管炎的概率等于因吸烟而得气管炎且患感冒的概率与因吸烟而得气管炎且未患感冒的概率之和。

因果推理 例

已知某人吸烟(S), 计算他患气管炎(T)的概率$P(T|S)$：　$P(T|S) = P(T,C|S) + P(T,\sim C|S)$

$$P(T, C \mid S) = P(T, C, S)/P(S)$$ (逆向使用概率的乘法公式)

$$= P(T|C, S)P(C, S)/P(S)$$ (乘法公式)

$$= P(T|C, S)P(C|S)$$ (对$P(C, S)/P(S)$使用乘法公式)

$$= P(T|C, S)P(C)$$ (因为C与S条件独立)

$$P(T, \sim C \mid S) = P(T \mid \sim C, S)P(\sim C)$$

$$P(T \mid S) = P(T \mid C, S)P(C) + P(T \mid \sim C, S)P(\sim C)$$

$$P(\text{气管炎}|\text{吸烟}) = P(\text{气管炎}|\text{感冒},\text{吸烟})P(\text{感冒}) + P(\text{气管炎}|\text{没感冒},\text{吸烟})P(\text{没感冒})$$

$$=0.35\times0.8+0.011\times0.2=0.2822$$

因此，吸烟可引起气管炎的概率为0.2822。

条件概率表

P(吸烟)=0.6，　P(不吸烟)=0.4；

P(感冒)=0.8，　P(没感冒)=0.2；

P(气管炎|吸烟,感冒)=0.35，

P(气管炎|不吸烟,感冒)=0.25，

P(气管炎|吸烟,没感冒)=0.011，

P(气管炎|不吸烟,没感冒)=0.002；

P(咳嗽|气管炎)=0.85，

P(咳嗽|非气管炎)=0.15；

P(气喘|气管炎)=0.50，

P(气喘|非气管炎)=0.10。

因果推理的思路和方法

首先, 根据问题表达询问节点的条件概率。

然后, 利用乘法公式、全概率公式对条件概率适当变形, 直到其中的所有概率值都可以贝叶斯网络的CPT中得到。

最后, 将相关概率值代入概率表达式进行计算即得询问节点的条件概率。

诊断推理

诊断推理

- 由结果到原因的推理。
- 已知网络中的后代节点而计算祖先节点的条件概率。
- 是一种自下而上的推理。

诊断推理的一般思路和方法：

- 利用贝叶斯公式将诊断推理问题转化为因果推理问题;
- 再用因果推理的结果, 导出诊断推理的结果。

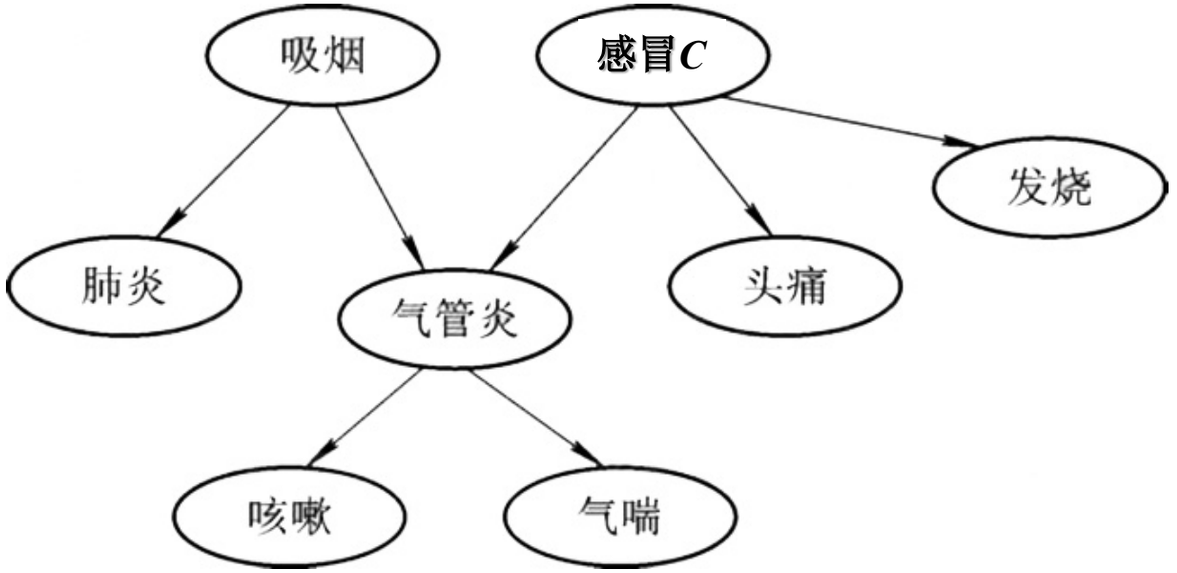
诊断推理 例

假设已知某人患了气管炎(T), 计算他吸烟(S)的后验概率$P(S|T)$。

由贝叶斯公式, 有

$$P(S \mid T) = \frac{P(T \mid S)P(S)}{P(T)}$$

- $P(T \mid S) = 0.2822$;
- $P(S) = 0.6$;
- **$P(T)$=？？**



诊断推理 例

由全概率公式

- $P(T) = P(T \mid S)\,P(S) + P(T \mid \sim S)\,P(\sim S)$

$$P(T \mid \sim S) = P(T, C \mid \sim S) + P(T, \sim C \mid \sim S) = P(T \mid C, \sim S)P(C) + P(T \mid \sim C, \sim S)P(\sim C)$$

$$= 0.25*0.8 + 0.002*0.2 = 0.2004$$

$$P(T) = 0.2822*0.6 + 0.2004*0.4 = 0.97082$$

$$P(T \mid S) = \frac{P(T \mid S)P(S)}{P(T)} = \frac{0.2822 \times 0.6}{0.97082} = 0.1744092$$

即该人的气管炎是由吸烟导致的概率为0.1744092。

贝叶斯网络意义

基于贝叶斯网络结构和条件概率,不仅可从祖先节点推算出后代节点的条件概率。

更重要的是利用贝叶斯公式还可以通过后代节点的概率反向推算出祖先节点的后验概率。

因此称这种因果网络为贝叶斯网络。

贝叶斯网络的建造涉及其拓扑结构和条件概率，比较复杂和困难。

采用机器学习的方法来解决贝叶斯网络的建造问题, 称为贝叶斯网络学习。

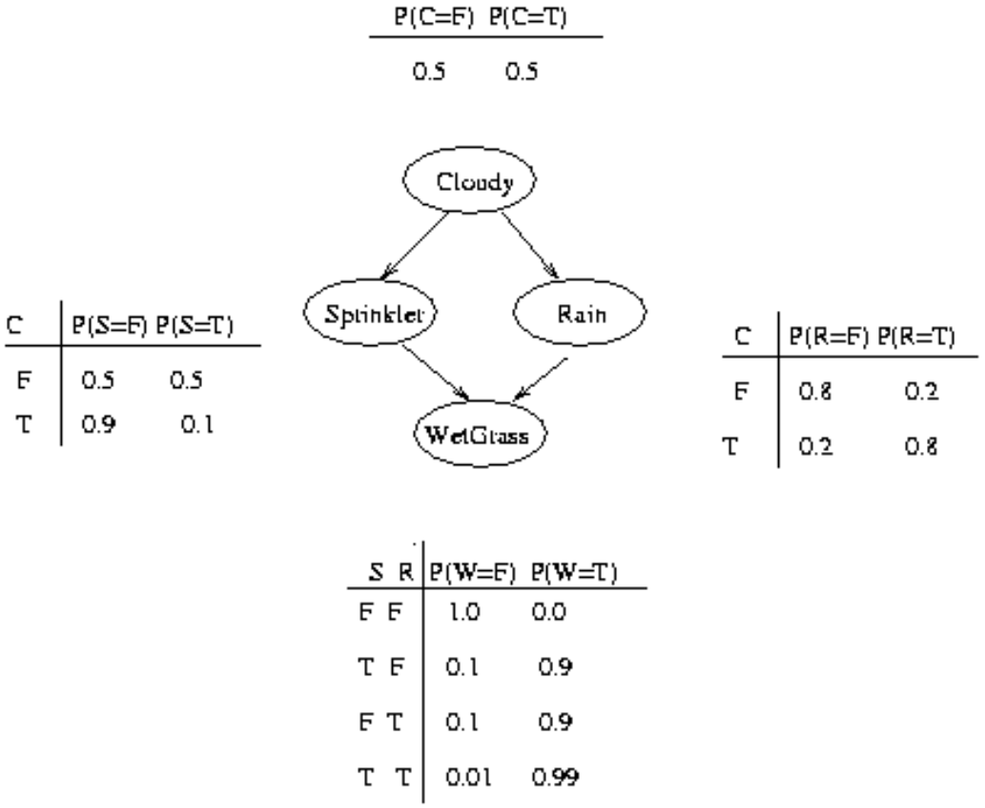
贝叶斯网络实现（基于Matlab）

Bayes Net Toolbox for Matlab

https://code.google.com/p/bnt/

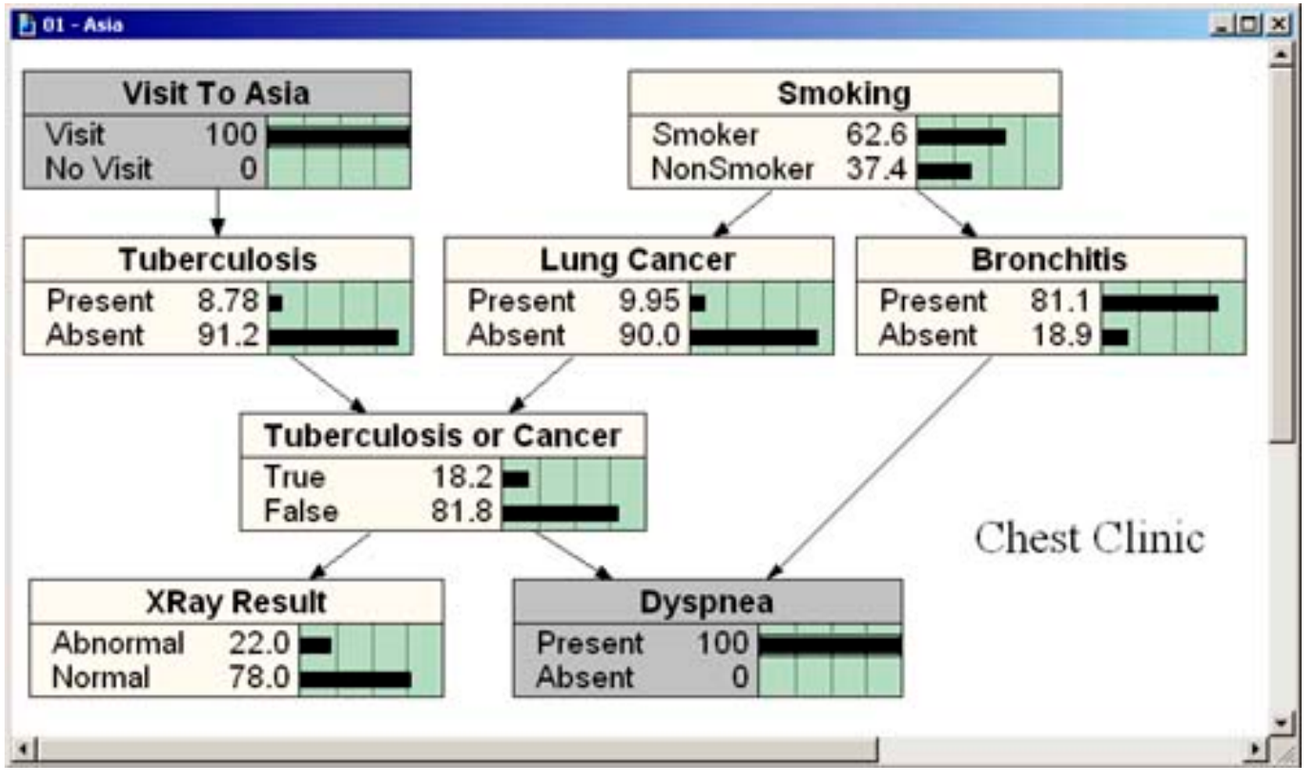
当前版本为 FullBNT-1.0.7

构造贝叶斯网络（基于Matlab）

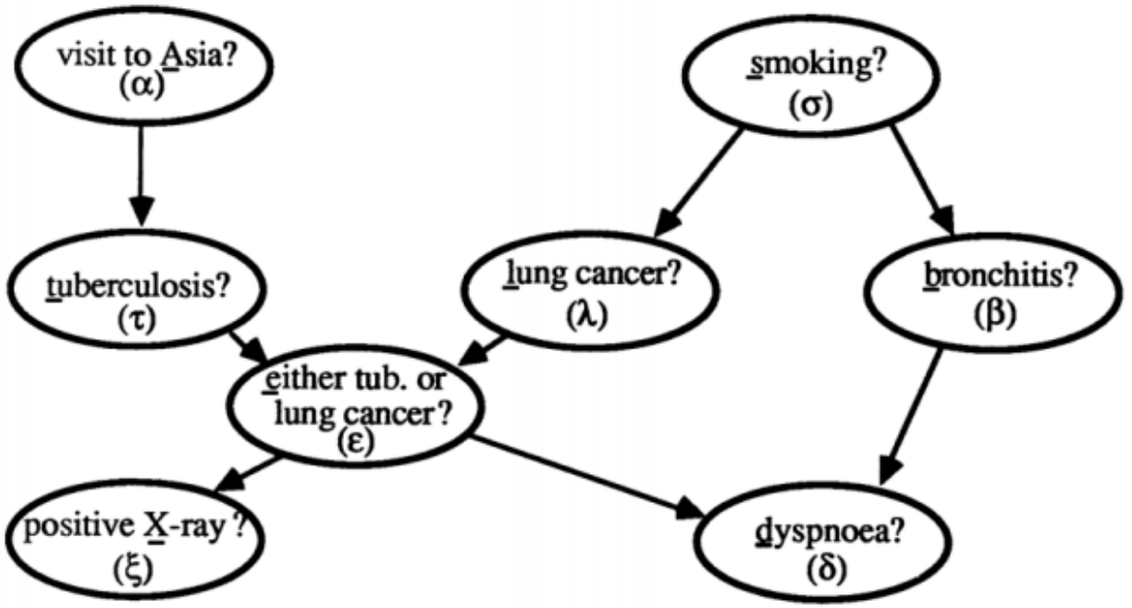



```
%建立贝叶斯网络结构，草地湿润模型
N = 4; %四个节点 分别是cloudy，sprinkler，rain，wetGrass
dag = zeros(N, N);
C = 1; S = 2; R = 3; W = 4;
dag(C,[R S]) = 1;                %节点之间的连接关系
dag(R,W) = 1;     dag(S,W) = 1;
discrete_nodes = 1:N;            %离散节点
node_sizes = 2*ones(1,N);     %节点状态数
bnet = mk_bnet(dag,node_sizes,'names',{'cloudy','sprinkler’,'rain','wetgrass'},'discrete',discrete_nodes);
bnet.CPD{C} = tabular_CPD(bnet,C,[0.5 0.5]);%手动输入的条件概率？？？
bnet.CPD{R} = tabular_CPD(bnet,R,[0.8 0.2 0.2 0.8]);
bnet.CPD{S} = tabular_CPD(bnet,S,[0.5 0.9 0.5 0.1]);
bnet.CPD{W} = tabular_CPD(bnet,W,[1 0.1 0.1 0.01 0 0.9 0.9 0.99]);
```

贝叶斯网络实现（基于Matlab）



具体见代码



$$
\begin{array}{llll}
\alpha: & p(a) = .01 & \varepsilon: & p(e \mid l, t) = 1 \\
 & & & p(e \mid l, \bar{t}) = 1 \\
\tau: & p(t \mid a) = .05 & & p(e \mid \bar{l}, t) = 1 \\
 & p(t \mid \bar{a}) = .01 & & p(e \mid \bar{l}, \bar{t}) = 0 \\
\sigma: & p(s) = .50 & \xi: & p(x \mid e) = .98 \\
 & & & p(x \mid \bar{e}) = .05 \\
\lambda: & p(l \mid s) = .10 & & \\
 & p(l \mid \bar{s}) = .01 & \delta: & p(d \mid e, b) = .90 \\
 & & & p(d \mid e, \bar{b}) = .70 \\
\beta: & p(b \mid s) = .60 & & p(d \mid \bar{e}, b) = .80 \\
 & p(b \mid \bar{s}) = .30 & & p(d \mid \bar{e}, \bar{b}) = .10
\end{array}
$$

Any Questions?