华中科技大学
电子信息与通信学院
School of Electronic Information and Communications

机器学习导论

– Introduction to Machine Learning

第 XX 章：贝叶斯模型

(Chapter XX: Bayes Model)

华中科技大学电信学院

王邦　博士, 教授博导

wangbang@hust.edu.cn

第 02 章：贝叶斯模型

贝叶斯理论 (1)

- 假设/类别 (hypothesis/class)：$h \in H$，H 为假设/类别集合。
- 样本/数据 (sample/data)：$\boldsymbol{x} = (x_1, ..., x_i, ..., x_d)$，其中 x_i 是 $\boldsymbol{x}$ 在第 i 个属性上的取值，共有 d 个属性。
- 样本空间/数据集 (sample space/dataset)：$\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_j, ..., \boldsymbol{x}_m\}$，共有 m 个样本。
- 如何从观察到的样本推导出其类别？

贝叶斯公式 Bayesian Rule

$$P(h|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|h)P(h)}{P(\boldsymbol{x})} \tag{1}$$

- 贝叶斯公式提供了由先验概率 $P(h)$, $P(\boldsymbol{x})$ 和条件概率 $P(\boldsymbol{x}|h)$ 计算后验概率 $P(h|\boldsymbol{x})$ 的方法。

贝叶斯理论 (2)

- 先验概率(Prior Probability)：是指根据历史的资料或主观判断所确定的各种事件发生的概率，该概率没有经过实验证实，属于检验前的概率。

  假设的先验概率 $P(h)$：反映了一个背景知识，与训练数据无关。也即在没有训练数据之前，h 是一个正确假设的可能性有多少。。

  样本的先验概率 $P(\boldsymbol{x})$：又称之为归一化的证据因子，即在任何假设都未知或不确定时 $\boldsymbol{x}$ 的概率。对于给定样本 $\boldsymbol{x}$，证据因子 $P(\boldsymbol{x})$ 与其假设或类别无关。

- 条件概率(Conditional Probability) $P(\boldsymbol{x}|h)$：当已知假设 h 成立时，出现样本 $\boldsymbol{x}$ 的概率，又称之为给定 h 时的样本 $\boldsymbol{x}$ 的**似然** (likelihood)。

- 后验概率(Posterior Probability)$P(h|\boldsymbol{x})$：经过学习之后，在给定数据 $\boldsymbol{x}$ 时假设 h 成立的概率，称为 h 的后验概率。

  后验概率是在数据上学习得到的结果，因此是与训练数据集 $\boldsymbol{X}$ 相关的。

  后验概率是学习的结果，可用于解决问题时的依据：对于给定数据的后验概率，做出相应的决策，如判断数据类别，执行某种行动等。

贝叶斯理论的应用

例子 (贝叶斯公式的应用)：某厂所用的电子元件由三家元件厂提供的，提供元件的份额分别为 0.15，0.8，0.05，且三个厂家的产品在仓库是均匀混合的，无区别标志。根据以往的记录，三个厂家的次品率分别为 0.02，0.01，0.03。

- 问题 1：在仓库中随机地取一个元件，求它是次品的概率。

贝叶斯理论的应用

例子 (贝叶斯公式的应用)：某厂所用的电子元件由三家元件厂提供的，提供元件的份额分别为 0.15，0.8，0.05，且三个厂家的产品在仓库是均匀混合的，无区别标志。根据以往的记录，三个厂家的次品率分别为 0.02，0.01，0.03。

- 问题 1：在仓库中随机地取一个元件，求它是次品的概率。
- 由全概率公式计算：
  $P(A) = \sum_{i=1}^{3} P(B_i)P(A|B_i) = 0.15 \times 0.02 + 0.80 \times 0.01 + 0.05 \times 0.03 = 0.0125.$
- 问题 2：在仓库中随机地取一个元件，**若已知它是次品**，为分析此次品出自何厂，需求出此元件由三个厂家分别生产的概率是多少?

贝叶斯理论的应用

例子 (贝叶斯公式的应用)：某厂所用的电子元件由三家元件厂提供的，提供元件的份额分别为 0.15，0.8，0.05，且三个厂家的产品在仓库是均匀混合的，无区别标志。根据以往的记录，三个厂家的次品率分别为 0.02，0.01，0.03。

- 问题 1：在仓库中随机地取一个元件，求它是次品的概率。
- 由全概率公式计算：
  $P(A) = \sum_{i=1}^{3} P(B_i)P(A|B_i) = 0.15 \times 0.02 + 0.80 \times 0.01 + 0.05 \times 0.03 = 0.0125.$
- 问题 2：在仓库中随机地取一个元件，**若已知它是次品**，为分析此次品出自何厂，需求出此元件由三个厂家分别生产的概率是多少?
- 由贝叶斯公式计算：

$$
\begin{aligned}
P(B_1|A) &= \frac{P(B_1)P(A|B_1)}{P(A)} = \frac{0.15 \times 0.02}{0.0125} = 0.24 \\
P(B_2|A) &= \frac{P(B_2)P(A|B_2)}{P(A)} = \frac{0.80 \times 0.01}{0.0125} = 0.64 \\
P(B_3|A) &= \frac{P(B_3)P(A|B_3)}{P(A)} = \frac{0.05 \times 0.03}{0.0125} = 0.12
\end{aligned}
$$

结果表明，对一个检测到的次品来说，其来自于第 2 家工厂的概率最大。

贝叶斯理论的应用 – 拼写检查 (1)

例如当你在 Google 中输入“julx”时，因为没有“julx”这个单词。系统会猜测你的意图，并会提示你是不是要输入某一个正确的单词比如“July”。

贝叶斯理论的应用 – 拼写检查 (2)

用户输入一个单词时，可能拼写正确，也可能拼写错误。如果把拼写正确的情况记做 c（代表 correct），拼写错误的情况记做 w（代表 wrong），那么" 拼写检查" 要做的事情就是：在发生 w 的情况下，试图推断出 c。换言之：已知 w，然后在若干个备选方案中，找出可能性最大的那个 c，也就是求 $P(c|w)$ 的最大值。

根据**贝叶斯定理**，有：

$$P(c|w) = \frac{P(w|c) \times P(c)}{P(w)}$$

由于对于所有备选的 c 来说，对应的都是同一个 w，所以它们的 $P(w)$ 是相同的，因此我们只要最大化

$$P(w|c) \times P(c).$$

$P(c)$ 表示某个正确的词的出现" 概率"，它可以用" 频率" 代替。如果我们有一个**足够大的文本库**，那么这个文本库中每个单词的出现频率，就相当于它的发生概率。某个词的出现频率越高，$P(c)$ 就越大。在你输入一个错误的词“Julx”时，系统更倾向猜测你可能想输入的词是“July”，而不是“Jult”，因为“July”更常见。

$P(w|c)$ 表示在试图拼写 c 的情况下，出现拼写错误 w 的概率。为了简化问题，假定两个单词在字形上越接近，就有越可能拼错，$P(w|c)$ 就越大。举例来说，相差一个字母的拼法，就比相差两个字母的拼法，发生概率更高。你想拼写单词"July"，那么错误拼成"Julw"（相差一个字母）的可能性，就比拼成"Jullw" 高（相差两个字母）。一般把这种问题称为**“编辑距离”**，

所以，我们**比较所有拼写相近的词在文本库中的出现频率，再从中挑出出现频率最高的一个**，即是用户最想输入的那个词。

贝叶斯最优分类器 (1)

对**分类**任务来说，在所有相关概率都已知的理想情况下，贝叶斯最优分类器考虑如何基于这些概率和误判损失来选择最优的类别标记。

- 假设有 N 种类别标记，$\mathcal{Y} = \{c_1, ..., c_N\}$，$\lambda_{ij}$ 是将一个真实标记为 c_i 的样本**错误分类**为 c_j 所产生的损失。
- 基于后验概率 $P(c_i|\boldsymbol{x})$ 将给定样本 $\boldsymbol{x}$ 分类为 c_i 所产生的期望损失(expected loss)，或称之为在样本 $\boldsymbol{x}$ 上的条件风险(conditional risk):

$$R(c_i|\boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\boldsymbol{x}). \tag{2}$$

- 寻找判定准则 $h : \mathcal{X} \mapsto \mathcal{Y}$ 以最小化总体风险

$$R(h) = \mathbb{E}_{\boldsymbol{x}}[R(h(\boldsymbol{x})|\boldsymbol{x})]. \tag{3}$$

- 如果对**每个样本**$\boldsymbol{x}$，若 h 能最小化条件风险 $R(h(\boldsymbol{x})|\boldsymbol{x})$，则总体风险 $R(h)$ 也将被最小化。

贝叶斯最优分类器 (2)

- **贝叶斯判定准则** (Bayes Decision Rule:)：为最小化**总体风险**，只需在每个样本上选择那个能使条件风险 $R(c|\boldsymbol{x})$ 最小的类别标记，即

$$h^*(\boldsymbol{x}) = \arg\min_{c\in\mathcal{Y}} R(c|\boldsymbol{x}). \tag{4}$$

  此时，分类器 h^* 称为 贝叶斯最优分类器 (Bayes Optimal Classifier)。与之对应的总体风险 $R(h^*)$ 称之为贝叶斯风险 (Bayes Risk)。注意：$1-R(h^*)$ 反映了分类器所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限。

- 若目标是最小化分类错误，则误判函数 λ_{ij} 可写为：$\lambda_{ij} = \begin{cases} 0, & \text{if } i=j \\ 1, & \text{otherwise.} \end{cases}$，此时条件风险为：$R(c|\boldsymbol{x}) = 1 - P(c|\boldsymbol{x})$。

- 最小化分类错误率的贝叶斯最优分类器 为：

$$h^*(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} \frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})}, \tag{5}$$

  即对每个样本 $\boldsymbol{x}$ 选择使后验概率 $P(c|\boldsymbol{x})$ 最大的类别标记。

贝叶斯最优分类器 (3)

例子 (贝叶斯最优分类器的应用)：在某类细胞化验中，有两个可选的假设：细胞正常 (c_1)、细胞异常 (c_2)；对应的先验概率分别为：$P(c_1)=0.9$，$P(c_2)=0.1$。现有一待识别细胞呈现出状态 x，且对应的类条件概率密度为：$P(x|c_1)=0.2$ 和 $P(x|c_2)=0.4$，试对该细胞 x 进行贝叶斯最优分类：

（1）基于最小分类错误，即代价函数为：$\begin{pmatrix}\lambda_{11}=0 & \lambda_{12}=1\\ \lambda_{21}=1 & \lambda_{22}=0\end{pmatrix}$；

基于贝叶斯公式，有：$P(c_1|x)=\frac{P(x|c_1)P(c_1)}{\sum_{j=1}^{2}P(x|c_j)P(c_j)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}\approx0.818$，

$P(c_2|x)=\frac{P(x|c_2)P(c_2)}{\sum_{j=1}^{2}P(x|c_j)P(c_j)}=\frac{0.4\times0.1}{0.2\times0.9+0.4\times0.1}\approx0.182$

因此，最小化分类错误率的贝叶斯最优分类器将 x 分类为：$h^*(x)=\arg\max_{\{c_1,c_2\}}P(c|x)=c_1$，即分类为正常细胞。

（2）代价函数为：$\begin{pmatrix}\lambda_{11}=0 & \lambda_{12}=6\\ \lambda_{21}=1 & \lambda_{22}=0\end{pmatrix}$

条件风险分别为：$R(c_1|x)=\sum_{j=1}^{2}\lambda_{1j}P(c_j|x)=\lambda_{12}P(c_2|x)=6\times0.182=1.092$

$R(c_2|x)=\sum_{j=1}^{2}\lambda_{2j}P(c_j|x)=\lambda_{21}P(c_2|x)=0.818$

因此，最小化总体风险的贝叶斯最优分类器将 x 分类为：$h^*(x)=\arg\min_{\{c_1,c_2\}}R(c|x)=c_2$，即分类为异常细胞。

极大似然估计 (1)

挑战： 如何获取贝叶斯最优分类器中类条件概率 $P(\boldsymbol{x}|c)$？

$P(\boldsymbol{x}|c)$ 涉及到关于 $\boldsymbol{x}$ 所有属性的联合概率，直接根据样本出现的频率来估计是很困难的。例如，假设样本 $\boldsymbol{x}$ 的 d 个属性都是二值的，则样本空间将有 2^d 种取值，可能远大于现实应用中的训练样本数目 m。很多样本取值在训练集中可能根本没有出现，直接使用出现频率来估计 $P(\boldsymbol{x}|c)$ 是不可行的。

注意：**未被观测到**与**出现概率为零**通常是不同的。

极大似然估计 (1)

挑战： 如何获取贝叶斯最优分类器中类条件概率 $P(\boldsymbol{x}|c)$?

$P(\boldsymbol{x}|c)$ 涉及到关于 $\boldsymbol{x}$ 所有属性的联合概率，直接根据样本出现的频率来估计是很困难的。例如，假设样本 $\boldsymbol{x}$ 的 d 个属性都是二值的，则样本空间将有 2^d 种取值，可能远大于现实应用中的训练样本数目 m。很多样本取值在训练集中可能根本没有出现，直接使用出现频率来估计 $P(\boldsymbol{x}|c)$ 是不可行的。

注意：**未被观测到**与**出现概率为零**通常是不同的。

对策： 假设 $P(\boldsymbol{x}|c)$ 服从某种确定的概率分布形式。

假设 $P(\boldsymbol{x}|c)$ 具有确定的形式，且被参数向量 $\boldsymbol{\theta}_c$ 唯一确定。则可以利用训练集数据 D 估计参数 $\boldsymbol{\theta}_c$。

极大似然估计 (1)

挑战： 如何获取贝叶斯最优分类器中类条件概率 $P(\boldsymbol{x}|c)$?

$P(\boldsymbol{x}|c)$ 涉及到关于 $\boldsymbol{x}$ 所有属性的联合概率，直接根据样本出现的频率来估计是很困难的。例如，假设样本 $\boldsymbol{x}$ 的 d 个属性都是二值的，则样本空间将有 2^d 种取值，可能远大于现实应用中的训练样本数目 m。很多样本取值在训练集中可能根本没有出现，直接使用出现频率来估计 $P(\boldsymbol{x}|c)$ 是不可行的。

注意：**未被观测到**与**出现概率为零**通常是不同的。

对策： 假设 $P(\boldsymbol{x}|c)$ 服从某种确定的概率分布形式。

假设 $P(\boldsymbol{x}|c)$ 具有确定的形式，且被参数向量 $\boldsymbol{\theta}_c$ 唯一确定。则可以利用训练集数据 D 估计参数 $\boldsymbol{\theta}_c$。

极大似然估计 (1)

挑战： 如何获取贝叶斯最优分类器中类条件概率 $P(\boldsymbol{x}|c)$?

$P(\boldsymbol{x}|c)$ 涉及到关于 $\boldsymbol{x}$ 所有属性的联合概率，直接根据样本出现的频率来估计是很困难的。例如，假设样本 $\boldsymbol{x}$ 的 d 个属性都是二值的，则样本空间将有 2^d 种取值，可能远大于现实应用中的训练样本数目 m。很多样本取值在训练集中可能根本没有出现，直接使用出现频率来估计 $P(\boldsymbol{x}|c)$ 是不可行的。

注意：**未被观测到**与**出现概率为零**通常是不同的。

对策： 假设 $P(\boldsymbol{x}|c)$ 服从某种确定的概率分布形式。

假设 $P(\boldsymbol{x}|c)$ 具有确定的形式，且被参数向量 $\boldsymbol{\theta}_c$ 唯一确定。则可以利用训练集数据 D 估计参数 $\boldsymbol{\theta}_c$。

参数估计(parameter estimation)：

- **频率主义学派：** 参数虽未知，但是客观存在的固定值。可以通过优化似然函数等准则来确定参数。
- **贝叶斯学派：** 参数也是未观察到的随机变量，其本身也可以有分布。可以假定参数也服从某种先验分布，然后基于观测数据来计算参数的后验分布。

极大似然估计 (2)

极大似然估计(Maximum Likelihood Estimation, MLE)：频率主义学派根据数据采样来估计概率分布参数的经典方法。

- 令 D_c 表示训练集 D 中第 c 类样本组成的集合，假设这些样本**独立同分布**，则参数 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然为：

$$P(D_c|\boldsymbol{\theta}_c) = \prod_{\boldsymbol{x}\in D_c} P(\boldsymbol{x}|\boldsymbol{\theta}_c) \tag{6}$$

- **极大似然估计**：寻找能最大化似然 $P(D_c|\boldsymbol{\theta}_c)$ 的参数 $\hat{\boldsymbol{\theta}}_c$。
- 上式的连乘操作易造成下溢，通常使用**对数似然** (log-likelihood)

$$LL(\boldsymbol{\theta}_c) = \log P(D_c|\boldsymbol{\theta}_c) = \sum_{\boldsymbol{x}\in D_c} \log P(\boldsymbol{x}|\boldsymbol{\theta}_c) \tag{7}$$

- 最大似然估计 $\hat{\boldsymbol{\theta}}_c$ 为：

$$\hat{\boldsymbol{\theta}}_c = \arg\max_{\boldsymbol{\theta}_c} LL(\boldsymbol{\theta}_c). \tag{8}$$

极大似然估计 (3)

在连续属性情况下，**假设**概率密度函数符合高斯分布：$p(\boldsymbol{x}|c) \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$，则参数 $\boldsymbol{\mu}_c$ 和 $\boldsymbol{\sigma}_c^2$ 的极大似然估计为：

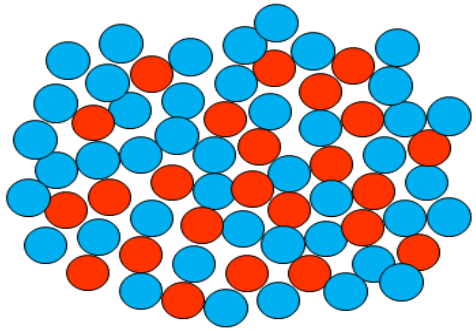
$$\begin{aligned} \hat{\mu}_c &= \frac{1}{|D_c|} \sum_{\boldsymbol{x} \in D_c} \boldsymbol{x} \\ \hat{\sigma}_c^2 &= \frac{1}{|D_c|} \sum_{\boldsymbol{x} \in D_c} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c)^{\mathsf{T}} \end{aligned}$$

这表明，通过极大似然法得到的正态分布均值就是样本均值，方差就是 $(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c)^{\mathsf{T}}$ 的均值。

注意：这种参数化的方法虽能是类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。在现实应用中，若仅凭“猜测”来假设概率分布形式，则很可能产生误导性的结果。此时，往往需要在一定程度上利用关于应用任务本身的经验知识。

极大似然估计 (4)

例子： 假设一个袋子装有红，白两种球，比例未知。每次从中随机取一球，记录其颜色后再次放回袋内（保证事件独立性）。假设抽取了 10 次，其中 7 次蓝球和 3 次红球。在此数据样本条件下，利用最大似然估计法求袋子中红球的比例。



假设每次取出的球颜色为一随机变量，且符合分布：$\boldsymbol{X} \sim \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}$

最大似然估计的思想是: 一次抽样有许多可能结果，如果某一结果在一次抽样中出现了，则认为这一结果是所有可能结果中概率最大的一个。设 10 次摸球得到样本值为 $(1,0,1,0,0,0,1,0,0,0)$。考虑到 $\boldsymbol{X}$ 是一离散型随机变量，并假设其分布为 $P(\boldsymbol{X}=1)=p(x;\theta)$。当一次抽样得到观测值 $(x_1,...,x_n)$ 时，得到此观测值的概率为：$P(\boldsymbol{X}=x_1,...,\boldsymbol{X}=x_n)$。本例中假设每次取球为独立同分布事件，故有：$P(\boldsymbol{X}=x_1,...,\boldsymbol{X}=x_n)= p(x_1;\theta)\cdots p(x_n;\theta) = \prod_{j=1}^{n} p(x_i;\theta)$

令 $L(\theta)=\prod_{j=1}^{n} p(x_i;\theta)$，为待估参数 θ 的函数，即为**似然函数**。若 $L(\theta)$ 在 $\hat{\theta}$ 处达到最大值，则称 $\hat{\theta}$ 为参数 θ 的最大似然估计。

本例中，$L(\theta)=\theta^3(1-\theta)^7$。令 $\frac{\partial L(\theta)}{\partial \theta}=0$，可得 $L(\theta)$ 在 $\hat{\theta}=0.3$ 处达到最大值，即 $\hat{\theta}=0.3$ 为参数 θ 的最大似然估计值。

极大似然估计 (5)

例子（高斯正态分布参数估计）： 假设 $\boldsymbol{X}$ 服从高斯分布：$\boldsymbol{X} \sim f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，求参数 μ 和 σ 的极大似然估计。

令 $\sigma^2 = \delta$，设样本观测值为 $x_1, x_2, ..., x_n$，似然函数为：

$$\begin{aligned} L(\mu, \delta) &= f(x_1; \mu, \sigma) f(x_2; \mu, \sigma) \cdots f(x_n; \mu, \sigma) \\ &= \frac{1}{\sqrt{2\pi\delta}} e^{-\frac{(x_1-\mu)^2}{2\delta}} \frac{1}{\sqrt{2\pi\delta}} e^{-\frac{(x_2-\mu)^2}{2\delta}} \cdots \frac{1}{\sqrt{2\pi\delta}} e^{-\frac{(x_n-\mu)^2}{2\delta}} \\ &= \frac{1}{(\sqrt{2\pi\delta})^n} e^{-\frac{(x_1-\mu)^2}{2\delta} - \frac{(x_2-\mu)^2}{2\delta} \cdots - \frac{(x_n-\mu)^2}{2\delta}} \\ &= (2\pi)^{-\frac{n}{2}} \delta^{-\frac{n}{2}} e^{-\frac{1}{2\delta}\sum_{i=1}^n (x_i-\mu)^2} \end{aligned}$$

两边取对数：$\ln L(\mu, \delta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\delta - \frac{1}{2\delta}\sum_{i=1}^n (x_i - \mu)^2$

令 $\frac{\partial \ln L(\mu,\delta)}{\partial \mu} = \frac{1}{\delta}(\sum_{i=1}^n x_i - n\mu) = 0$，得到 $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$，此即为 μ 的极大似然估计。

令 $\frac{\partial \ln L(\mu,\delta)}{\partial \sigma} = -\frac{n}{2\delta} + \frac{1}{2\delta^2}\sum_{i=1}^n (x_i - \mu)^2 = 0$，并将 $\hat{\mu}$ 带入，得到 $\hat{\delta} = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})^2$，此即为 δ 的极大似然估计。

朴素贝叶斯分类器 (1)

- **朴素贝叶斯分类器** (naïve Bayes classifier, NB) 采用了**属性条件独立性假设** (attribute conditional independence assumption)：对已知类别，假设所有属性相互独立，即假设每个属性独立地对分类结果产生影响。基于属性条件独立性假设，贝叶斯公式可以写为：

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})} \prod_{i=1}^{d} P(x_i|c), \tag{9}$$

  其中 d 为属性数目，x_i 为 $\boldsymbol{x}$ 在第 i 个属性上的取值。

- **朴素贝叶斯分类器**：

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} P(x_i|c). \tag{10}$$

- 朴素贝叶斯分类器的训练过程就是基于训练集 D 来估计类先验概率 $P(c)$，并为每个属性估计条件概率 $P(x_i|c)$.

朴素贝叶斯分类器 (2)

- 令 D_c 表示训练集 D 中第 c 类样本组成的集合，若有**充足的独立同分布**样本，则可以依据样本出现频率来估计类先验概率：

$$P(c) = \frac{|D_c|}{|D|}$$

- 对离散属性而言，令 D_{c,x_i} 表示 D_c 中第 i 个属性上取值为 x_i 的样本组成的集合，则条件概率 $P(x_i|c)$ 可估计为：

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

- 对连续属性可考虑概率密度函数，假定 $p(x_i|c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$，其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的均值和方差，则有

$$p(x_i|c) = \frac{1}{\sqrt{\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

朴素贝叶斯分类器 (3)

例子：利用西瓜数据集 3.0 训练一个朴素贝叶斯分类器，对训练例“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
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	**0.697**	0.46	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	0.36	0.37	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
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.46	?

西瓜数据集 3.0 (来源：周志华，《机器学习》，清华大学出版社)

首先估计类先验概率 $P(c)$: $P(\text{好瓜} = \text{是}) = \frac{8}{17} \approx 0.471$, $P(\text{好瓜} = \text{否}) = \frac{9}{17} \approx 0.529$.

朴素贝叶斯分类器 (4)

估计每个属性的条件概率 $P(x_i|c)$

$$P_{\text{青绿}|\text{是}} = P(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375,$$
$$P_{\text{青绿}|\text{否}} = P(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333,$$
$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂} = \text{蜷缩} \mid \text{好瓜} = \text{是}) = \frac{5}{8} = 0.625,$$
$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂} = \text{蜷缩} \mid \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333,$$
$$P_{\text{浊响}|\text{是}} = P(\text{敲声} = \text{浊响} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750,$$
$$P_{\text{浊响}|\text{否}} = P(\text{敲声} = \text{浊响} \mid \text{好瓜} = \text{否}) = \frac{4}{9} \approx 0.444,$$
$$P_{\text{清晰}|\text{是}} = P(\text{纹理} = \text{清晰} \mid \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875,$$
$$P_{\text{清晰}|\text{否}} = P(\text{纹理} = \text{清晰} \mid \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222,$$
$$P_{\text{凹陷}|\text{是}} = P(\text{脐部} = \text{凹陷} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750,$$
$$P_{\text{凹陷}|\text{否}} = P(\text{脐部} = \text{凹陷} \mid \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222,$$
$$P_{\text{硬滑}|\text{是}} = P(\text{触感} = \text{硬滑} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750,$$
$$P_{\text{硬滑}|\text{否}} = P(\text{触感} = \text{硬滑} \mid \text{好瓜} = \text{否}) = \frac{6}{9} \approx 0.667,$$

- $P_{\text{青绿}|\text{是}} = \frac{3}{8}$：在训练数据集中，属于好瓜的样本有 8 条，这 8 条中色泽 = 青绿的样本有 3 条；
- $P_{\text{青绿}|\text{是}} = \frac{3}{9}$：在训练数据集中，属于好瓜的样本有 9 条，这 9 条中色泽 = 青绿的样本有 3 条；
- $P_{\text{密度：0.697}|\text{是}} \sim \mathcal{N}(\mu, \sigma)$：在训练数据集中，**粗体的数据统计得到均值和标准偏差。**

$$P_{\text{密度：0.697}|\text{是}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi}\cdot 0.129} exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right) \approx 1.959,$$
$$P_{\text{密度：0.697}|\text{否}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.195} exp\left(-\frac{(0.697-0.496)^2}{2\cdot 0.195^2}\right) \approx 1.203,$$
$$P_{\text{含糖：0.460}|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi}\cdot 0.101} exp\left(-\frac{(0.697-0.279)^2}{2\cdot 0.101^2}\right) \approx 0.788,$$
$$P_{\text{含糖：0.460}|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.108} exp\left(-\frac{(0.697-0.154)^2}{2\cdot 0.108^2}\right) \approx 0.066,$$

朴素贝叶斯分类器 (5)

$P(\text{好瓜} = \text{是} \mid \text{测 1}) = P(\text{好瓜} = \text{是}) \times P_{\text{青绿} \mid \text{是}} \times P_{\text{蜷缩} \mid \text{是}} \times P_{\text{浊响} \mid \text{是}} \times P_{\text{清晰} \mid \text{是}} \times P_{\text{凹陷} \mid \text{是}} \times P_{\text{硬滑} \mid \text{是}} \times P_{\text{密度:0.697} \mid \text{是}} \times P_{\text{含糖:0.460} \mid \text{是}} \approx 0.038$。

$P(\text{好瓜} = \text{否} \mid \text{测 1}) = P(\text{好瓜} = \text{否}) \times P_{\text{青绿} \mid \text{否}} \times P_{\text{蜷缩} \mid \text{否}} \times P_{\text{浊响} \mid \text{否}} \times P_{\text{清晰} \mid \text{否}} \times P_{\text{凹陷} \mid \text{否}} \times P_{\text{硬滑} \mid \text{否}} \times P_{\text{密度:0.697} \mid \text{否}} \times P_{\text{含糖:0.460} \mid \text{否}} \approx 6.8 \times 10^{-5}$

由于 $0.063 > 6.80 \times 10^{-5}$，根据朴素贝叶斯分类器的判断准：选择使后验概率最大的类别，因此将“测 1”判别为“好瓜”

实践中常通过取对数的方式来将“连乘”转化为“连加”以避免数值下溢

零概率问题

需注意，若某个属性值在训练集中没有与某个类同时出现过，直接根据公式 $P(x_i|c) = \frac{D_{c,x_i}}{D_c}$ 和 $h_{nb} = \arg\max_{c \in \gamma} P(c) \prod_{i=1}^{d} P(x_i|c)$ 计算将出现零概率问题.

例如，在西瓜数据集 3.0 训练朴素贝叶斯分类器时，对一个 $\forall$ “敲声 = 清脆”的测试例 x，有

$$P_{\text{清脆} \mid \text{是}} = P(\text{敲声} = \text{清脆} \mid \text{好瓜} = \text{是}) = \frac{0}{8} = 0,$$

因此即使该样本在其他属性上明显像“好瓜”，分类的结果都是

$P(\text{好瓜} = \text{是}|x) = P_{\text{清脆} \mid \text{是}} \times \ldots \equiv 0 \leq P(\text{好瓜} = \text{否}|x)$，这显然不合理.

拉普拉斯修正 (Laplacian correction)

- 若某属性的取值在训练集中没有与某个类同时出现过，即 $D_{c,x_i} = \emptyset$, 则利用 $P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$ 计算得到 $P(x_i|c) = 0$。
- **原因**：是因为数据集过小，该属性**未被观测到**？还是因为该属性**出现概率为零**？通常假定是第一个原因。
- **对策**：为了避免某个属性携带的信息被训练集中未出现的属性值“抹去”吗，在估计概率值时通常进行“平滑”(smoothing)。
- 拉普拉斯修正(Laplacian correction)：令 N 表示训练集 D 中可能的类型数，N_i 表示第 i 个属性可能的取值数，则：

$$\hat{P} = \frac{|D_c| + 1}{|D| + N} \tag{11}$$

$$\hat{P}(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \tag{12}$$

注意：拉普拉斯修正避免了因训练集样本不充分而导致概率估值为零的问题。当训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，使得估值逐渐趋于实际概率值。

贝叶斯网络 – 简介 (1)

- 贝叶斯网络 (Bayesian Network) 又被称为贝叶斯信念网络、因果网络，起源于贝叶斯统计分析。
- 贝叶斯网络将因果知识和概率知识相结合的信息表示框架，使得不确定性推理在逻辑上变得更为清晰．理解性更强。
- 贝叶斯网络将概率理论和图论相结合，利用图结构描述随机变量 (事件) 之间的依赖关系。
- 贝叶斯网络最早由 Judea Pearl 于 1986 年提出，并逐渐发展起来，主要用于表示不确定性知识和解决推理问题。
- 贝叶斯网络已经成为数据库中的知识发现和决策支持系统的有效方法：从大量数据中构造贝叶斯网络模型，从而进行不确定性知识的发现。

贝叶斯网络 – 简介 (2)

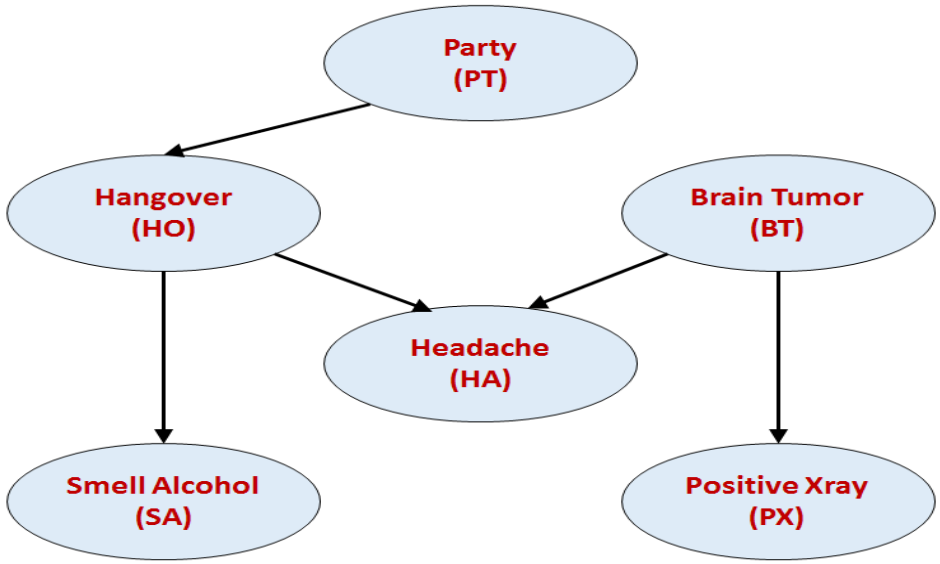
引例 (概率推理)： 所示贝叶斯网络包括 6 个结点：参加晚会 (party，PT)、宿醉 (hangover，HO)、患脑瘤 (brain tumor，BT)、头疼 (headache，HA)、有酒精味 (smell alcohol，SA) 和 X 射线检查呈阳性 (positive xray，PX)。想象这样一个场景：一个中学生回家后，其父母猜测她参加了晚会，并且喝了酒；第二天这个学生感到头疼，她的父母带她到医院做头部的 X 光检查。通过长期的观察，或者从别人那里了解，这个中学生的父母知道他们的女儿参加晚会的概率。通过长时间的数据积累，他们也知道他们的女儿参加晚会后宿醉的概率。因此，结点 party 和结点 hangover 之间有一条连线。同样，**有明显的因果关系或相关关系的结点之间都有一条连线，并且连线**从原因结点出发，指向结果结点。

原因推理结果:

1) 如果父母已知他们的女儿参加了晚会，那么第二天一早，她呼出的气体中有酒精味的概率有多大? 也就是说，当 party 发生时，smell alcohol 发生的概率有多大?

2) 如果他们的女儿头疼，那么她患脑瘤的概率有多大? 这时，如果他们又知道昨晚她参加了晚会，那么综合这些情况，她患脑瘤的可能性有多大?

结果反推原因: 如果父母早晨闻到他们的女儿呼出的气体中有酒精味，那么她昨晚参加晚会的概率有多大?



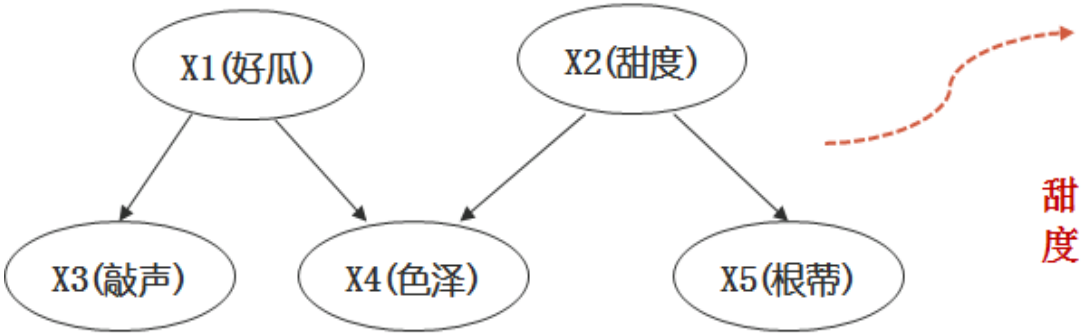
基于结点间概率关系的推理案例。

贝叶斯网络 – 简介 (3)

贝叶斯网(Bayesian Network) 亦称“信念网”(belief network)，它借助于**有向无环图** (Directed Acyclic Graph, DAG) 来刻画属性之间的依赖关系，并用**条件概率表** (Conditional Probability Table, CPT) 来描述属性的联合分布。

贝叶斯网 $B =< G, \Theta >$ 由**结构**G 和**参数**Θ 两部分组成。

- 结构 G 是一个**有向无环图**，其每个结点对应于一个属性，若两个属性有直接依赖关系，则它们由一条边连接起来。
- 参数 Θ 定量描述这种依赖关系。假设属性 x_i 在 G 中的父结点集为 π_i，则 Θ 包含了每个属性的条件依赖关系表 $\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$



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

(来源，周志华，《机器学习》，清华大学出版社)

贝叶斯网络 – 结构 (1)

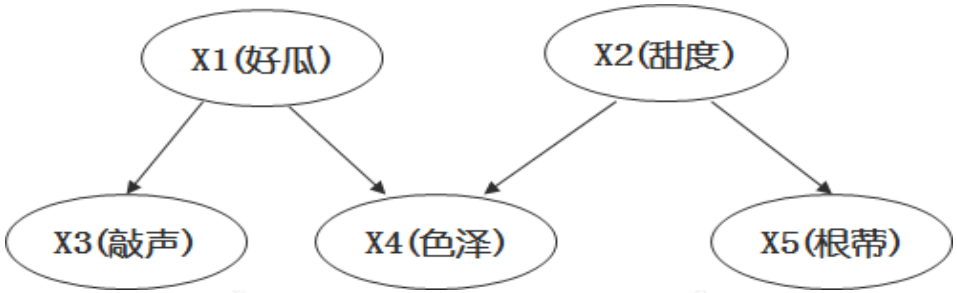
贝叶斯网结构有效地表达了属性间的条件独立性，给定父结点集，贝叶斯网**假设每个属性与它的非后裔属性独立**，于是 $B =< G, \Theta >$ 将属性 $x_1, ..., x_d$ 的联合概率分布定义为：

$$P_B(x_1, x_2, ..., x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$

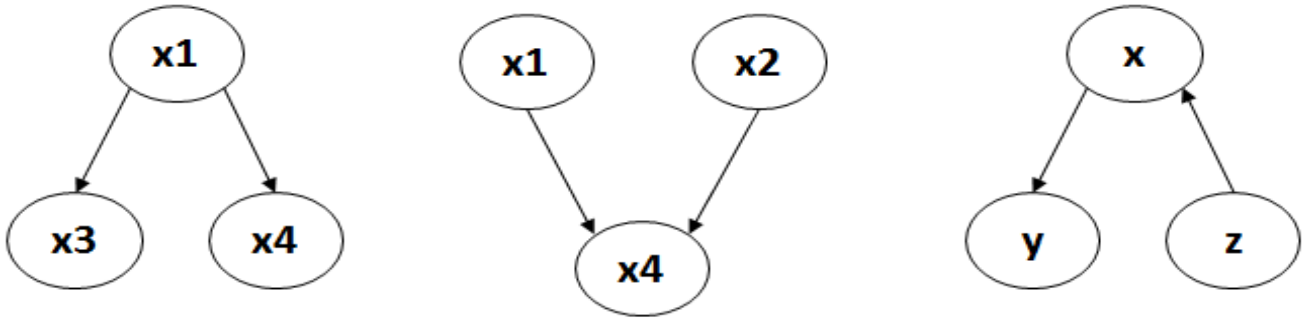
前图中的联合概率分布定义为：

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)$$

右图中，x_3 和 x_4 在给定 x_1 的取值时独立，x_4 和 x_5 在给定 x_2 的取值时独立，分别记为 $x_3 \perp x_4|x_1$ 和 $x_4 \perp x_5|x_2$。



贝叶斯网络 – 结构 (2)



贝叶斯网中三个变量之间的典型依赖关系

（来源，周志华，《机器学习》，清华大学出版社）

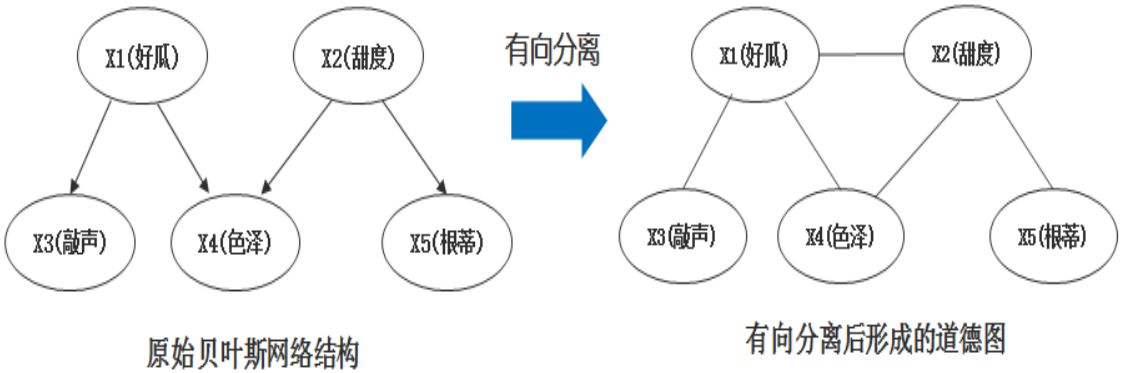
贝叶斯网中三个变量的典型关系：

- **同父结构**：给定父节点 x_1 的取值，则 x_3 与 x_4 条件独立；
- **顺序结构**：给定 x 的取值，则 y 与 z 条件独立；
- **V 型结构**：给定子节点 x_4 的取值，x_1 与 x_4 必不独立。但是，若 x_4 的取值完全未知，则 V 型结构下 x_1 与 x_2 确是相互独立的。这样的独立性称之为“边际独立性”(marginal independence)。

贝叶斯网络 – 结构 (3)

为了分析有向图中变量间的条件独立性，可使用“**有向分离**”(D-separation)，将无向图转变为一个无向图。

- 找出有向图中的所有 V 型结构，在 V 型结构的两个父结点之间加上一条无向边。
- 将所有有向边改为无向边。



(来源：周志华，《机器学习》，清华大学出版社)

假定道德图中有变量 x, y 和变量集合 $\boldsymbol{z} = \{z_i\}$，若变量 x 和 y 能在图上被 $\boldsymbol{z}$ 分开，即从道德图中将变量集合 $\boldsymbol{z}$ 去除后，x 和 y 分属两个连通分枝，则称变量 x 和 y 被 $\boldsymbol{z}$ 有向分离，$x \perp y|\boldsymbol{z}$ 成立。从上面转换后的道德图中，能找到的所有条件独立关系：$x_3 \perp x_4|x_1$, $x_4 \perp x_5|x_2$, $x_3 \perp x_2|x_1$, $x_1 \perp x_5|x_1$, $x_3 \perp x_5|x_2$。

贝叶斯网络 – 学习 (1)

- 若网络结构已知，则属性间的依赖关系已知，只需通过对训练样本“计数”，估计出每个结点的条件概率表即可。但实际中并不知道网络结构。贝叶斯网学习的**首要任务**：根据训练数据集找出结构最恰当的贝叶斯网。常用解决方法：**评分搜索**。具体做法：先定义一个**评分函数**，以此来评估贝叶斯网和训练数据的契合程度，然后**基于这个评分函数来寻找结构最优的贝叶斯网**。
- 常用评分函数通常基于信息论准则，将学习问题看作一个数据压缩任务，学习的目标是找一个**能以最短编码长度描述训练数据的模型**，此时编码的长度包括了**描述模型自身所需的字节长度**，以及**使用该模型数据所需的字节长度**。
- **“最小描述长度”准则：**(Minimal Description Length, 简称 MDL) 选择综合编码长度（包括描述网络和编码数据）最短的贝叶斯网。

贝叶斯网络 – 学习 (2)

给定训练集 $D = \{\boldsymbol{x}\}_{j=1}^{m}$，定义贝叶斯网 $B =< G, \Theta >$ 的对数似然为：

$$LL(B|D) = \sum_{j=1}^{m} \log P_B(\boldsymbol{x}_i).$$

定义贝叶斯网 B 在 D 上的评分函数：

$$s(B|D) = f(\theta)|B| - LL(B|D).$$

其中，B 是贝叶斯网的参数个数，$f(\theta)$ 表示描述每个参数 θ 所需的字节数。第一项计算编码贝叶斯网 B 所需的字节，第二项计算 B 所对应的概率分布 P_B 需要多少字节来描述 D。将学习任务转换为一个优化任务：

寻找一个贝叶斯网 B 使得评分函数 $s(B|D)$ 最小!

贝叶斯网络 – 学习 (3)

- 若 $f(\theta) = 1$，即每个参数用 1 个字节描述，则得到 AIC (Akarkie Information Criterion) 评分函数

$$AIC(B|D) = |B| - LL(B|D).$$

- 若 $f(\theta) = \frac{1}{2}\log m$，即每个参数用 $\frac{1}{2}\log m$ 字节描述，则得到 BIC (Bayesian Information Criterion) 评分函数

$$BIC(B|D) = \frac{1}{2}\log m|B| - LL(B|D)$$

- 若 $f(\theta) = 0$，即不计算对网络进行编码的长度，则评分函数退化为负对数似然；相应的，学习任务退化为极大似然估计。

贝叶斯网络 – 学习 (4)

若贝叶斯网 $B =< G, \Theta >$ 的网络结构 G 固定，则评分函数 $s(B|D)$ 的第一项为常数。此时，最小化 $s(B|D)$ 等价于对参数 Θ 的极大似然估计。同时，参数 $\theta_{x_i|\pi_i}$ 能直接在训练数据集 D 上通过经验估计获得，即

$$\theta_{x_i|\pi_i} = \hat{P}_D(x_i|\pi_i),$$

其中，$\hat{P}_D(\cdot)$ 是 D 上的经验分布。为了最小化评分函数 $s(B|D)$，只需对网络结构进行搜索，而候选结构的最优参数可直接在训练集上计算得到。
然而，从所有可能的网络结构空间搜索最优贝叶斯网结构是一个**NP 难问题**，难以快速求解。两种常用策略能在有限时间内求得近似解：

- 贪心法：例如从某个网络结构出发，每次调整一条边（增加、删除或调整方向），直到评分函数值不能再降低为止。
- 对网络结构增加约束来削减搜索空间，如将网络结构限定为树形结构。

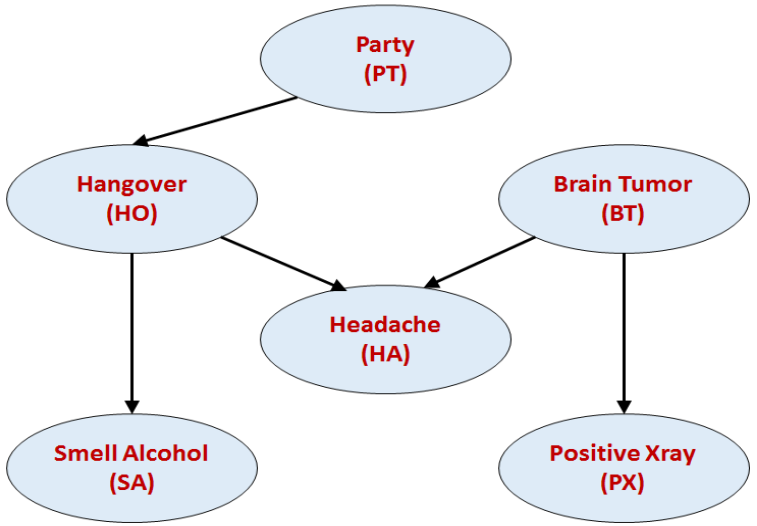
贝叶斯网络 – 学习 (5)

引例回顾：假定贝叶斯网络的 6 个结点分别为：

PT (party，参加晚会)	HO (hangover，宿醉)	BT (brain tumor，患脑瘤)
HT (headache，头疼)	SA (smell alcohol，有酒精味)	PX (positive xray，X 射线检查呈阳性)

贝叶斯网络的建立：

- 首先，要把实际问题中的事件抽象为网络中的结点；每个结点必须有明确的意义，至少有是、非两个状态或者多个状态，并且这些状态在概率意义上是完备的和互斥的。也就是说，所有状态在某一时刻只能发生一个，并且这些状态的概率之和为 1。
- 其次，在两个或多个结点之间的建立连线。基本原则：有明确因果关系的结点之间应建立连线；没有明确因果关系的结点之间尽量不要建立连线，以防止网络过于复杂而不能把握问题的实质。注意：在两个结点之间建立连线时，**要防止环的出现，因为贝叶斯网络必须是无环图**。



基于结点间概率关系的推理案例。

贝叶斯网络 – 学习 (6)

贝叶斯网络的训练： 指学习得到结点的概率分布和结点间的条件概率分布。可以通过专家经验填入，但使用更多的方法是通过历史数据训练得到。

Sample	PT	HO	BT	HA	SA	PX
1	1	1	0	1	1	0
2	0	0	1	1	0	1
3	1	0	0	0	0	0
4	1	1	1	0	1	1
5	0	0	0	0	0	1
6	1	1	0	1	1	0
7	1	0	1	0	1	0
8	0	0	1	0	0	0
9	1	0	0	0	1	0
10	1	1	1	1	1	1

可以用统计的方式得到任意结点的概率分布。假设结点 V 有 d 个状态 $V_1, V_2, ..., V_d$，则有

$$P(V_d) = \frac{\#Sample(V_d)}{\#SampleAll},$$

其中：$\#Sample(V_d)$：$P(V_d) = V_d$ 出现的数据条数；$\#SampleAll$：总的数据条数

例如，对于结点 PT，有 $P(+PT) = 7/10 = 0.7, P(-PT) = 3/10 = 0.3$。

贝叶斯网络 – 学习 (7)

贝叶斯网络的训练： 指学习得到结点的概率分布和结点间的条件概率分布。可以通过专家经验填入，但使用更多的方法是通过历史数据训练得到。

如果 U_s 表示结点 U 的一个状态，V_s 表示结点 V 的一个状态，则 U_s 发生时 V_s 发生的概率为：

$$P(V_s|U_s) = \frac{\#Sample(U_s, V_s)}{\#Sample(U_s)}$$

其中：$\#Sample(U_s, V_s)$：U_s 和 V_s **共同出现**的数据条数；$\#Sample(U_s)$：U_s 发生的次数。

例如，$+PT$ 共发生了 7 次，$+PT$ 和 $+HO$ 共同发生了 4 次，因此有 $P(+HO|+PT) = 4/7$。

同理，可以计算出多个结点间的联合条件分布。假设 U_s, V_s, W_s 分别表示结点 U, V, W 的一个状态。那么 U_s 和 V_s 发生时 W_s 发生的概率为：

$$P(W_s|U_s, V_s) = \frac{\#Sample(U_s, V_s, W_s)}{\#Sample(U_s, V_s)},$$

$\#Sample(U_s, V_s, W_s)$ 为 U_s, V_s, W_s 共同发生的次数，$\#Sample(U_s, V_s)$ 为 U_s, V_s 共同发生的次数。

例如，$+HO$ 和 $+BT$ 共发生了 2 次，而 $+HO$，$+BT$ 和 $+HA$ 共发生了 1 次，因此 $P(+HA|+HO, +BT) = 1/2 = 0.5$。

如果某个结点是结果结点或中间结点，那么得到这个结点的概率分布的方式有如下两种。(1) 直接从训练数据集中通过统计获得。(2) 先从表格数据中通过统计获得原因结果的概率分布，再从训练数据集中通过统计获得条件概率分布或联合条件概率分布，最后用全概率公式计算中间结点或结果结点的概率分布。可以验证，这两种方式获得的概率分布是一致的。

贝叶斯网 – 应用 (1)

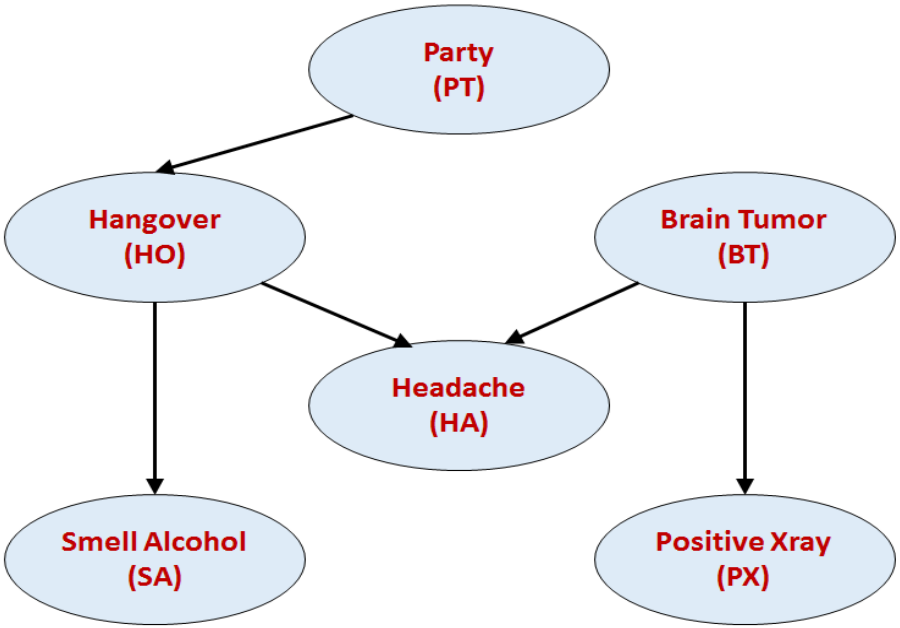
回顾：贝叶斯网是一种概率推理技术，结合概率理论和图结构来描述不同知识成分之间的条件而产生的不确定性。 **贝叶斯网络学习 (训练)：**利用现有数据对先验知识进行修正的过程，每一次学习都对贝叶斯网络的先验概率进行调整，使得新的贝叶斯网络更能反映数据中所蕴含的知识。贝叶斯网络能够持续学习．上次学习得到的后验贝叶斯网络变成下一次学习的先验贝叶斯网络，每一次学习前用户都可以对先验贝叶斯网络进行调整，使得新的贝叶斯网络更能体现数据中蕴涵的知识。
基于训练数据集构建了贝叶斯网后，可进行预测和诊断。

- **贝叶斯网预测：** 贝叶斯网络的预测是指从起因推测一个结果的推理，也称为由顶向下的推理。目的是由原因推导出结果。已知一定的原因 (证据)，利用贝叶斯网络的推理计算，求出由原因导致的结果发生的概率。
- **贝叶斯网诊断：** 贝叶斯网络的诊断是指从结果推测一个起因的推理，也称为由底至上的推理。目的是在已知结果时，找出产生该结果的原因。已知发生了某些结果，根据贝叶斯网络推理计算造成该结果发生的原因和发生的概率。该诊断作用多用于病理诊断、故障诊断中，目的是找到疾病发生、故障发生的原因。

贝叶斯网络 – 应用 (2)

引例回顾：假定贝叶斯网络已经训练完毕。6 个结点分别为：

PT (party，参加晚会)	HO (hangover，宿醉)	BT (brain tumor，患脑瘤)
HT (headache，头疼)	SA (smell alcohol，有酒精味)	PX (positive xray，X 射线检查呈阳性)



基于结点间概率关系的推理案例。

图中 Party 和 Brain Tumor 两个结点是原因结点，没有连线以它们为终点。它们的无条件概率如下表所示

	P(PT)	P(BT)
True	0.200	0.001
False	0.800	0.999

该表中给出了这两个事件发生的概率：第二列是关于 Party(参加晚会) 的概率：参加晚会的概率是 0.2，不参加晚会的概率是 0.8。第三列是关于患脑瘤的概率：患脑瘤的概率是 0.001，不患脑瘤的概率是 0.999。

贝叶斯网络 – 应用 (3)

其他几组条件概率如下：

已知结点 PT 时 HO 的条件概率

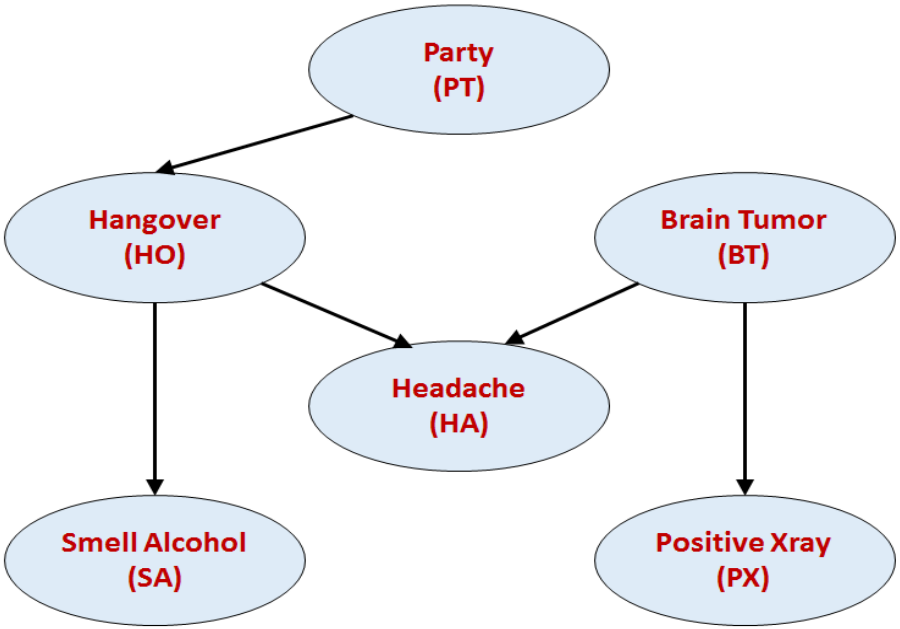
P(HO|PT)	PT=True	PT=False
True	0.700	0
False	0.300	1.000

已知结点 HO 时 SA 的条件概率

P(SA|HO)	HO=True	HO=False
True	0.800	0.100
False	0.200	0.900

已知结点 BT 时 PX 的条件概率

P(PX|BT)	BT=True	BT=False
True	0.980	0.010
False	0.020	0.990



基于结点间概率关系的推理案例。

已知 HO 和 BT 时 HA 的概率

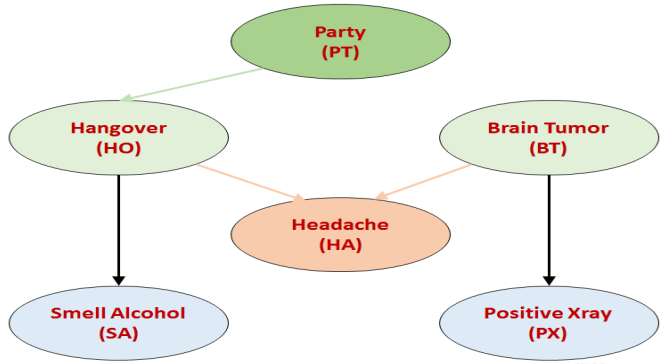
P(HA|HO,BT)	HO=True		HO=False	
	BT=True	BT=False	BT=True	BT=False
True	0.99	0.7	0.9	0.02
False	0.01	0.3	0.1	0.98

- 第 1 列解释：当宿醉发生和有脑瘤的情况下，头疼的概率是 0.99，不头疼的概率是 0.01。
- 第 3 列解释：当没有宿醉但患有脑瘤的情况下，头疼的概率是 0.9，不头疼的概率是 0.01。

贝叶斯网络 – 应用 (4)

贝叶斯网络的预测算法： 贝叶斯网络的功能之一就是在已知某些条件结点的情况下，预测结果结点的概率。贝叶斯网络也可以在不知任何结点信息的情况下计算某个结果结点的发生概率。图中，如果不知道任何结点发生与否的信息，仍然可以估算结点 HA 的概率。**约定：对于一个结点 Point，P(+Point) 表示 Point 发生的概率，P(-Point) 表示不发生的概率。**

例子 1：计算结点 HA 的概率：由贝叶斯网络图可知，HA 与 HO 和 BT 有关，而 HO 与 PT 有关。因此首先求 HO 对应的概率，由全概率公式，有：



$$
\begin{aligned}
P(+HO) &= P(+HO|+PT)P(+PT) + P(+HO|-PT)P(-PT) \\
&= 0.7 \times 0.2 + 0 \times 0.8 = 0.14 \\
P(-HO) &= P(-HO|+PT)P(+PT) + P(-HO|-PT)P(-PT) \\
&= 0.3 \times 0.2 + 1.0 \times 0.8 = 0.86
\end{aligned}
$$

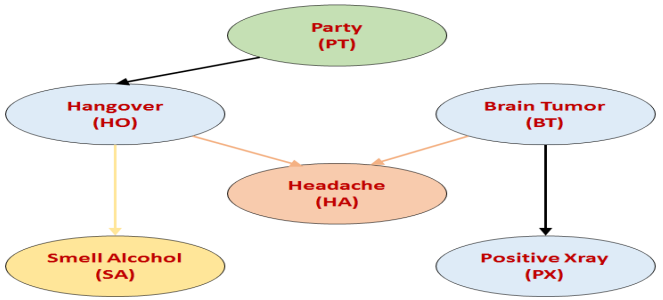
$$
\begin{aligned}
P(+HA) &= P(+HA|+BT,+HO)P(+BT)P(+HO) + P(+HA|+BT,-HO)P(+BT)P(-HO) \\
&\quad + P(+HA|-BT,+HO)P(-BT)P(+HO) + P(+HA|-BT,-HO)P(-BT)P(-HO) \\
&= 0.99 \times 0.001 \times 0.14 + 0.9 \times 0.001 \times 0.86 + 0.7 \times 0.999 \times 0.14 + 0.02 \times 0.999 \times 0.86 \\
&= 0.0001386 + 0.000774 + 0.097902 + 0.0171828 = 0.1159944 \approx 0.116 \\
P(-HA) &= 1 - P(+HA) = 0.884
\end{aligned}
$$

在没有任何结点信息 (称为证据) 的情况下，头疼的概率是 0.116，不头疼的概率是 0.884。用同样的方式，可以计算所有结点的概率，这样进一步完善贝叶斯网络。事实上，完善结点概率也是预测贝叶斯网络预测的一种情况，即在不知结点明确信息 (证据) 情况下的预测。

贝叶斯网络 – 应用 (5)

例子 2： 计算已知参加晚会的情况下，第二天早晨呼吸有酒精味的概率。

由贝叶斯网络图可以看出，SA 发生的概率只与 HO 有关，而 HO 与是否参加 Party 有关。现在已知参加了 Party，由前面的条件概率表可知：$P(+HO|+PT)=0.7$ 以及 $P(-HO|+PT)=0.3$。故 $P(+SA)$ 可由全概率公式可得：



$$\begin{aligned} P(+SA) &= P(+SA|+HO)P(+HO)+P(+SA|-HO)P(-HO) \\ &= 0.8\times 0.7+0.1\times 0.3=0.59 \end{aligned}$$

例子 3： 计算已知参加晚会的情况下，头疼发生的概率：

由贝叶斯网络图可知，HA 发生的概率与 HO 和 BT 有关。而 HO 与是否参加 Party 有关。现在已知参加了 Party，由前面的条件概率表可知：$P(+HO|+PT)=0.7$ 以及 $P(-HO|+PT)=0.3$。BT 为原因节点，与其他概率无关；BT 发生的概率为 0.001；BT 不发生的概率为 0.999。故 HO 发生的概率，由全概率公式：

$$\begin{aligned} P(+HA) &= P(+HA|+HO,+BT)P(+HO)P(+BT)+P(+HA|+HO,-BT)P(+HO)P(-BT) \\ &\quad P(+HA|-HO,+BT)P(-HO)P(+BT)+P(+HA|-HO,-BT)P(-HO)P(-BT) \\ &= 0.99\times 0.7\times 0.001+0.7\times 0.7\times 0.999+0.9\times 0.3\times 0.001+0.02\times 0.3\times 0.999 \\ &= 0.000693+0.48951+0.00027+0.005994=0.496467 \\ P(-HA) &= 1-P(+HA)=0.503533 \end{aligned}$$

也就是说，如果知道已经参加了晚会，而没有其他方面的任何证据，则这个人头疼的概率是 0.4965，不头疼的概率是 0.5035。**注意与例子 1 对比：**在没有任何结点信息 (称为证据) 的情况下，头疼的概率是 0.116，不头疼的概率是 0.884。

贝叶斯网 – 应用 (6)

贝叶斯网络预测算法的步骤：

输入：给定贝叶斯网络 B (包括网络结构 N 个结点以及结点连边、原因结点到中间结点的条件概率或联合条件概率)，给定若干个原因结点发生与否的事实向量 $\boldsymbol{E}$ (或者称为证据向量)，给定待预测的某个结点 v。

输出：结点 v 发生的概率。

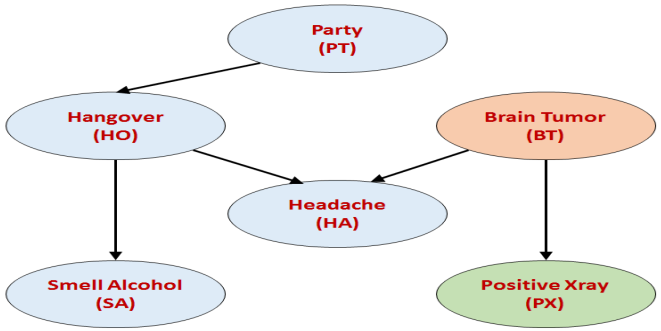
1. 把证据向量输入到贝叶斯网络 B 中。
2. 对于 B 中的每一个没处理过的结点 n，如果它具有发生的事实 (证据)，则标记它为已经处理过；否则继续下面的步骤。
3. 如果它的所有父结点中有一个没有处理过，则不处理这个结点；否则，继续下面的步骤。
4. 根据结点 n 的所有父结点的概率以及条件概率或联合条件概率计算结点 n 的概率分布，并把结点 n 标记为已处理。
5. 重复步骤 (2) ～ (4)，共 T 次。此时，结点 n 的概率分布就是它的发生 / 不发生的概率。算法结束。需要注意的是，第 (5) 步的作用是使得每个结点都有被计算概率分布的机会。

贝叶斯网络 – 应用 (7)

贝叶斯网络的诊断算法: 在已知结果结点发生与否的情况下推断条件结点发生的概率。

例子 4: 计算已知 X 光检查呈阳性的情况下，患脑瘤的概率。即求 $P(+BT|+PX)$ 的概率，依据条件概率公式:

$$
\begin{aligned}
P(+BT|+PX) &= \frac{P(+PX|+BT)P(+BT)}{P(+PX)} \\
&= \frac{0.98 \times 0.001}{0.011} = 0.089
\end{aligned}
$$



其中:

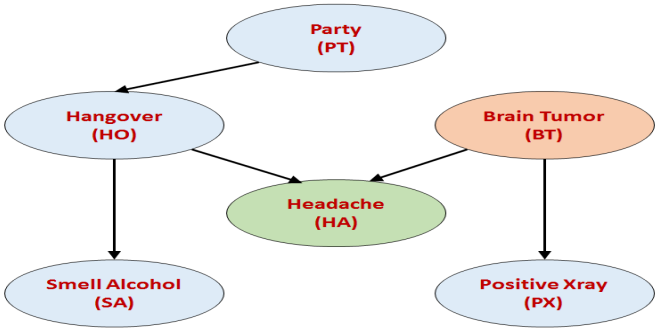
$$
\begin{aligned}
P(+PX) &= P(+PX|+BT)P(+BT) + P(+PX|-BT)P(-BT) \\
&= 0.980 \times 0.001 + 0.010 \times 0.999 \approx 0.011
\end{aligned}
$$

也就是说，当 X 光检查呈阳性的情况下，患脑瘤的概率是 0.089，不患脑瘤的概率是 0.911。

贝叶斯网络 – 应用 (8)

例子 5：计算已知头疼的情况下，患脑瘤的概率，即求 $P(+BT|+HA)$ 的概率。依据条件概率公式：$P(+BT|+HA)=\frac{P(+HA|+BT)P(+BT)}{P(+HA)}$，除了 $H(+BT)=0.001$ 概率已知外，还需先求 $P(+HA)$ 以及 $P(+HA|+BT)$ 的概率。

从例 1 可知，HA 与 HO 和 BT 有关，而 HO 仅与 PT 有关。由全概率公式求得 $P(+HO)=0.14$，$P(-HO)=0.86$，并可求得 $P(+HA)=0.116$，$P(-HA)=0.884$。再次利用全概率公式，有



P(HA|HO,BT)	HO=True		HO=False	
	BT=True	BT=False	BT=True	BT=False
True	0.99	0.7	0.9	0.02
False	0.01	0.3	0.1	0.98

$$\begin{aligned} P(+HA|+BT) &= P(+HA|+BT,+HO)P(+HO) \\ &\quad +P(+HA|+BT,-HO)P(-HO) \\ &= 0.99\times 0.14+0.9\times 0.86=0.9126 \end{aligned}$$

上面的计算得到了已知患脑瘤的情况下头疼的概率是 0.913。这个条件概率是一个边缘分布，它是从联合条件概率分布 (H0，BT→HA) 去掉一个条件 HO 得到的。

最后计算得到在头疼的情况下，患脑瘤的概率：

$$\begin{aligned} P(+BT|+HA) &= \frac{P(+HA|+BT)P(+BT)}{P(+HA)} \\ &= \frac{0.9126\times 0.001}{0.116}\approx 0.007867 \end{aligned}$$

贝叶斯网 – 应用 (9)

贝叶斯网络诊断算法的步骤：

输入：给定贝叶斯网络 B，给定若干个结果结点发生与否的事实向量 $\boldsymbol{E}$(或者称为证据向量)，给定待诊断的某个结点 v。

输出：结点 v 发生的概率。

1. 把证据向量输入到贝叶斯网络 B 中。
2. 对于 B 中的每一个没处理过的结点 n，如果它具有发生的事实 (证据)，则标记它为已经处理过；否则继续下面的步骤。
3. 如果它的所有子结点中有一个没有处理过，则不处理这个结点；否则，继续下面的步骤。
4. 根据结点 n 的所有子结点的概率以及条件概率或联合条件概率计算结点 n 的概率分布，并把结点 n 标记为已处理。
5. 重复步骤 (2) ～ (4) 共 T 次。此时，原因结点 v 的概率分布就是它的发生 / 不发生的概率。算法结束。需要注意的是，第 (5) 步的作用是使得每个结点都有被计算概率分布的机会。

小结

本节主要讲解了贝叶斯模型的原理与应用，主要包括以下内容：

- 贝叶斯理论：先验概率，后验概率以及贝叶斯公式；
- 贝叶斯最优分类器：最小化总体风险以及最小化分类错误的贝叶斯最优分类器；
- 极大似然估计：基本原理，以及如何用于求取类条件概率；
- 朴素贝叶斯分类器：属性条件独立性假设下的贝叶斯公式；
- 贝叶斯网络：基本概念，网络结构与学习，其预测与诊断应用。