【人工智能】— 不确定性、先验概率/后验概率、概率密度、贝叶斯法则、朴素贝叶斯

不确定性

我们考虑一个不确定推理的例子：诊断牙病患者的牙痛。诊断——无论是医疗、汽车修理、或者其他——几乎总是包含不确定性。我们试着使用命题逻辑写出牙病诊断的规则，以便让我们看看逻辑方法是如何失败的。考虑下面的简单规则：

$$Toothache \Rightarrow Cavity$$

问题是，上面这条规则是错误的。不是所有的牙痛（toothache）都是因为牙齿有洞（cavity），有时牙痛是因为牙龈疾病（gum disease）、牙龈脓肿（abscess）、或其他几种问题中的一种：

$$Toothache \Rightarrow Cavity \vee GumProblem \vee Abscess \cdots$$

不幸的是，为了使得规则正确，我们不得不增加一个几乎无限长的可能原因的列表。我们可以尝试把上面的规则改成一条因果规则：

$$Cavity \Rightarrow Toothache$$

但这条规则也不正确，因为不是所有的牙洞都会引起牙痛。修正该规则的唯一途径是从逻辑上穷举各种可能的情形：用一个牙洞引起牙痛所需的所有**限制**（qualifications）扩充规则的左边。试图使用逻辑处理像医疗诊断这样的问题域之所以会失败，有以下三个主要原因：

- **惰性**：为了确保得到一个没有任何意外的规则，需要列出前提和结论的完整集合，这个工作量太大，这样的规则也难以使用。
- **理论的无知**：对于该领域，医学科学还没有完整的理论。
- **实践的无知**：即使我们知道所有的规则，对于一个特定的病人我们也可能无法确定，

牙痛和牙洞之间的联系并不是一方对另一方的逻辑结果。这是医学领域的典型情况，大多数其他判断性的领域也是如此：包括法律、商业、设计、汽车修理、园艺、年代测定，等等。Agent 的知识顶多能提供对相关语句的**信念度**（degree of belief）。我们处理信念度的主要工具是**概率理论**（probability theory）。在 8.1 节的术语中，逻辑和概率理论的**本体约束**（ontological commitments）是相同的——世界是由在某种特定情形下成立或不成立的事实组成的——但**认识约束**（epistemological commitments）是不同的：逻辑 Agent 相信每个语句是正确的或错误的，或不做评价，而概率 Agent 为每条语句赋予一个 0 到 1 之间的数值作为其信念度。

概率提供了一种方法以**概括**由我们的惰性和无知产生的不确定性，由此解决**限制问题**。也许我们不能确定是什么病在折磨一个特定的病人，但我们相信牙痛病人有牙洞的可能性，比如 80%的可能性——即 0.8 的概率。也就是说，我们期望在所有与当前情形无法区别的情形中，根据 Agent 的知识，有 80%的病人有牙洞。这种信念可由统计数据获得——目前为止所见过的牙痛患者中 80%有牙洞——或由一般性的牙科知识获得，或结合多种证据获得。

不确定性与理性决策

再次考虑去机场的规划 A_{90}。假设规划 A_{90} 让我们有 97%的机会赶上航班，这意味着这个规划是一个理性的选择吗？不一定：可能其他规划有更高的概率，比如 A_{180}。如果绝对不允许错过航班，那么在机场的长时间等待是值得的。A_{1440} 是一个提前 24 小时出门的规划，这个规划怎么样呢？在大多数情况下，这个规划不是一个好的选择：尽管这个规划几乎能确保按时到达机场，但也造成难以忍受的等待——更不用说难以下咽的机场饮食。

为了做出这些选择，Agent 首先必须在各种规划的不同**结果**（outcomes）之间有所**偏好**（preferences）。一个结果是一个完全特定的状态，包括 Agent 是否按时到达机场、在机场等待多长时间等诸如此类的要素。我们使用**效用理论**（utility theory）来对偏好进行表示和推理（这里的 **utility** 一词是“功用、效用”的意思，而不是“电力和水等公共事业”的意思）。效用理论认为，每个状态对一个 Agent 而言都有一定程度的有用性，即效用，而 Agent 会偏好那些效用更高的状态。

在被称为**决策理论**的理性决策通用理论中，由效用表示的偏好是与概率理论相结合的：

决策理论 = 概率理论 + 效用理论

决策理论的基本思想是：一个 Agent 是理性的，当且仅当它选择能产生最高期望效用的行动，这里的期望效用是行动的所有可能结果的平均。这称为**期望效用最大化**（Maximum Expected Utility，MEU）原则。也许“期望”看似是一个含糊的、不确定的术语，但此处它有精确的含义：它是指“平均”或结果的“统计平均”（结果的概率加权平均）。在第 5 章中当我们简短地接触西洋双陆棋的优化决策时，我们见识了这条原则所发挥的作用。它事实上完全是一条通用原则。

基本概率符号

先验概率(无条件概率)/后验概率(条件概率)

例如，掷两个一样的色子时，$P(Total=11)=P((5,6))+P((6,5))=1/36+1/36=1/18$。注意，概率理论并不要求知道每个可能世界的概率。例如，如果我们相信两个色子“密谋”产生相同的数，我们可以说 $P(doubles)=1/4$，而不需知道两个色子偏向于产生两个 6 还是偏向于产生两个 2。就像逻辑断言，这个概率断言在没有完全确定概率模型的情况下给出了基本的概率模型的限制。

称 $P(Total=11)$和 $P(doubles)$这样的概率为**无条件概率**（unconditional probabilities）或**先验概率**（prior probabilities，有时简写为 priors）：它们是指不知道其他信息的情况下对命题的信念度。然而，在大多数情况下，我们会有一些已经为我们所知的信息——通常称为**证据**（evidence）。例如，我们已经看到第一个色子结果是 5，此时我们还在等待第二个色子停止翻滚。在这种情况下，我们不再关心得到相同数的先验概率，而是关心在第一色子是 5 的前提下两个色子结果相同的概率。这个概率写作 $P(doubles|Die_1=5)$，其中“|”读作“给定（given）”。这样的概率称为**条件概率**（conditional probabilities）或**后验概率**（posterior probabilities,有时简写为 posteriors）。类似地，如果我要去牙医那里做日常检查，那么概率 $P(cavity)=0.2$ 可能我感兴趣的；但如果我是因为有牙痛而到牙医那里检查，那么 $P(cavity \mid toothache)=0.6$ 是我感兴趣的。注意，符号“|”的优先级是：任何 P(…|…)形式的表达式是指 P((...)|(...))。

数学上，条件概率是由无条件概率定义的：对于任何命题 a 和 b

$$P(a|b)=\frac{P(a\wedge b)}{P(b)} \qquad (13.3)$$

只要 $P(b)>0$，这个公式是成立的。例如

条件概率的定义（见等式（13.3））可以写成**乘法规则**（product rule）：

$$P(a\wedge b)=P(a|b)P(b)$$

乘法规则也许更容易理解：为了使 a 和 b 都成立，就需要 b 成立，且需要在给定 b 的前提下 a 也成立。

随机变量

概率理论中变量被称为**随机变量**（random variables），变量的名字以大写字母开头。这样，在掷色子的例子中，*Total* 和 Die_1 就是随机变量。每个随机变量有一个**定义域**（domain）——这个变量能取的所有可能值组成的集合。两个色子的问题中，*Total* 的定义域是 $\{2,\cdots,12\}$，Die_1 的定义域是 $\{1,\cdots,6\}$。一个布尔变量的定义域是 $\{true, false\}$（注意，变量的值总是小写）；例如，“两个色子产生相同数”的命题可以写作 *Doubles=true*。按照约定，具有 A=*true* 形式的命题可以简写为 *a*，而 A=*false* 简写为 $\neg a$。（前面一节中的 *doubles*、*cavity* 和 *toothache* 是这种形式的简写。）像在 CSP（约束满足问题）中一样，定义域可以是一些记号组成的集合；我们可以将 Age 的定义域定义为 {*juvenile*, *teen*, *adult*}，将天气的定义域定义为 {*sunny*, *rain*, *cloudy*, *snow*}。在不引起歧义的情况下，通常可以用一个值本身表示“一个特定的变量取这个值”的命题；这样，*sunny* 就可以代表 *Weather=sunny*。

前面这些例子的变量都有有限定义域。变量也可以有无限定义域——离散的（如整数）或连续的（如实数）。对于任何具有有序定义域（ordered domain）的变量，可以允许类似 $NumberOfAtomsInUniverse \geq 10^{70}$ 这样的不等式。

最后，我们可以用命题逻辑中的连接符号来组合这些基本的命题（包括布尔变量的简写形式）。例如，我们可以将“如果患者是一个没有牙痛的青少年，那么她有牙洞的概率是 0.1”表示为：

$$P(cavity|\neg toothache \wedge teen)=0.1$$

有时，我们要讨论一个随机变量每个可能取值的概率。我们可以写：

$$P(Weather = sunny) = 0.6$$
$$P(Weather = rain) = 0.1$$
$$P(Weather = cloudy) = 0.29$$
$$P(Weather = snow) = 0.01$$

这可以简写为：

$$\mathbf{P}(Weather) = \langle 0.6, 0.1, 0.29, 0.01 \rangle$$

其中黑体的 **P** 表示结果是由一些数组成的向量，同时我们也假设 *Weather* 的定义域中的值有一个预定义的顺序 <*sunny*, *rain*, *cloudy*, *snow*>。我们说 **P** 定义了随机变量 *Weather* 的一个**概率分布**（probability distribution）。符号 **P** 也被用于条件分布（conditional distributions）：**P**(*X*|*Y*)给出每个可能的 *i*、*j* 组合下的值 $P(X=x_i|Y=y_j)$。

概率密度

对于连续变量，我们不可能用一个向量写出整个分布，因为有无限多的值。然而，我们可以把一个随机变量取某个值 *x* 的概率定义为一个以 *x* 为参数的函数。例如，语句

$$P(NoonTemp = x) = Uniform_{[18C,26C]}(x)$$

表示“中午的温度均匀分布在 18～26 摄氏度之间”的信念。我们称此为**概率密度函数**（probability density function，有时简写为 pdf）。

概率密度函数与离散概率分布的含义不同。说“概率密度在 18～26*C* 之间均匀分布”是指温度 100%地落在这个 8*C* 宽的区间范围内的某个位置，而落在其中任何一个 4*C* 宽的区间范围内的可能性是 50%，诸如此类。我们将一个连续随机变量 *X* 在值 *x* 处的概率密度写作 $P(X = x)$或简写为 $P(x)$；$P(x)$的直观定义是 *X* 落在以 *x* 开始的一个相当小的区域内的概率除以这个区间的宽度：

$$P(x) = \lim_{dx \to 0} P(x \leq X \leq x + dx) / dx$$

对于 *NoonTemp*，有

$$P(NoonTemp = x) = Uniform_{[18C,26C]}(x) = \begin{cases} \frac{1}{8C} & \text{当 } 18C \leq x \leq 26C \\ 0 & \text{其他} \end{cases}$$

其中 *C* 代表摄氏度，不是常量。注意 $P(NoonTemp = 20.18C) = \frac{1}{8C}$ 中，$\frac{1}{8C}$ 不是概率，而是概率密度。*NoonTemp* 恰好等于 20.18*C* 的概率是 0，因为 20.18*C* 是一个宽度为 0 的区间。有的作者用不同的符号来区分离散分布与密度函数；我们对于两种情况都用 *P* 表示，因为很少会引起混淆，且公式都是一样的。注意，概率是无单位的数值，而密度函数是用单位来度量的，上面的这个例子中单位是 $\frac{1}{C}$。

联合概率分布

除了单个变量的分布外，还需要符号表示多个变量的分布，我们使用逗号分割多个变量。例如，**P**(*Weather*, *Cavity*)表示 *Weather* 和 *Cavity* 的取值的所有组合的概率。这是一个 4×2 的概率表，称为 *Weather* 和 *Cavity* 的**联合概率分布**（joint probability distribution）。我也可以将变量与值搭配：**P**(*sunny*, *Cavity*)是一个二元向量，给出晴天且有牙洞的概率和晴天且无牙洞的概率。符号 **P** 使得某些表示比起不使用 **P** 时的表示更精练。例如，*Weather* 和 *Cavity* 所有可能取值的乘法规则可以写成一个单一的等式：

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\ \mathbf{P}(Cavity),$$

而不必写成如下的 4×2=8 个等式（此处使用缩写 *W* 和 *C*）：

$$P(W=sunny \wedge C=true) = P(W=sunny|C=true)\ P(C=true)$$
$$P(W=rain \wedge C=true) = P(W=rain|C=true)\ P(C=true)$$
$$P(W=cloudy \wedge C=true) = P(W=cloudy|C=true)\ P(C=true)$$
$$P(W=snow \wedge C=true) = P(W=snow|C=true)\ P(C=true)$$
$$P(W=sunny \wedge C=false) = P(W=sunny|C=false)\ P(C=false)$$
$$P(W=rain \wedge C=false) = P(W=rain|C=false)\ P(C=false)$$
$$P(W=cloudy \wedge C=false) = P(W=cloudy|C=false)\ P(C=false)$$
$$P(W=snow \wedge C=false) = P(W=snow|C=false)\ P(C=false)$$

除了单个变量的分布外，还需要符号表示多个变量的分布，我们使用逗号分割多个变量。例如，**P**(*Weather*, *Cavity*)表示 *Weather* 和 *Cavity* 的取值的所有组合的概率。这是一个 4×2 的概率表，称为 *Weather* 和 *Cavity* 的**联合概率分布**（joint probability distribution）。我也可以将变量与值搭配：**P**(*sunny*, *Cavity*)是一个二元向量，给出晴天且有牙洞的概率和晴天且无牙洞的概率。符号 **P** 使得某些表示比起不使用 **P** 时的表示更精练。例如，*Weather* 和 *Cavity* 所有可能取值的乘法规则可以写成一个单一的等式：

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\,\mathbf{P}(Cavity),$$

而不必写成如下的 4×2=8 个等式（此处使用缩写 *W* 和 *C*）：

$$P(W=sunny \wedge C=true) = P(W=sunny|C=true)\,P(C=true)$$
$$P(W=rain \wedge C=true) = P(W=rain|C=true)\,P(C=true)$$
$$P(W=cloudy \wedge C=true) = P(W=cloudy|C=true)\,P(C=true)$$
$$P(W=snow \wedge C=true) = P(W=snow|C=true)\,P(C=true)$$
$$P(W=sunny \wedge C=false) = P(W=sunny|C=false)\,P(C=false)$$

$$P(W=rain \wedge C=false) = P(W=rain|C=false)\,P(C=false)$$
$$P(W=cloudy \wedge C=false) = P(W=cloudy|C=false)\,P(C=false)$$
$$P(W=snow \wedge C=false) = P(W=snow|C=false)\,P(C=false)$$

公理

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

这个公式不难理解：*a* 成立的情况加上 *b* 成立的情况，无疑包含了 $a \vee b$ 成立的情况，但这两个集合相加将它们的交集计算了两次，所以需要减去 $P(a \wedge b)$。

完全联合分布

我们从一个简单例子开始：一个由三个布尔变量 *Toothache*，*Cavity* 以及 *Catch*（由于牙医的钢探针不洁而导致的牙龈感染）组成的问题域。其完全联合分布是一个 2 × 2 × 2 的表格，如图 13.3 所示。

	toothache		¬*toothache*	
	catch	¬*catch*	*catch*	¬*catch*
cavity	0.108	0.012	0.072	0.008
¬*cavity*	0.016	0.064	0.144	0.576

图 13.3　关于 *Toothache*，*Cavity*，*Catch* 世界的完全联合分布

注意，概率公理要求联合分布中的所有概率之和为 1。公式（13.2）为我们提供了计算任何命题（无论是简单命题还是复合命题）概率的一种直接方法：只需识别使命题为真的那些可能世界，然后把它们的概率加起来。例如，使命题 $cavity \vee toothache$ 成立的可能世界有 6 个：

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

一个特别常见的任务是提取关于随机变量的某个子集或者某单个变量的概率分布。例如，将图 13.3 中第一行的条目加起来就得到 *cavity* 的无条件概率，或者称为**边缘概率**[1]：

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

这个过程称为**边缘化**（marginalization），或者称**求和消元**（summing out）——因为是将除了 *Cavity* 以外的其他变量取每个可能值的概率相加，所以它们都被从公式中消除了。

概率演算

在多数情况下，我们会对已知一些变量的证据而计算另一些变量的条件概率感兴趣。条件概率可以如此计算：首先使用公式（13.3）得到一个基于无条件概率的表达式，然后再由完全联合分布对表达式求值。例如，已知有牙痛的证据，我们可以计算有牙洞的概率：

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

为了验算，我们还可以计算已知牙痛的证据时没有牙洞的概率：

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

这两个计算出来的值相加等于 1，应该如此。注意，这两次计算中的项 $1 / P(toothache)$ 是不变的，与我们计算的 *Cavity* 的值无关。事实上我们可以把它视为 **P**(*Cavity* | *toothache*) 的一个**归一化**（normalization）常数，保证其中的概率加起来等于 1。贯穿于处理概率的章节，我们将用 α 来表示这样的常数。有了这个符号，我们可以把前面的两个公式合并为一个：

$$\begin{aligned}\mathbf{P}(Cavity \mid toothache) &= \alpha\,\mathbf{P}(Cavity, toothache)\\ &= \alpha\,[\,\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)\,]\\ &= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle\end{aligned}$$

换句话说，即使我们并不知道 $P(toothache)$ 的值，我们仍可以计算出 **P**(*Cavity* | *toothache*)！我们暂时不管 $P(toothache)$，而对 *Cavity* 分别取 *cavity* 和 ¬*cavity* 时进行求和得到 0.12 和 0.08。这两个数代表了有关比例，但它们相加不等于 1，所以将这两个数都除以 0.12+0.08 而进行归一化。在许多概率演算中，归一化被证明是有用的捷径，不但使得计算更简单，而且在某些概率（如 $P(toothache)$）无法估算时可以使概率演算照样进行下去。

从这个例子，我们可以提取出一个通用推理过程。我们只考虑查询仅涉及一个变量的情况，假设这个变量为 *X*（这个例子中是 *Cavity*）。假设 **E** 为证据变量集合（这个例子中只有 *Toothache*），**e** 表示其观察值；并假设 **Y** 为其余的未观测变量（这个例子中是 *Catch*）。查询为 **P**(*X* | **e**)，它的值计算为：

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha\,\mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}) \qquad (13.9)$$

独立性

两个命题 a 和 b 之间独立可以写作：

$$P(a|b) = P(a) \quad \text{or} \quad P(b|a) = P(b) \quad \text{or} \quad P(a \wedge b) = P(a)P(b)$$

这个三种形式是等价的　　变量 X 和 Y 之间独立可以写作（三种形式也是等价的）：

$$\mathbf{P}(X|Y) = \mathbf{P}(X) \quad \text{或} \quad \mathbf{P}(Y|X) = \mathbf{P}(Y) \quad \text{或} \quad \mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$$

贝叶斯法则

$P(a \wedge b) = P(a \mid b)\,P(b)$　和　$P(a \wedge b) = P(b \mid a)\,P(a)$

这两个式子的右边相等，然后同时除以 $P(a)$，可得到

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$$

这个公式是著名的**贝叶斯规则**（Bayes' Rule；也称为贝叶斯定律，Bayes' Law；或者贝叶斯定理，Bayes' Theorem）。这个简单的公式是大多数进行概率推理的现代人工智能系统的基础。

对于多值随机变量的更一般情况，可以用符号 **P** 写成如下形式：

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\mathbf{P}(Y)}{\mathbf{P}(X)}$$

像前面一样，这个式子表示一组公式，每个公式处理变量的特定取值。还有某些场合，我们需要以某个背景证据 **e** 为条件的更通用的公式：

$$\mathbf{P}(Y \mid X, \mathbf{e}) = \frac{\mathbf{P}(X \mid Y, \mathbf{e})\mathbf{P}(Y \mid \mathbf{e})}{\mathbf{P}(X \mid \mathbf{e})}$$

在类似这样的医疗诊断中，我们经常有因果关系的条件概率（也就是说，医生知道 $P(symptoms \mid disease)$）而想得出诊断 $P(disease \mid symptoms)$。例如，医生知道脑膜炎会引起病人脖子僵硬，比如说有 70%的机会。医生还了解一些无条件事实：病人患脑膜炎的先验概率是 1/50 000，而任何一个病人脖子僵硬的先验概率为 1%。令 s 表示“病人脖子僵硬”的命题，m 表示“病人患有脑膜炎”的命题，则有

$$P(s \mid m) = 0.7$$
$$P(m) = 1 / 50000$$
$$P(s) = 0.01$$
$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014 \qquad (13.14)$$

也就是说，我们期望 700 个有脖子僵硬症状的病人中只有不到 1 个人患有脑膜炎。注意，尽管脑膜炎相当强烈地预示着会有脖子僵硬的症状（概率为 0.7），但脖子僵硬的病人患脑膜炎的概率却依然很低。这是因为脖子僵硬的先验概率大大高于患脑膜炎的先验概率。

13.3 节阐述了一个可以避免对证据的概率（此处是 $P(s)$）进行估算的过程，而只需要计算查询变量的每个值的后验概率（此处是 m 和 $\neg m$），然后对结果进行归一化。当使用贝叶斯规则时，同样可以应用这个过程。我们有：

$$\mathbf{P}(M \mid s) = \alpha \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) \rangle$$

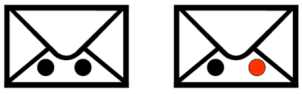
这样，为了使用这种方法我们需要估计 $P(s \mid \neg m)$而不是 $P(s)$。天下没有免费的午餐——有时这很容易，但有时却很困难。使用归一化的贝叶斯规则的一般形式为

$$\mathbf{P}(Y \mid X) = \alpha\, \mathbf{P}(X \mid Y)\, \mathbf{P}(Y) \qquad (13.15)$$

其中 α 是使 $\mathbf{P}(Y \mid X)$中所有条目总和为 1 所需的归一化常数。

例1

你有两个信封可供选择。一个信封里有一个红球（价值100美元）和一个黑球，另一个信封里有两个黑球（价值为零）。



你随机选择一个信封，然后从该信封中随机取出一个球，结果是黑色的。

此时，你可以选择是否换另一个信封。问题是，你应该换还是不换？

E: envelope, 1表示有一个红球的信封，2表示都是黑球的信封$1 = (R, B), 2 = (B, B)$

B: the event of drawing a black ball 拿到一个黑棋的事件

贝叶斯法则：

$$P(E|B) = \frac{P(B|E)P(E)}{P(B)}$$

We want to compare 比较：$P(E=1|B)$ vs. $P(E=2|B)$

在红球信封拿到黑球：$P(B|E=1) = 0.5$

在黑球信封拿到黑球：$P(B|E=2) = 1$

拿到1、2信封的概率相同：$P(E=1) = P(E=2) = 0.5$

抽到黑球的概率：

B在E取值上的边缘概率

$$\begin{aligned} P(B) &= P(B|E=1)P(E=1) + P(B|E=2)P(E=2) \\ &= (0.5)(0.5) + (1)(0.5) \\ &= 0.75 \end{aligned}$$

已经抽到一个黑球，此信封是红球信封的概率：

$$P(E=1|B) = \frac{P(B|E=1)P(E=1)}{P(B)} = \frac{(0.5)(0.5)}{(0.75)} = \frac{1}{3}$$

已经抽到一个黑球，此信封是黑球信封的概率：

$$P(E=2|B) = \frac{P(B|E=2)P(E=2)}{P(B)} = \frac{(1)(0.5)}{(0.75)} = \frac{2}{3}$$

通过计算可得，抽到黑球后信封为1的概率为$1/3$，

信封为2的概率为$2/3$。因此，更换信封可以提高获得红球的概率。

例2

一位医生进行一项测试，该测试有99%的可靠性，即99%的生病者测试结果为阳性，99%的健康者测试结果为阴性。这位医生估计整个人口中有1%的人是生病的。
因此，对于测试结果为阳性的患者，他是生病的概率是多少呢?

我们可以使用贝叶斯定理来计算患者生病的条件概率。设事件 S 表示患者生病，事件 T 表示测试结果为阳性。则所求的条件概率为：

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

其中，$P(T|S)$ 表示患者生病的条件下，测试结果为阳性的概率，$P(S)$ 表示患者生病的**先验概率**，$P(T)$ 表示测试结果为阳性的概率。

根据题目中给出的数据，我们有：

$$P(T|S) = 0.99$$
$$P(S) = 0.01$$
$$P(T) = P(T|S)P(S) + P(T|\overline{S})P(\overline{S})$$

其中，$\overline{S}$表示患者不生病。

根据测试的可靠性，我们可以得到

$$P(T|\overline{S}) = 1 - P(T|S) = 0.01$$

因此

$$\begin{aligned} P(T) &= P(T|S)P(S) + P(T|\overline{S})P(\overline{S}) \\ &= (0.99)(0.01) + (0.01)(0.99) \\ &= 0.0198 \end{aligned}$$

代入贝叶斯公式，我们可以计算出患者生病的条件概率：

$$P(S|T) = \frac{(0.99)(0.01)}{0.0198} \approx 0.50$$

因此，测试结果为阳性的患者生病的概率约为50

使用贝叶斯规则：合并证据

当我们有两条或者更多证据时会怎么样？例如，如果牙医的不清洁的钢探针引起病人疼痛的牙齿感染，她能得出什么结论？如果我们知道完全联合分布（图 13.3），则可以读出答案：

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha \langle 0.108, 0.016 \rangle \approx \langle 0.871, 0.129 \rangle$$

然而我们知道，这种方法不能扩展到有大量的变量的情况。我们也可以试着使用贝叶斯规则重新对问题形式化：

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\, \mathbf{P}(toothache \wedge catch \mid Cavity)\, \mathbf{P}(Cavity) \quad (13.16)$$

为了使用这个式子，我们需要知道在 *Cavity* 每个取值下合取式 $toothache \wedge catch$ 的条件概率。这对于只包含两个证据变量的情形可能是可行的，但同样不允许变量太多。如果有 n 个可能的证据变量（X 射线透视、日常饮食、卫生保健、等等），观察到的值就有 2^n 个可能组合，我们需要知道每个可能组合下的条件概率。我们也许还不如回到使用完全联合分布的方法。这是最初导致研究人员远离概率理论而寻求对证据进行组合的近似方法的原因，虽然近似方法可能给出不正确的答案，但为了得到任何答案需要的数据量更少。

如果我们不采用这条路线，那么就需要找到关于问题域的、使我们能够简化表达式的附加断言。13.4 节介绍的**独立性**的概念提供了一条线索，但需要完善。要是 *Toothache* 和 *Catch* 彼此独立就好了，但是它们并非如此：如果探针引起牙齿感染，那么牙齿可能有洞，而这个牙洞引起牙痛。不过，如果已知病人是否有牙洞，这两个变量就是相互独立的。每个变量取值都是由牙洞导致的，但是它们彼此之间没有直接影响：牙痛依赖于牙神经的状态，而使用探针的精确度取决于牙医的技术，牙痛与此无关 [1]。数学上，这个性质可以写作：

$$\mathbf{P}(toothache \wedge catch \mid Cavity) = \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity) \quad (13.17)$$

这个公式表达了当给定 *Cavity* 时 *toothache* 和 *catch* 的**条件独立性**（conditional independence）。我们可以把它代入到公式（13.16）中得到有牙洞的概率：

$$\begin{aligned} &\mathbf{P}(Cavity \mid toothache \wedge catch) \\ &= \alpha\, \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity)\, \mathbf{P}(Cavity) \end{aligned} \quad (13.18)$$

这时所需的信息就和单独使用每条证据进行推理是一样的了：查询变量的先验概率 $\mathbf{P}(Cavity)$，以及给定原因下各种结果的条件概率。

给定第三个随机变量 Z 后，两个随机变量 X 和 Y 的条件独立性的一般定义是：

$$\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\, \mathbf{P}(Y \mid Z)$$

例如在牙科问题域中，给定 *Cavity*，断言变量 *Toothache* 和 *Catch* 的条件独立性看来是合理的：

$$\mathbf{P}(Toothache,\ Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\, \mathbf{P}(Catch \mid Cavity) \quad (13.19)$$

注意这个断言比公式（13.17）要强一些，公式（13.17）只断言了 *Toothache* 和 *Catch* 在特定取值下的条件独立性。与公式（13.11）所表达的绝对独立性相类似，也可以使用以下条件独立性的等价形式（习题 13.17）：

$$\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z) \quad 和 \quad \mathbf{P}(Y \mid X, Z) = \mathbf{P}(Y \mid Z)$$

13.4 节说明，绝对独立性断言允许将完全联合分布分解成很多更小的分布。这对于条件独立性断言同样也是成立的。例如，给定公式（13.19）中的断言，我们得到如下分解形式：

$$\begin{aligned} &\mathbf{P}(Toothache, Catch, Cavity) \\ &= \mathbf{P}(Toothache, Catch \mid Cavity)\, \mathbf{P}(Cavity) && （乘法原则） \\ &= \mathbf{P}(Toothache \mid Cavity)\, \mathbf{P}(Catch \mid Cavity)\, \mathbf{P}(Cavity) && （使用公式（13.19）） \end{aligned}$$

朴素贝叶斯

这个牙科的例子说明了一类普遍存在的模式，其中单一原因直接影响许多结果，这些结果在给定这个原因时都是彼此条件独立的。这时，完全联合分布可以写为：

$$\mathbf{P}(Cause, Effect_1, \cdots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

这样的一个概率分布被称为一个**朴素贝叶斯**（naive Bayes）模型——“朴素”是因为这个模型经常用于（作为模型的简化假设）“结果”变量在给定原因变量下实际上不是条件独立的情况。（朴素贝叶斯模型有时被称为**贝叶斯分类器**（Bayesian classifier），一个多少有些欠考虑的用法，因此一些真正的贝叶斯支持者们将其称为**傻瓜贝叶斯**（idiot Bayes）模型。）在实际中，基于朴素贝叶斯模型的系统工作得出奇地好——即使条件独立性假设不成立时。

最大似然估计

最大似然估计（Maximum Likelihood Estimation，简称MLE）是一种常用的参数估计方法，用于根据已知的样本数据来估计模型的参数。它的核心思想是选择能够使观测到的数据出现的概率最大的参数作为估计值。

具体来说，在最大似然估计中，我们假设样本数据来自于某个概率分布，但是该分布的参数是未知的。我们的目标是通过样本数据来估计这些参数，使得该分布能够最好地解释观测到的数据。

假设我们有一个样本集合 $X = x_1, x_2, \ldots, x_n$，每个样本都是来自于某个分布 $f(x|\theta)$ 的观测值，其中 θ 是分布的参数。我们要找到能够最大化样本集合 X 的联合概率密度函数 $L(X|\theta)$ 的参数值 θ。这个联合概率密度函数可以表示为：

$L(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$

我们的目标是找到能够最大化 $L(X|\theta)$ 的 θ 值。因此，最大似然估计的计算可以表示为：

$\hat{\theta}_{MLE} = \arg\max_{\theta} L(X|\theta)$

有时候我们需要对上式取对数来避免计算机计算下溢，得到的式子为：

$\hat{\theta}_{MLE} = \arg\max_{\theta} \log L(X|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta)$

最大似然估计方法是一种常用的参数估计方法，具有计算简单、理论基础好等优点。它在统计学、机器学习、信号处理等领域都得到了广泛应用。

小结

以下是对概率论中重要的公式的整理：

---

1. 条件概率公式：

对于事件 A 和事件 B，其条件概率表示为 $P(A|B)$，表示在事件 B 发生的条件下，事件 A 发生的概率。条件概率公式为：

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

---

2. 乘法规则公式：

对于事件 A 和事件 B，其联合概率表示为 $P(A,B)$，表示事件 A 和事件 B 同时发生的概率。乘法规则公式为：

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

---

3. 链式规则公式：

对于多个事件 A, B, C, D，其联合概率表示为 $P(A,B,C,D)$，链式规则公式可以表示为：

$P(A,B,C,D) = P(A|B,C,D)P(B|C,D)P(C|D)P(D)$

---

4. 条件化的链式规则公式：

对于事件 A 和事件 B，其联合概率表示为 $P(A,B)$，条件化的链式规则公式可以表示为：

$P(A,B|C) = P(A|B,C)P(B|C)$

$P(A,B|C)P(C) = P(A,B,C)$

$\frac{P(A|B,C)P(B|C)P(B,C)}{P(B|C)} = P(A,B,C)$

$\mathbf{P(A,B|C)} = \frac{P(A|B,C)P(B|C)P(B,C)}{P(B|C)P(C)}$

$= \frac{P(A|B,C)P(B|C)P(B|C)}{P(B|C)} = \mathbf{P(A|B,C)P(B|C)}$

5. 贝叶斯定理公式：

贝叶斯定理是根据先验概率和条件概率来计算后验概率的一种方法，可以用于分类、预测等任务。贝叶斯定理公式为：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

---

6. 条件化的贝叶斯定理公式：

对于事件 A 和事件 B，条件化的贝叶斯定理公式可以表示为：

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$$

---

7. 加法/条件概率公式：

对于事件 A 和事件 B，加法/条件概率公式可以表示为：

$$P(A) = P(A,B) + P(A,\neg B) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$$

这些公式在概率论中非常重要，可以应用于统计学、机器学习、信号处理、金融领域、医学领域等各个领域的问题中。熟练掌握这些公式可以帮助我们更好地理解和解决实际问题。