

作者	生姜 DrGinger
脚本	生姜 DrGinger
视频	崔崔 CuiCui
开源学习资源	https://github.com/Visualize-ML
平台	https://www.youtube.com/@DrGinger_Jiang https://space.bilibili.com/3546865719052873 https://space.bilibili.com/513194466

1.4 贝叶斯定理



本节你将掌握的核心技能：

- ▶ 掌握在已知事件发生的前提下重新评估另一个事件概率；
- ▶ 根据已知事件收缩样本空间并在子空间内重新评估概率；
- ▶ 联合概率与条件概率关系；
- ▶ 利用互不相容事件对样本空间的分割，将一个事件的概率分解为所有可能路径的加权和；
- ▶ 根据先验概率和新证据更新后验概率，体现“已知条件下推测未知”的核心思想。

生活中条件概率无处不在

在现实生活中，**条件概率** (conditional probability) 几乎无处不在。

所谓**条件概率**，是指在已知某个事件已经发生的前提下，计算另一个事件发生的概率。**条件概率**描述的是在获取新信息后，我们对事件发生可能性的更新判断。

数学上，这种思维方式的核心，是通过引入“条件”来缩小样本空间，从而重新计算概率。

下面举几个例子。

假设我们关心某所学校学生身高超过 180 厘米的概率。如果没有任何附加信息，我们可以直接统计学生身高情况。

但如果我们想知道男生身高超过 180 厘米的概率，那么样本空间就缩小到了“男生”这一群体，身高超过 180 厘米的概率就变成了“在已知学生是男生的条件下，身高超过 180 厘米”的条件概率。

再比如，某门职业考试的总体通过率是 60%。这意味着随机选一个学生，他通过考试的概率是 0.6。然而，如果我们进一步知道这个学生参加了补习班，那么“他通过考试”的概率可能更高。这里，“参加补习班”是条件，它改变了我们评估“通过考试”这件事的基础。于是，“在已知参加补习班的前提下，通过考试”的概率就是一个条件概率。

条件概率同样适用于日常生活中看似普通的场景。例如，在一个事故频发的路口，“夜间发生交通事故”的概率往往比白天更高。这里“夜间”这一条件限定了环境。

甚至在最经典的概率实验中也能看到条件概率的影子。掷一颗色子时，如果我们事先知道点数是偶数，那么样本空间只剩下 $\{2, 4, 6\}$ 。此时，在“点数为偶数”的条件下得到点数 6 的概率就不再是 $1/6$ ，而是 $1/3$ 。

总的来说，条件概率是一种在获得额外信息后重新评估不确定性的方式。它体现了概率论中“知识改变信念”的思想——当我们获得新的条件信息时，我们对事件的判断也随之更新。掌握条件概率不仅是学习统计和机器学习的基础，更是理解世界中各种“已知条件下推测未知”问题的关键。

下面，就让我们用一节内容聊聊条件概率，以及贝叶斯定理。

B 发生条件下，A 发生的概率

条件概率描述的是这样一种情形：在我们已经知道某个事件发生的前提下，另一个事件发生的可能性会变成多少。

⚠ 注意，条件概率本身并不要求事件有时间先后关系，更不要求因果关系。它只是条件信息更新。在本册最后大家会看到具有时间先后顺序的条件概率计算。

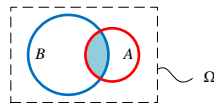


图 1. A 和 B 为样本空间 Ω 中的两个事件

如图 1 所示，A 和 B 为样本空间 Ω 中的两个事件，其中 $\Pr(B) > 0$ 。事件 B 发生的条件下事件 A 发生的条件概率可以通过下式计算得到：

$$\underbrace{\Pr(A|B)}_{\text{Conditional}} = \frac{\overbrace{\Pr(A, B)}^{\text{Joint}}}{\underbrace{\Pr(B)}_{\text{Marginal}}} \quad (1)$$

其中， $\Pr(A, B)$ 为 A 和 B 事件的联合概率，表示 A 和 B 事件同时发生的概率； $\Pr(B)$ 也叫 B 事件**边缘概率** (marginal probability)。

⚠ 注意， $\Pr(B)$ 、 $\Pr(A, B)$ 的样本空间都是 Ω 。

这一公式的直观含义是：当我们只考虑那些“B 已经发生”的情况时，我们需要把样本空间缩小，只在 B 的范围内重新计算 A 的概率。

此时，如图 2 所示，事件 B 所对应的区域就成了一个新的样本空间，记作 Ω_B 。因此，条件概率 $\Pr(A|B)$ 实际上是事件 A 在这个新的样本空间 Ω_B 中的概率。

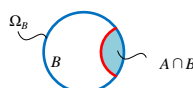


图 2. 条件概率 $\Pr(A|B)$ 的样本空间变为 Ω_B

换句话说，条件概率反映了“局部视角”下的概率计算。当我们得到新的信息，比如事件 B 已经发生，就相当于我们不再从整个 Ω 观察，而是只在 Ω_B 这个子集中重新评估事件 A 的可能性。

⚠ 注意， $\Pr(B)$ 是在整个样本空间 Ω 上定义的。因此，我们也可以将其写作 $\Pr(B|\Omega)$ ，表示在整个样本空间中 B 发生的概率。

从几何直观来看，可以把样本空间 Ω 想象成一块区域，而事件 A 与 B 是其中的两个子区域。当 B 已知发生时，我们只看 B 这部分区域，此时 $A \cap B$ 的面积占 B 的面积的比例，就是 $\Pr(A|B)$ 。

这正是条件概率的本质：在限定信息下重新度量事件发生的可能性。

反过来看，我们可以把 (1) 写成

$$\underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(B)}_{\text{Marginal}} \underbrace{\Pr(A|B)}_{\text{Conditional}} \quad (2)$$

也就是说，联合概率 $\Pr(A, B)$ 等于 $\Pr(B)$ 乘条件概率 $\Pr(A|B)$ 。

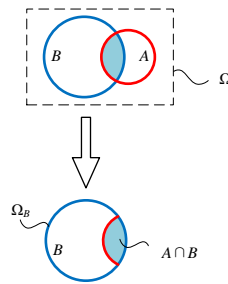


图 3. 逐层缩小观察范围

从直观上看，(2) 这个计算过程就像在样本空间中“逐层改变观察范围”，具体如图 3 所示。第一步，我们在整个样本空间 Ω 中找到事件 B 对应的区域；第二步，我们在这块区域 Ω_B 内再考察事件 A 所占的比例。两步结合起来，就得到了 A 与 B 同时出现的概率。

因此，式 (2) 的意义可以理解为，我们先在样本空间 Ω 中“挑出”事件 B ，计算得到 $\Pr(B)$ ；再在 B 的范围中“挑出”事件 A ，计算得到 $\Pr(A|B)$ 。将两者相乘，就能得到 A 与 B 同时发生的概率，即联合概率 $\Pr(A, B)$ 。

A 发生条件下，B 发生的概率

类似地，事件 A 发生的条件下事件 B 发生的条件概率为：

$$\underbrace{\Pr(B|A)}_{\text{Conditional}} = \frac{\overbrace{\Pr(A, B)}^{\text{Joint}}}{\underbrace{\Pr(A)}_{\text{Marginal}}} \quad (3)$$

其中， $\Pr(A)$ 为 A 事件边缘概率， $\Pr(A) > 0$ 。

当我们说“在 A 发生的条件下， B 发生的概率”时，强调的是：我们对世界的了解已经发生了变化，因为 A 的发生减少了原来的不确定性，使得样本空间缩小成“只包含 A 发生时所有可能结果”的那一部分。这个被缩小后的样本空间通常记为 Ω_A ，它只保留满足事件 A 的所有样本点。

如图 4 所示，条件概率 $\Pr(B|A)$ 就可以理解为：在新的样本空间 Ω_A 中，事件 B 占据了多大的比例。原本 B 的概率需要在整个样本空间 Ω 中计算，但现在我们必须在一个更狭窄的条件 (Ω_A) 下重新评估它。

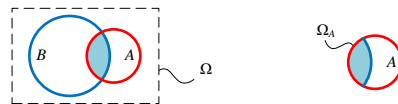


图 4. 条件概率 $\Pr(B|A)$ 的样本空间变为 Ω_A

(3) 也可以写成

$$\underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(A)}_{\text{Marginal}} \underbrace{\Pr(B|A)}_{\text{Conditional}} \quad (4)$$

举个例子

为了更直观地理解条件概率，我们来看一个最常见的例子——掷一颗均匀色子。假设色子每一面出现的可能性完全相同，原始样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 。

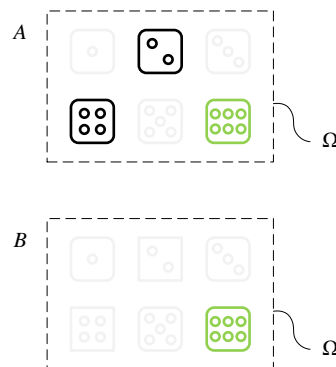


图 5. 掷一颗均匀色子，事件 A 为结果点数为偶数；事件 B 为结果点数为 6

现在，我们引入事件 A ：点数为偶数。也就是说，只要结果是 2、4 或 6，都属于事件 A 。根据等概率原理，事件 A 对应的概率为

$$\Pr(A) = \frac{3}{6} = \frac{1}{2} \quad (5)$$

接下来定义事件 B ：点数为 6。单独来看，事件 B 只有一个样本点，即 $\{6\}$ 。如果不考虑任何额外信息，那么从六个等可能的结果中得到 6 的概率是 $1/6$ ，即

$$\Pr(B) = \frac{1}{6} \quad (6)$$

图 5 直观展示了事件 A 和事件 B 在原始样本空间中的位置。

我们发现 B 是 A 的子集，对于这一特殊情况

$$\Pr(A, B) = \Pr(B) = \frac{1}{6} \quad (7)$$

如果想要计算“点数是偶数 (A) 条件下结果为 6 (B) 的条件概率”，我们就需要本节前文介绍的数学工具。

由于给定了“点数是偶数”这一额外信息，我们对世界的认知发生了改变，原本的六种可能结果被缩小成三个，因此新的样本空间变成了 $\{2, 4, 6\}$ 。这也意味着接下来的所有概率计算都必须在这个新的背景下进行。

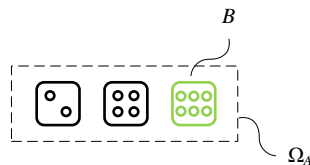


图 6. 掷一颗均匀色子，事件 A 为条件看事件 B

如图 6 所示，一旦知道事件 A 已经发生，就等于我们已经排除了 1、3、5 这三个奇数点数，色子的可能结果只剩三个，因此 6 的相对位置发生了变化。在新的样本空间 $\{2, 4, 6\}$ 中，事件 B 只有一个样本点，因此在“点数为偶数”的条件下得到 6 的概率变为 $1/3$ ，即

$$\Pr(B|A) = \frac{\Pr(A, B)}{\Pr(A)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3} \quad (8)$$

图 6 展示了当 A 已经发生时，对事件 B 的观察方式如何改变。

为了再进一步加深理解，我们再定义事件 C ：点数小于 3，也就是 $\{1, 2\}$ 。

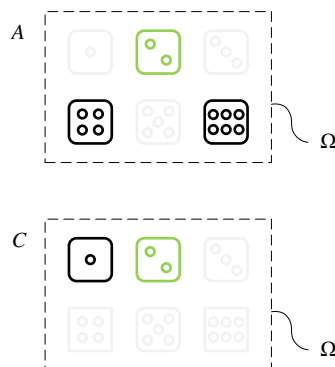


图 7. 掷一颗均匀色子，事件 A 为结果点数为偶数；事件 C 为结果点数小于 3

图7显示了事件 A (点数为偶数) 与事件 C (点数小于 3) 在原始样本空间中的关系。很容易看出, A 和 C 之间的重叠只有一个点, 即 2。这意味着,

$$\Pr(A, C) = \frac{1}{6} \quad (9)$$

如果我们依旧以事件 A 已经发生作为条件, 那么新的样本空间仍然是 $\{2, 4, 6\}$ 。在这样的背景下观察事件 C , 你会发现图8表达的内容非常直接: 事件 C 在新的样本空间中其实只有一个可能结果 2, 因此 $\Pr(C|A)$ 变成了 $1/3$, 即

$$\Pr(C|A) = \frac{\Pr(A, C)}{\Pr(A)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3} \quad (10)$$

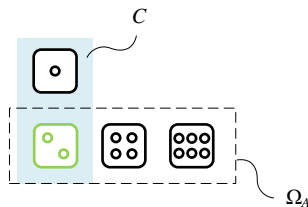


图8. 掷一颗均匀色子, 事件 A 为条件看事件 C

为了完整地展示条件概率在不同类型事件上的表现, 我们再来看最后一个例子。定义事件 D : 点数为奇数, 也就是 $\{1, 3, 5\}$ 。

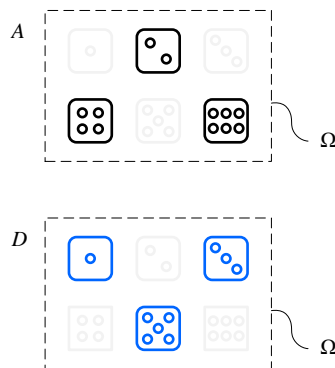


图9. 掷一颗均匀色子, 事件 A 为结果点数为偶数; 事件 D 为结果点数为奇数

图9展示了事件 A (点数为偶数) 与事件 D (点数为奇数) 在原始样本空间中的位置。由于偶数与奇数本身就是互斥的, 两个事件之间没有任何重叠部分。换句话说, A 和 D 不可能同时发生, 它们的交集为空集。这意味着,

$$\Pr(A, D) = \frac{0}{6} \quad (11)$$

只要色子掷出的结果是偶数，它一定不会是奇数；反之亦然。因此，事件 A 和事件 D 属于典型的互斥事件。

接下来，我们依旧将“事件 A 已经发生”作为新的条件。在这个前提下，新的样本空间依然缩小为 $\{2, 4, 6\}$ 。当我们在这样的背景下去观察事件 D 时，会发现图 10 所展示的状况非常直观：事件 D 的所有样本点 $\{1, 3, 5\}$ 在新的样本空间中完全消失了，没有一个落在 $\{2, 4, 6\}$ 内。

在新的样本空间中找不到任何属于事件 D 的结果，也就意味着事件 D 在条件 A 下发生的概率为 0，即 $\Pr(D|A)$ 为 0，对应运算为

$$\Pr(D|A) = \frac{\Pr(A, D)}{\Pr(A)} = \frac{0}{\frac{1}{2}} = 0 \quad (12)$$

也就是说，只要我们已知点数是偶数，那么“点数是奇数”这一事件就失去了任何可能性。

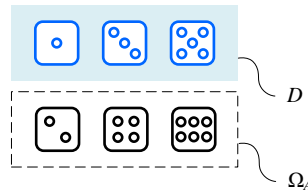


图 10. 掷一颗均匀色子，事件 A 为条件看事件 D

这个例子说明了一个重要的思想：条件概率不是改变事件本身，而是改变我们看事件的“视角”。当我们获得新的信息（例如知道点数是偶数）时，评估概率所依赖的样本空间会随之改变，这个改变直接影响条件概率。

当已知点数为偶数时，原来的样本空间 $\{1, 2, 3, 4, 5, 6\}$ 缩小为 $\{2, 4, 6\}$ 。在新的样本空间中，共有 3 个等可能的结果，其中只有一个是 6。因此条件概率为 $1/3$ 。如图 11 所示，根据大数定律，当试验次数 n 足够大时，事件发生的频率趋近于其理论概率。

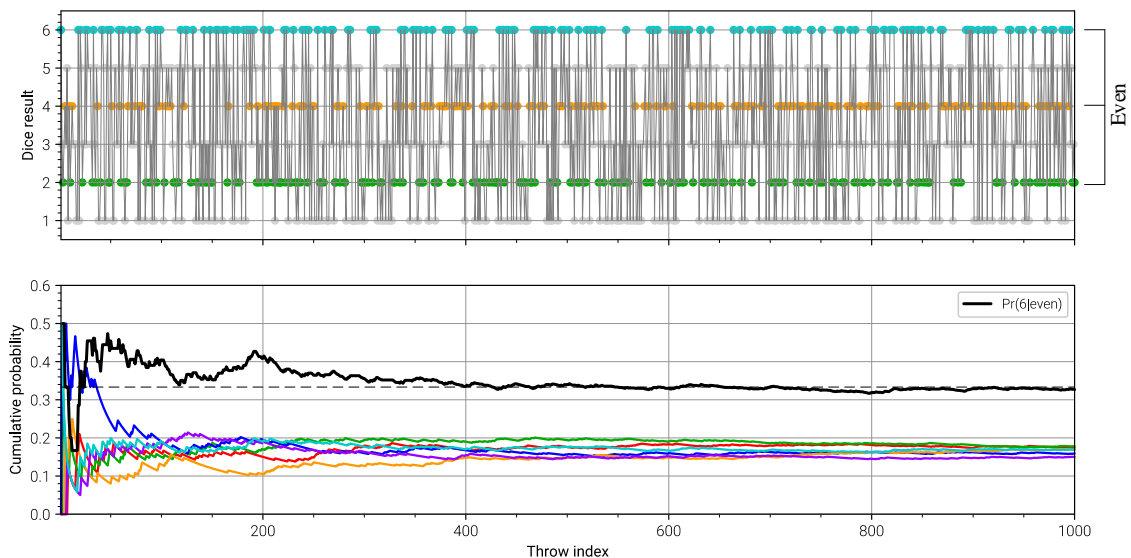


图 11. 掷一颗色子 500 次，模拟 $\Pr(6 | \text{even})$ 条件概率

贝叶斯定理

在介绍了条件概率之后，我们就能够顺理成章地推导出**贝叶斯定理** (Bayes' theorem)。

由前面的内容可知，事件 A 和事件 B 的联合概率既可以写成 $\Pr(A, B) = \Pr(A)\Pr(B|A)$ ，也可以写成 $\Pr(A, B) = \Pr(B)\Pr(A|B)$ 。将这两个表达式联立起来，便得到等式

$$\underbrace{\Pr(A|B)}_{\text{Conditional}} \underbrace{\Pr(B)}_{\text{Marginal}} = \underbrace{\Pr(B|A)}_{\text{Conditional}} \underbrace{\Pr(A)}_{\text{Marginal}} = \underbrace{\Pr(A, B)}_{\text{Joint}} \quad (13)$$

回顾上式的每个概率值：

- ◀ $\Pr(A|B)$ 是指在 B 发生条件下 A 发生的**条件概率** (conditional probability)；也就是说， $\Pr(A|B)$ 的样本空间为 Ω_B ；
- ◀ $\Pr(B|A)$ 是指在 A 发生条件下 B 发生的**条件概率**；也就是说， $\Pr(B|A)$ 的样本空间为 Ω_A ；
- ◀ $\Pr(A)$ 是 A 的**边缘概率** (marginal probability)，对应的样本空间为 Ω ；
- ◀ $\Pr(B)$ 是 B 的**边缘概率**，对应的样本空间也是 Ω ；
- ◀ $\Pr(A, B)$ 是事件 A 和 B 的**联合概率**，样本空间为 Ω 。

对 (13) 稍作调整，便得到著名的**贝叶斯定理** (Bayes' theorem)：

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \quad (14)$$

这一定理由十八世纪的数学家**托马斯·贝叶斯** (Thomas Bayes) 提出。虽然它的形式看上去非常简洁，但却在现代统计推断和机器学习中发挥着深远的影响。可以说，机器学习和深度学习中相当一部分概率模型，都直接或间接依赖这一基本等式。

如图 12 所示，贝叶斯定理的基本思想是根据**先验概率** (prior) 和新的**证据** (evidence) 来计算**后验概率** (posterior)。在实际应用中，我们通常根据一些已知的先验知识，来计算事件的先验概率。然后，当我们获取新的证据时，就可以利用贝叶斯定理来计算事件的后验概率，从而更新我们的信念或概率。

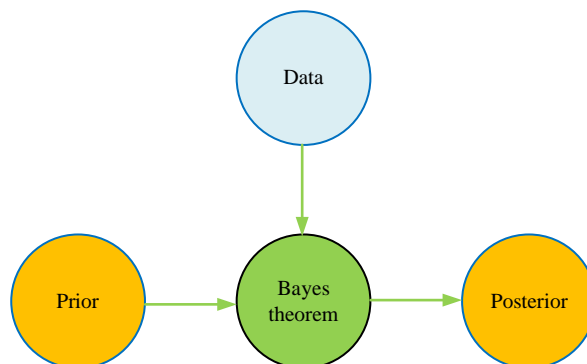


图 12. 贝叶斯推断



本书最后专门介绍如何利用贝叶斯定理完成贝叶斯推断。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

全概率定理

在理解了条件概率与贝叶斯定理之后，我们可以更深入地讨论另一个经常与它们同时出现的重要工具：**全概率定理** (Law of total probability)。为了说明它的意义，我们首先回到样本空间 Ω 的结构。

如果一组事件 A_1, A_2, \dots, A_n 两两互不重叠，并且它们的并集恰好覆盖了整个样本空间，那么我们就称这组事件对 Ω 构成了一个**分割** (partition)。

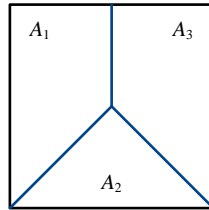


图 13. A_1, A_2, A_3 对样本空间 Ω 构成分割

大白话来说， Ω 就像是一个月饼，而 A_1, A_2, \dots, A_n 就是把这一个月饼切开的几块。

这些“切法”有两个关键要求：第一，不能重叠（每一块都是独立的一块，不能你这块里有一部分同时也属于另一块）；第二，必须切完整（不能有哪一小块月饼没被分到任何一份里）。

所以结果就是，这一个月饼被干干净净地切成了 n 块，不多不少、不重不漏。无论你随便取月饼上的一个点，它一定落在某一块里，而且只会落在这一块里，不可能同时属于两块。

换回概率的语言，就是说：无论随机试验的结果是什么，它一定对应其中某一个事件，而且只对应这一个事件。也正因为这样，这种“分割”特别有用——我们可以把复杂的问题拆开，在每一块里分别分析，然后再把结果加起来。

如图 14 所示，掷一颗均匀色子时， A_1 表示结果为奇数， A_2 表示结果为偶数；显然，事件 A_1, A_2 对样本空间 Ω 构成分割，即

$$A_1 \cap A_2 = \emptyset, A_1 \cup A_2 = \Omega \quad (15)$$

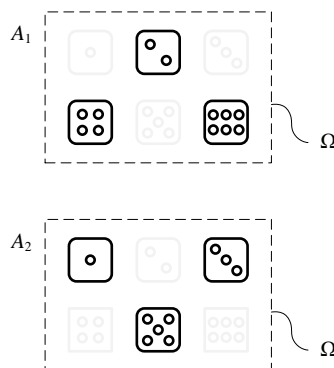


图 14. 事件 A_1, A_2 对样本空间 Ω 构成分割

在这样的前提下，只要 $\Pr(A_i) > 0$ ，那么对于样本空间 Ω 中的任意事件 B ，都可以通过下式计算其概率：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\underbrace{\Pr(B)}_{\text{Marginal}} = \sum_{i=1}^n \underbrace{\Pr(A_i, B)}_{\text{Joint}} = \Pr(A_1, B) + \Pr(A_2, B) + \cdots + \Pr(A_n, B) \quad (16)$$

这就是全概率定理。

虽然形式上只是一个求和公式，但它的本质是“从所有可能路径进行穷举”。每一项 $\Pr(A_i)\Pr(B|A_i)$ 都代表“先落入 A_i ，再在 A_i 的条件下发生 B ”的整体概率，而对全部路径求和，就是将所有可能导致 B 的方式全部考虑进去。

举个例子，图 15 给出的例子是三个互不相容事件 A_1 、 A_2 、 A_3 对 Ω 形成分割。通过全概率定理，我们可以将 $\Pr(B)$ 写成由三条路径组成的加总，每条路径都以某个 A_i 为起点，从而穷举得到完整的结果

$$\underbrace{\Pr(B)}_{\text{Marginal}} = \underbrace{\Pr(A_1, B)}_{\text{Joint}} + \underbrace{\Pr(A_2, B)}_{\text{Joint}} + \underbrace{\Pr(A_3, B)}_{\text{Joint}} \quad (17)$$

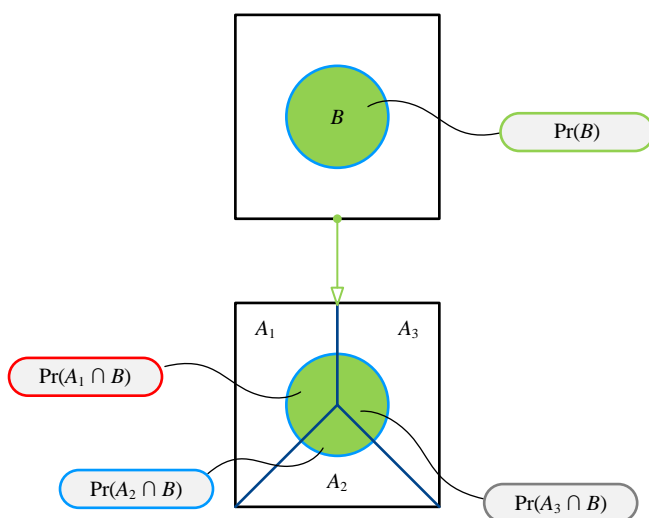


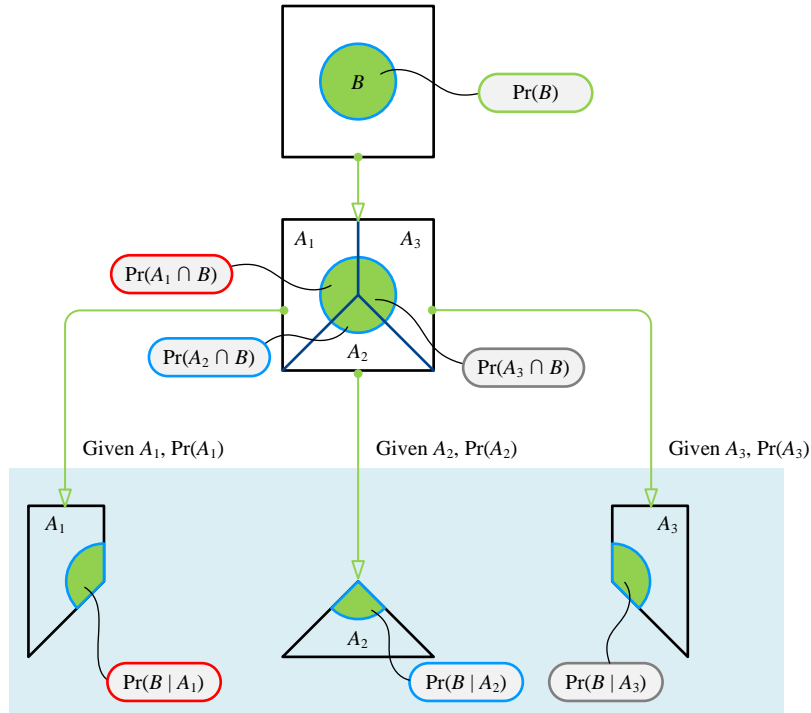
图 15. A_1, A_2, A_3 对空间 Ω 分割

引入贝叶斯定理

在此基础上，我们可以再次引入贝叶斯定理。将贝叶斯定理写成 $\Pr(A_i)\Pr(B|A_i)$ 的形式代入全概率公式，就可以得到 $\Pr(B)$ 的另一种展开方式

$$\begin{aligned} \Pr(B) &= \sum_{i=1}^n \underbrace{\Pr(A_i, B)}_{\text{Joint}} \\ &= \sum_{i=1}^n \underbrace{\Pr(B|A_i)}_{\text{Conditional}} \underbrace{\Pr(A_i)}_{\text{Marginal}} \\ &= \Pr(B|A_1)\Pr(A_1) + \Pr(B|A_2)\Pr(A_2) + \cdots + \Pr(B|A_n)\Pr(A_n) \end{aligned} \quad (18)$$

图 16 所示为分别给定 A_1, A_2, A_3 条件下，事件 B 是如何在各个分支上发生的。

图 16. 分别给定 A_1, A_2, A_3 条件下，事件 B 发生的情况

更有意义的是，我们也可以反过来，用贝叶斯定理来求在事件 B 已经发生的前提下 ($\Pr(B) > 0$)，各个 A_i 的条件概率。换句话说，当我们观察到 B 发生之后，原本的“分割”之间的平衡会发生变化，不同 A_i 发生的可能性会根据它们与 B 的关系被重新加权。由此得到的公式如下：

$$\Pr(A_i | B) = \frac{\Pr(A_i, B)}{\Pr(B)} = \frac{\Pr(B | A_i) \cdot \Pr(A_i)}{\Pr(B)} \quad (19)$$

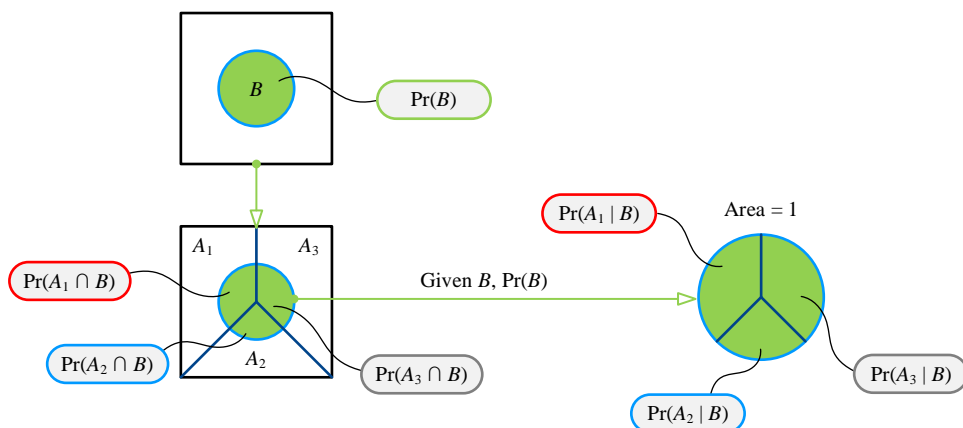
利用贝叶斯定理，以为 B 条件，进一步展开得到：

$$\begin{aligned} \Pr(B) &= \sum_{i=1}^n \underbrace{\Pr(A_i, B)}_{\text{Joint}} = \sum_{i=1}^n \underbrace{\Pr(A_i | B)}_{\text{Conditional}} \underbrace{\Pr(B)}_{\text{Marginal}} \\ &= \Pr(A_1 | B) \Pr(B) + \Pr(A_2 | B) \Pr(B) + \dots + \Pr(A_n | B) \Pr(B) \end{aligned} \quad (20)$$

(20) 等式左右消去 $\Pr(B)$ ($\Pr(B) > 0$)，得到：

$$\sum_{i=1}^n \Pr(A_i | B) = \Pr(A_1 | B) + \Pr(A_2 | B) + \dots + \Pr(A_n | B) = 1 \quad (21)$$

图 17 所示为给定 B 条件下，事件 A_1, A_2, A_3 发生的情况。

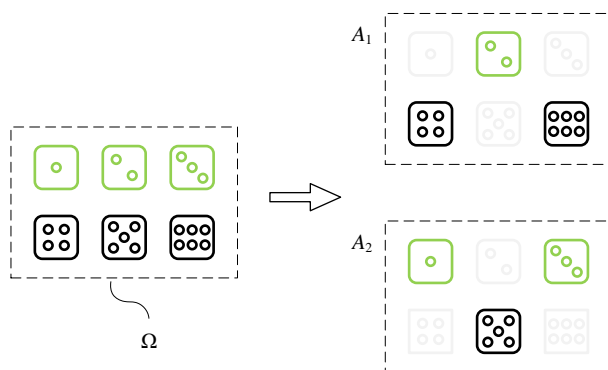
图 17. 给定 B 条件下，事件 A_1 、 A_2 、 A_3 发生的情况

类似前文，把整个样本空间 Ω 想象成一只完整的月饼，而事件 B 则是月饼中心的蛋黄。当我们用一组不相容事件 A_1, A_2, \dots, A_n 对整个样本空间 Ω 进行了完整、无重叠的切分时，就好像把月饼切成几块，每一块都属于这只月饼，没有遗漏也没有交叠。

由于蛋黄本身就是月饼的一部分，因此同样的切法也自然把蛋黄切成对应的几块：蛋黄中的每一部分都恰好落在某一块月饼切块里，不会落到外面。换句话说，既然 A_1, A_2, \dots, A_n 完整地划分了整个月饼，那么它们必然也完整地划分了蛋黄。

为了进一步帮助理解，我们回到掷一颗均匀色子这个例子。 A_1 表示点数为奇数，即 $\{1, 3, 5\}$ ； A_2 表示点数为偶数，即 $\{2, 4, 6\}$ ，两者对应的概率为

$$\begin{aligned} \Pr(A_1) &= \frac{1}{2} \\ \Pr(A_2) &= \frac{1}{2} \end{aligned} \quad (22)$$

图 18. A_1 、 A_2 对样本空间 Ω 进行分割，也完成了对 B 的分割

前文提过，这两个事件完全覆盖整个样本空间且互不重叠，因此构成对 Ω 的一次完整分割。再定义事件 B 为点数小于 4，即 $\{1, 2, 3\}$ ， B 对应的概率为

$$\Pr(B) = \frac{1}{2} \quad (23)$$

根据图 18，我们很容易计算条件概率

$$\begin{aligned}\Pr(B|A_1) &= \frac{1}{3} \\ \Pr(B|A_2) &= \frac{2}{3}\end{aligned}\tag{24}$$

这两个条件概率之和为 1，即构成了完整的“蛋黄”。

这样，我们可以验证

$$\begin{aligned}\Pr(B) &= \Pr(B|A_1)\Pr(A_1) + \Pr(B|A_2)\Pr(A_2) \\ &= \frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{2} \\ &= \frac{1}{2}\end{aligned}\tag{25}$$

如果读者暂时仍然对全概率定理和贝叶斯定理的关系感到抽象，不必担心。本书后续会通过大量实例，从不同角度反复讲解这两个定理的含义，让它们成为理解现代统计方法和机器学习算法的基石。

独立

我们借助条件概率 $\Pr(A|B)$ 来描述这样一种场景：当我们已经知道事件 B 发生时，事件 A 发生的可能性会如何变化。条件概率反映的正是“在额外信息出现之后，我们对事件发生概率的更新”。

然而，在许多情况下，事件 A 的发生与否并不会受到事件 B 的影响。换句话说，即使我们事先知道 B 是否发生，我们对 A 发生概率的判断也不会改变。如果在这种情况下条件概率满足

$$\underbrace{\Pr(A|B)}_{\text{Conditional}} = \underbrace{\Pr(A)}_{\text{Marginal}}\tag{26}$$

那么我们就称事件 A 和事件 B 是独立的。独立性意味着，从 B 身上获得的信息对于预测 A 完全没有帮助；反之也一样，事件 A 的发生状态对事件 B 也没有任何影响，即

$$\underbrace{\Pr(B|A)}_{\text{Conditional}} = \underbrace{\Pr(B)}_{\text{Marginal}}\tag{27}$$

当事件 A 和 B 独立时，我们可以把上式与条件概率的定义结合起来，得到一个更常用、也更直观的表达式：

$$\underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(A)}_{\text{Marginal}} \cdot \underbrace{\Pr(B)}_{\text{Marginal}}\tag{28}$$

以“抛一枚硬币”和“抛一颗骰子”为例：抛硬币的结果（无论是正面还是反面）和抛骰子的结果（无论是 1 点还是 6 点）是彼此独立的。

我们抛硬币得到“正面”，既不会让骰子更倾向于抛出某个数字，也不会让某个数字变得不可能出现。反之亦然。因此，计算这两个独立事件同时发生的概率时，只需要将它们各自发生的概率相乘即可，这也正是“相互独立”最直观的体现——它们各走各的道，互不干扰。

⚠ 注意，“独立”不同于“条件独立”。

⚠ 注意，“独立”不同于“互斥”。如果两个事件 A 和 B 独立，意味着一个事件的发生不影响另一个事件的发生概率。比如，掷两枚骰子，第一枚的结果不会影响第二枚的结果。如果两个事件 A 和 B 互斥，表示它们不可能同时发生。比如，掷一枚骰子，点数不可能同时为奇数 (A) 和偶数 (B)。



请大家用 DeepSeek/ChatGPT 等工具完成本节如下习题。

Q1. 掷一颗均匀色子，定义事件 A 为“点数为偶数”，事件 B 为“点数为 6”。请写出样本空间 Ω ，事件 A 、 B ，以及事件 B 在事件 A 条件下的新样本空间 Ω_A 。

Q2. 某班学生男女比例为 3:2，男生中身高超过 180cm 的概率为 0.4，女生中身高超过 180cm 的概率为 0.05。随机选一个学生，已知其身高超过 180cm，计算他是男生的条件概率。

Q3. 掷两颗均匀色子，事件 A 为“总点数大于 8”，事件 B 为“第一颗色子为 6”。计算 $\Pr(A|B)$ 并解释结果。

Q4. 某工厂有三条生产线，分别生产产品的合格率为 0.9、0.8、0.85，三条生产线产量占比分别为 0.3、0.5、0.2。随机抽取一个产品，计算它合格的概率。

Q5. 用 Python 模拟掷一颗色子 1000 次，统计事件 A (点数为偶数) 条件下事件 B (点数为 6) 发生的概率，并与理论值比较。