第0章 CHAPTER 0 [绪论]

人工智能在过去的60多年里取得了令人瞩目的进展，也在社会的各个领域得到了广泛应用，深刻改变了众多行业的运行方式，并被许多国家列为未来重点发展的科技领域之一。为了向读者们系统地介绍人工智能的核心原理与算法，本书精选了人工智能的9个核心方向，包括搜索、机器学习、线性回归、决策树、集成学习、神经网络、计算机视觉、自然语言处理以及强化学习，并对这些方向的基础原理与具体算法进行详细讲解。

第2章介绍人工智能中基本的搜索问题以及四类基础算法，包括：盲目搜索，即不利用问题定义本身之外的知识，而是根据事先确定好的规则依次调用动作，以探求到达目标的路径；启发式搜索，利用问题定义之外的知识引导搜索，主要通过访问启发函数来估计每个节点到目标点的代价或损耗；局部搜索；多智能体对博弈中常出现的对抗搜索。

第3章介绍监督学习的框架，包括如何定义训练数据集与测试数据集，寻找合适的模型，定义合适的损失函数；保证模型能同时在训练数据集与测试数据集上表现优秀，即拥有出色的泛化能力。还介绍了创建数据集的基本思路以及无监督学习框架下的 K 平均算法以及谱聚类算法。

第4章介绍了线性回归，并介绍了使用(随机)梯度下降法对目标函数进行优化。在线性回归的基础上，介绍了使用Sigmoid函数或者Softmax函数，输出针对二分类或者多分类的概率分布。还介绍了分类问题中常用的损失函数交叉熵，并学习了正则化的方法，包括岭回归与套索回归。最后，介绍了支持向量机及其与高维核空间的配合使用。

第5章与第6章介绍了在机器学习中应用广泛的决策树模型及集成学习方法，包括决策树的定义、构建及预防过拟合的方法，以及如何通过结合多个简单的决策树模型得到比单个模型更优的继承学习算法。

第7章介绍神经网络，包括常用的激活函数，如ReLU函数与Sigmoid函数等，以及用于神经网络优化中计算导数的反向传播算法。此外，讨论了一些常见的优化神经网络的方法，包括初始化、权值衰减等。最后，探讨了神经网络权值共享的最常见结构——卷积神经网络和循环神经网络。

第8章主要介绍四个知识点。首先是图像的形成，包括小孔相机成像原理与数字图像原理。接下来，介绍了线性滤波器，包括其定义、常见的线性滤波器及其用途。然后介绍了图像边缘的含义、形成原因及检测边缘的方法。最后，介绍了卷积神经网络，包括神经网络卷积层的定义以及卷积神经网络的设计。

第9章介绍了多种语言模型的建模和计算方法。首先从最基础的 n-gram 开始，给出

了语义计算，以及基于循环神经网络的语言模型。接着围绕 Seq2Seq 模型引入了注意力机制，并进一步介绍了前沿自然语言处理的基本模型 Transformer。最后，简单阐述了基于 Transformer 的预训练方法。

第 10 章介绍了马尔可夫决策过程与强化学习，包括马尔可夫决策过程的定义及几个重要的算法——值迭代、策略迭代、线性规划及强化学习算法 Q-learning。最后，简单介绍了深度强化学习的原理及几个常见的核心算法。

随着人工智能的快速发展，其包含的细分领域也越来越多。同时许多学科在研究当中也引入了人工智能的思想与方法，因此产生了许多人工智能的前沿应用（AI＋X），如人工智能与生物医药、人工智能与交通、人工智能与通信和计算等。本书并没有尝试覆盖所有子领域与应用，不过书中章节精选的人工智能技术在很多人工智能的子领域及前沿应用中得到了广泛的应用。希望读者在学习书中人工智能领域基础知识的同时，能同时学习其他子领域的关键技术，并进一步加深对其他学科方向的理解。

第1章 CHAPTER 1 [数学基础]

微积分、概率论和矩阵运算是学习人工智能必要的数学基础。本章旨在为读者介绍本书中涉及的数学知识，使有初等数学基础的读者无需进行额外的阅读与学习便可直接阅读学习后续的章节。为便于读者理解，本章的正文部分主要介绍必要的定义与基础，对知识点的扩展与解释将放在附录中。

1.1 导数

导数，也称微商，在自然科学、计算机科学、工程学科等诸多领域均有着广泛的应用。导数研究的是函数在某一点附近的局部性质，用以刻画曲线或曲面的弯曲程度。在本节中，将会介绍导数的基本概念、计算方法及一些简单应用，以便后续章节使用。

1.1.1 导数的定义

在日常生活及科学研究中，我们经常会遇到需要表示某种量变化快慢的问题。例如，汽车行进过程中位置随时间变化的快慢；吹气球时，气球的半径随吹入气体的量变化的快慢；登山过程中，山的高度随水平位置的变化快慢（即陡峭程度）等。那么，我们应该如何描述这些变化的快慢呢？

可以看出，上述问题均涉及两个量的变化关系：一个是我们关心的正在变化的量（汽车的位置、气球的半径、山的高度等），称作因变量，通常记为 y；另一个是引起这个变化的原因（汽车行驶的时间、吹入气体的量、相对于山的位置），称为自变量，记为 x。这两个量之间存在函数关系 f，写作 $y=f(x)$。

如图 1.1 所示，两个不同的自变量 x_1 和 x_2，分别对应不同的因变量 y_1 和 y_2。不难想到，要表示 $y=f(x)$ 在 x_1 和 x_2 之间变化的快慢，只需将因变量的变化量（$\Delta y=y_2-y_1=f(x_2)-f(x_1)$）除以自变量的变化量（$\Delta x=x_2-x_1$），即

$$\frac{\Delta y}{\Delta x}=\frac{y_2-y_1}{x_2-x_1}=\frac{f(x_2)-f(x_1)}{x_2-x_1}$$

上式称为函数 $y=f(x)$ 从 x_1 到 x_2 的平均变化率，也可写成如下形式：

$$\frac{f(x_1+\Delta x)-f(x_1)}{\Delta x}$$

从图 1.1 中不难看出，平均变化率在函数图像中的意义为 x_1 和 x_2 对应的点 P_1 和 P_2 之间连线（函数的割线）的斜率。

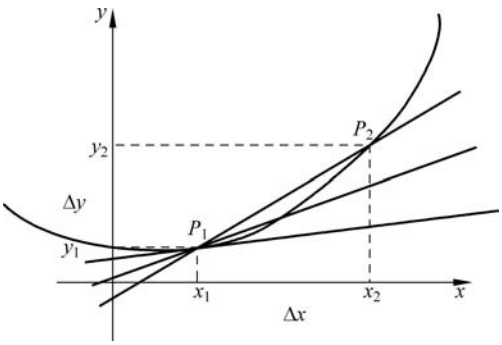


图 1.1　函数斜率的逼近

例 1.1　平均变化率计算

对于一次线性函数 $y=f(x)=ax+b$，它在 $x=x_0$ 的平均变化率为

$$\frac{\Delta y}{\Delta x}=\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}=a$$

与 x_0 无关，因为从图 1.1 可以看出，线性函数的平均变化率即它的斜率。

上文中定义的平均变化率是在 Δx 内对应函数 $y=f(x)$ 的变化速率。如果令 Δx 越取越小，不断逼近 0，那么 x_1 和 x_2 也会越来越近，直至几乎变为同一点。在这种情况下，原先定义的平均变化率也渐渐变为了 $y=f(x)$ 在 (x_1,y_1) 这一点处瞬间所具有的变化速度，称为瞬间变化率，记为

$$\lim_{\Delta x\to 0}\frac{f(x_1+\Delta x)-f(x_1)}{\Delta x}$$

其中，极限符号 $\lim_{x\to c}f(x)$ 表示 x 趋向于 c 时函数 $f(x)$ 的值。一般情况下，在 Δx 趋近于 0 时，这个式子的值趋向于一个定值，即为函数在这一点处的导数。

例如，在车辆行驶的过程中，速度仪表盘上的读数就代表了行驶距离关于时间的函数 $s(t)$ 在这一时刻的导数，即我们常说的（瞬时）速率。

例 1.2　瞬时变化率计算

考虑一个二次函数 $y=f(x)=x^2$ 在 $x=1$ 附近的变化率，有

$$\frac{f(1+\Delta x)-f(1)}{\Delta x}=\frac{(1+\Delta x)^2-1^2}{\Delta x}=2+\Delta x$$

而当 Δx 趋近于 0 时，不难看出，这个式子的值就趋于一个定值 2，即

$$\lim_{\Delta x\to 0}\frac{f(1+\Delta x)-f(1)}{\Delta x}=2$$

基于上述介绍，可以将导数按如下方式定义：

定义［导数］：假设函数 $y=f(x)$ 在某区间上的导数存在，则在此区间上某点 $(x_1,f(x_1))$ 处的导数定义为

$$f'(x_1)=\lim_{\Delta x\to 0}\frac{f(x_1+\Delta x)-f(x_1)}{\Delta x}$$

此区间上所有点的导数构成以 x 为自变量的函数，称为导函数（有时也简称为导数），记为 $f'(x)\left(\text{或 } y', \frac{\mathrm{d}y}{\mathrm{d}x}, \frac{\mathrm{d}f}{\mathrm{d}x}\right)$。寻找已知的函数在某点的导数或其导函数的过程称为求导。

从图 1.1 中不难看出，当 Δx 趋近于 0 时，两点无限接近，原本的割线变为了函数图像的切线，因此导数的几何意义为函数 $y=f(x)$ 的图像在点 $(x_1, f(x_1))$ 处的切线斜率。下面是导数计算的一些基本性质：

两函数和差：$(u\pm v)'=u'\pm v'$

两函数积：$(uv)'=u'v+uv'$

两函数商：$\left(\frac{u}{v}\right)'=\frac{u'v-uv'}{v^2}$

复合函数：$\{f[\varphi(x)]\}'=f'[\varphi(x)]\varphi'(x)$ 或 $\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\mathrm{d}y}{\mathrm{d}u}\cdot\frac{\mathrm{d}u}{\mathrm{d}x}$（链式法则）

1.1.2 高阶导数与偏导数

导数 $f'(x)$ 本身也可以视作自变量 x 的函数，因此可以对函数的导函数再次求导，得到高阶导数。例如，一阶导数 $y'=f'(x)$ 的导数为 $y=f(x)$ 的二阶导数，记作 y''，或 $f''(x), \frac{\mathrm{d}^2y}{\mathrm{d}x^2}$。

举个例子，物体的位移对时间进行求导可以得到速度，速度是位移的一阶导数。而速度可以对时间再求一次导数，得到加速度，这是一个用来衡量物体运动速度的变化的物理量，那么加速度就是速度的一阶导数。同时，加速度也可以看作对位移求导之后再进行求导得到的，所以加速度也是位移的二阶导数。同样地，可以再对加速度进行求导，得到所谓的加加速度。依此类推，可以不断地求导，从而得到一个函数的高阶导数，$y=f(x)$ 的 n 阶导数记为 $y^{(n)}$，或者 $f^{(n)}(x), \frac{\mathrm{d}^ny}{\mathrm{d}x^n}$。

设函数 $y=f(x)$ 在 x_0 处的值为 $f(x_0)$，则当 x_0 增加一个小量 Δx 时（即 $\Delta x\ll 1$），$f(x_0+\Delta x)$ 与 $f(x_0)$ 的关系近似表达为

$$f(x_0+\Delta x)\approx f(x_0)+f'(x_0)\Delta x+\frac{f''(x_0)}{2}\Delta x^2$$

特别地，当 $x_0=0$ 时，有

$$f(\Delta x)\approx f(0)+f'(0)\Delta x+\frac{f''(0)}{2}\Delta x^2$$

更一般地，有

$$f(x_0+\Delta x)=f(x_0)+f'(x_0)\Delta x+\frac{f''(x_0)}{2}\Delta x^2+\frac{f'''(x_0)}{6}\Delta x^3+\cdots$$

这个展开式叫做函数 $y=f(x)$ 在 x_0 处的泰勒展开。下面是一些常用的泰勒展开（在 $x=0$ 处），大多数情况下只需展开到第一阶。

(1) $(1+x)^n\approx 1+nx+\frac{n(n-1)}{2}x^2$

(2) $e^x \approx 1+x+\frac{x^2}{2}$

(3) $\sin x \approx x-\frac{x^3}{6}, \cos x \approx 1-\frac{x^2}{2}$

(4) $\ln(1+x) \approx x-\frac{x^2}{2}, \ln(1-x) \approx -x-\frac{x^2}{2}$

上述讨论的函数 $y=f(x)$均为只有一个自变量 x 的一元函数。此时该函数对于自变量 x 的变化率即为它的导数。而对于自变量多于一个的多元函数 $y=f(x_1,x_2,\cdots)$,研究它的变化率同样是一个有意义的问题,例如,植物的生长与所处环境的温度、湿度、光照强度均有关系,我们该如何刻画生长速度与不同因素的关联呢?下面我们将引入偏导数。

在数学中,一个多变量函数关于某一变量的偏导数,是在保持其他变量不变的情况下函数关于该变量的导数。具体来说,函数 $z(x,y)$关于变量 x 的偏导数写作 z'_x 或$\partial z/\partial x$。此时我们将变量 y 视为常数,而只对变量 x 进行求导,如函数 $z=x^2+3xy+2y^2$ 关于变量 x 和 y 的偏导数分别为

$$\frac{\partial z}{\partial x}=2x+3y$$

$$\frac{\partial z}{\partial y}=3x+4y$$

偏导数的作用与价值在向量分析和微分几何以及机器学习领域中受到广泛认可。

1.1.3 导数与函数极值

若一个函数 $y=f(x)$在 x_0 处的导数 $f'(x_0)=0$,则函数图像在该处的切线平行于 x 轴。不难看出,很多时候该点处的函数值是附近的一个小区间中最大或最小的函数值(存在例外情况,如图 1.2(c)所示),此时称 $y=f(x)$在 x_0 处有极大值或者极小值。

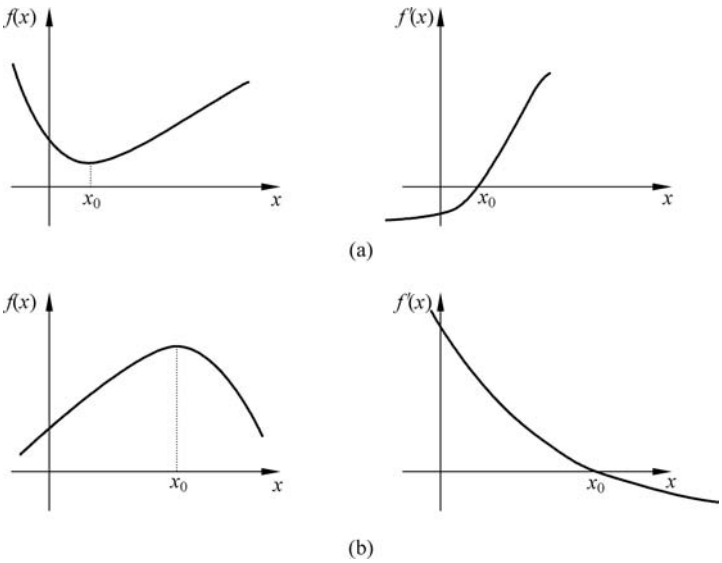
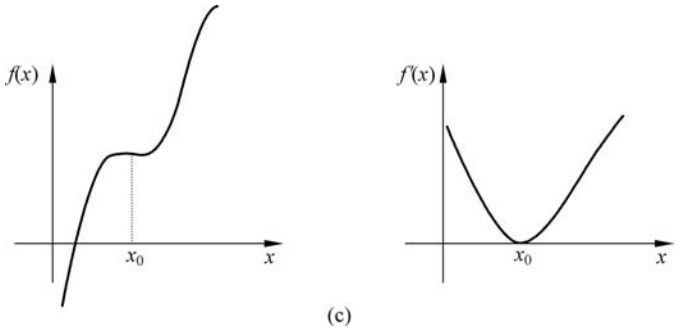


图 1.2 不同取值的导数

(a) $f''(x_0)>0$; (b) $f''(x_0)<0$; (c) $f''(x_0)=0$(例如,$y=ax^3$)



(c)

图 1.2(续)

判断函数是否取到极值可通过导数是否为零来判断，而判断是极大值还是极小值可通过二阶导数的正负来判断(如图 1.2 所示)：若 $f''(x_0)>0$，则为极小值；若 $f''(x_0)<0$，则为极大值；若 $f''(x_0)=0$，则二者皆可能，也可能二者皆非，具体需要分析更高阶的导数。

1.2 概率论基础

概率论是研究随机现象的数量规律的数学分支，是统计学、统计推断和统计机器学习的基础。在本节中，将简单地介绍概率论中的一些基本概念，并提及一些数学分析中的基本知识。

1.2.1 事件与概率

概率论研究的基本对象是随机的、偶然的自然现象或社会现象，它与必然现象是相对的。然而，随机现象本身也有规律可循。如著名的 Galton 钉板实验，如图 1.3 所示。在这个实验中，每次在顶端放下一个小球。假设小球质量均匀，钉子光滑，则小球从上端落下碰到钉子后往哪边下落具有很强的随机性。进行一次实验后，小球会落到下方某一个球槽中，那么进行 n 次实验后，不同球槽中出现的小球数量占落下小球总数的比重就会形成一个分布，我们将该分布称为频率。当实验的次数 n 趋近于无穷大时，频率的值便会趋于稳定。我们称该极限为频率的稳定值。

我们相信，每次实验中小球落在每一个球槽中的可能性大小是客观存在的，因此可以对其进行量度。我们将存在客观可能性的实验过程叫做随机实验(stochastic experiment)，可能性的量度指标称为概率。

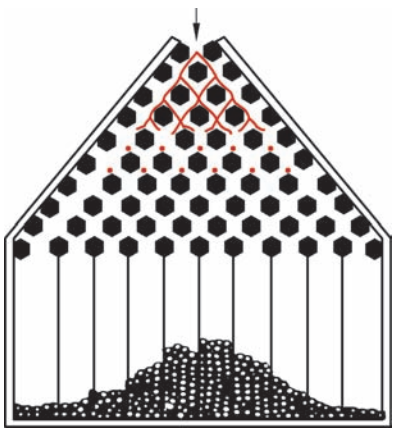


图 1.3 Galton 钉板实验

为了更准确地对随机实验和概率进行描述，我们引入(随机)事件的概念。所谓事件，可以粗略地理解为随机实验的结果。例如，抛一枚硬币结果朝上就是一个事件，在 Galton 钉板实验中，小球落在第 i 个球槽中也是一个事件。随机实验里最基本的不能再分解的结果叫做基本事件。所有基本事件构成的合集记为

Ω。一个事件就是一些基本事件的合集，即 Ω 的一个子集。根据集合的性质我们可以定义出所有可能事件构成的事件体$\mathcal{F}$和事件的运算(并、交、逆)。对于同一事件体中的事件 A 与 B，定义 $A\cup B$ 或者 $A+B$ 为 A 与 B 的和事件，而 $A\cap B$ 或者 AB 为积事件。定义 ϕ 为不可能事件，Ω 为必然事件。那么如果 $AB=\phi$，我们说 A 事件与 B 事件互斥，或不相容。严格的事件定义需要依赖于集合论(σ-代数)中的概念，但以上关于事件集合的定义已经能基本满足人工智能学习的要求，因此我们不对其进行更加详细的介绍。

概率的严格定义是事件体$\mathcal{F}$上定义的一个非负的、和为 1 的(规范的)、可列可加的实值(测度)函数，而较为容易理解的模糊定义可以参见高中教科书。记 $P(A)$为事件 A 的概率。可列可加性是测度的基本要求。我们将观测的对象 Ω、事件体$\mathcal{F}$和概率 P 构成的三元体$(\Omega,\mathcal{F},P)$称为概率空间。对于一般事件的概率计算，可以直接利用集合论的结论进行。

在日常生活中，以下两种基本事件等可能的概型最为常见：古典概型和几何概型。在古典概型中，基本事件是个数有限且等可能的概率模型，也是我们之后讨论的重点。我们还能碰到一些情况，基本事件是连续分布的，事件空间 Ω 是一个连续空间。此时我们可以将 Ω 与一个几何区域的面积联系起来，这样的概型称为几何概型。例如，在 Galton 钉板实验中，每一次小球碰钉子后都有向左落下和向右落下两种可能的结果，概率各为 1/2，因而是古典概型。而如果钉板有 n 层，那么总的 Galton 钉板实验不是古典概型(由于其不是等概率)，而是 n 次连续古典概型实验的结果。

在古典概型中，我们记事件空间为 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_N\}$，其中每个 ω_i 为一个基本事件，其概率为 $P(\omega_i)=1/N$，N 为基本事件的总数。而事件体$\mathcal{F}$由 Ω 的所有子集构成。如果 A 事件中包含 n_A 个基本事件，那么 $P(A)=n_A/N$。我们以摸红色和绿色小球为例，说明古典概型的作用。

考虑一个箱子里有 a 个红球，b 个绿球。那么对于摸到的每个球的颜色，其构成一个古典概型。具体来说，假设我们进行 n 次古典概型实验，可以引出以下两种常见的概型：

(1) 二项式概型(独立重复古典概型实验)

如果每次进行实验之后，就放回摸出的小球，那么两轮之间的结果相互独立。摸出 k 次红球的概率不难通过排列组合方法得出：

$$p_k=C_n^k\left(\frac{a}{a+b}\right)^k\left(\frac{b}{a+b}\right)^{n-k}$$

我们称 k 满足的分布为二项分布 $B\left(k;n,\frac{a}{a+b}\right)$。这种有放回的摸小球的概型为二项概型。这里 C_n^k 是组合数，定义为从 n 个元素中取出 k 个元素，k 个元素的组合数量，即 $C_n^k=n!/[k!(n-k)!]$。

(2) 超几何概型

如果每次进行实验之后，不放回摸出的小球，那么后一轮摸小球的结果会有所变化。此时，摸出 k 次红球的概率可以通过古典概型计算得出：

$$p_k=\frac{C_a^kC_b^{n-k}}{C_{a+b}^n}$$

称 k 满足的分布为超几何分布 $H(k;n,a,a+b)$。这种没有放回的摸小球的概型为超几何概型。

1.2.2 随机变量与概率分布

随机变量是可以随机地取不同值的变量，它可以是离散或连续的。在此为了简化讨论，仅考虑离散型随机变量（连续型变量的讨论将在附录中给出，感兴趣的读者可自行查阅）。概率分布描述随机变量在每个可能取到的值的可能性大小，离散型变量的概率分布可以用概率质量函数来描述。概率质量函数可以同时描述多个随机变量，这种多变量的概率分布称为联合概率分布。例如，$P(X=x,Y=y)$表示 $X=x,Y=y$ 同时发生的概率，可简写为 $P(x,y)$。

为了丰富对样本空间 Ω 的描述方法，可以引入实值函数对概率分布进行描述。设 $(\Omega,\mathcal{F},P)$为概率空间，X 为 Ω 上的实值函数，满足对任意的 $x\in R$，

$$P(X\leqslant x):=P(\{\omega:X(\omega)\leqslant x\})$$

其中，$\{\omega:X(\omega)\leqslant x\}\in\mathcal{F}$，那么可以说 X 为空间$(\Omega,\mathcal{F})$上的随机变量，并且称

$$F_X(x):=P(X\leqslant x),\quad x\in R$$

为 X 的分布函数。由以上定义可见，如果随机变量给定，那么分布函数是存在并且唯一的。

概率和分布函数具有以下关系：

$$P(a<X\leqslant b)=F_X(b)-F_X(a),\quad P(X>x)=1-F_X(x)$$

$$P(X<x)=F_X(x-0),\quad P(X=x)=F_X(x)-F_X(x-0)$$

其中，$P(x-0):=\lim\limits_{n\to\infty}\left(x-\dfrac{1}{n}\right)$为 x 的左极限。

接下来，介绍条件概率与条件分布的概念。设 $A,B\in\mathcal{F}$，且 $P(A)>0$，记

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

为已知 A 事件发生的条件下，事件 B 发生的条件概率。而 $P(AB)=P(A)P(B|A)$称为条件概率的乘法公式，可以拓展为以下的一般形式：

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2\mid A_1)\cdots P(A_n\mid A_1A_2\cdots A_{n-1})$$

如果有一组有限多个或者可列无穷个事件$\{A_i,i=1,2,\cdots\}$，满足 $A_i\in\mathcal{F},P(A_i)\geqslant 0,i=1,2,\cdots$且$\bigcup_i A_i=\Omega$，并有$\{A_i\}$两两相斥，我们称其为 Ω 的完备事件群。由概率的可加性和条件概率的乘法公式，可以得到以下的全概率公式：

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

该公式提供了在已知 A 事件情况下 B 事件发生概率的计算方法。

由条件概率定义、乘法公式和全概率公式，可以得到：

$$P(A_i\mid B)=\frac{P(A_iB)}{P(B)}=\frac{P(A_i)P(B\mid A_i)}{\sum\limits_k P(A_k)P(B\mid A_k)}$$

这个公式叫做逆概率公式或贝叶斯公式，它是统计学习和统计推断的基础。以下是一个贝叶斯公式应用的例子。

例 1.3 已知在所有男子与女子中分别有 5%与 0.25%的人患有色盲症。假设男女的比例为 1∶1。现在随机抽查一人发现其患有色盲症，计算其为男子的概率。

可以设变量 A 表示性别（0/1 分别对应男/女），B 表示是否色盲（0/1 为否/是色盲）。

由条件有

$$P(B=1 \mid A=0)=0.05, \quad P(B=1 \mid A=1)=0.0025$$

因为男女比例为 1∶1，在随机抽查的条件下，应有 $P(A=0)=P(A=1)=1/2$。此时，由贝叶斯公式即可得到：

$$\begin{aligned} P(A=0 \mid B=1) &= P(B=1 \mid A=0)\frac{P(A=0)}{P(B=1)} \\ &= \frac{P(B=1 \mid A=0)P(A=0)}{\sum_i P(B=1 \mid A=i)P(A=i)} \approx 95\% \end{aligned}$$

例 1.4 （三门问题）有三扇关闭的门，其中一扇门的后面有辆跑车，而另外两扇门的后面各藏有一只山羊，跑车在哪一扇门的后面是完全随机的。参赛者需要从中选择一扇门，如果参赛者选中后面有车的那扇门就可以赢得这辆跑车。参赛者随机选定了一扇门，但未去开启它的时候，节目主持人会开启剩下两扇门的其中一扇，其门后是一只山羊。此时参赛者是否应该保持他的原来选择，还是应该转而选择剩下的那一道门？

这个问题乍一看似乎没有换门的必要。我们现在用贝叶斯公式来看看结论是否如此。不妨假设参赛者选 1 号门，而主持人打开了 2 号门。记随机变量 $A=i$ 为第 i 扇门后面有汽车，由于随机性，有 $P(A=i)=1/3, i=1,2,3$。

现在再定义随机变量 B 为主持人是否打开 2 号门：如果主持人打开 2 号门，则 $B=1$；否则，$B=0$。这里注意到如果 2 号门的背后有跑车，主持人是不能打开该门的。根据 A 和 B 的定义可得：

$$\begin{aligned} P(B=1 \mid A=1)&=0.5 \\ P(B=1 \mid A=2)&=0 \\ P(B=1 \mid A=3)&=1 \end{aligned}$$

这里第一个式子是因为 $A=1$（参赛者已经选对了），因此主持人可能选 2 或者 3，且选 2 的可能性是 1/2；第二种情况不可能发生，因为主持人不能打开正确的门；而在第三种情况下，主持人只能打开 2 号门，所以 $B=1$ 一定成立。于是有全概率公式

$$P(B=1)=\sum_i P(B=1 \mid A=i)P(A=i)=0.5$$

那么，

$$\begin{aligned} P(A=1 \mid B=1) &= P(B=1 \mid A=1)\frac{P(A=1)}{P(B=1)}=\frac{1}{3} \\ P(A=2 \mid B=1) &= P(B=1 \mid A=2)\frac{P(A=2)}{P(B=1)}=0 \\ P(A=3 \mid B=1) &= P(B=1 \mid A=3)\frac{P(A=3)}{P(B=1)}=\frac{2}{3} \end{aligned}$$

所以参赛者应该改变想法选 3 号门。

接下来，我们介绍事件的独立性与条件变量的独立性。对于事件 A 与事件 B，若 $P(AB)=P(A)P(B)$，则称它们相互独立。拓展到多个事件的情况，称事件 $A_i, i=1,2,\cdots,n$ 相互独立。如果对其中任意 k 个事件，均满足