第 1 章　概率论基础

本章简要介绍概率论中与贝叶斯网密切相关的一些基本概念. 与一般的概率论教材不同，我们将侧重于概念的直观含意. 深入理解诸如条件独立、贝叶斯定理等概念和结果，对以后理解贝叶斯网至关重要. 另外，本章还将简述概率方法在人工智能研究中的崛起过程，并介绍在第三部分中会被用到的一些信息论知识.

1.1　随机事件与随机变量

世界上许多事情都具有不确定性. 例如掷硬币，其结果可能正面朝上，也可能反面朝上，在抛掷之前无法预知. 又如赌马，理论上每匹马都有跑第一的可能，事先无法预料哪匹马一定会赢. 再如火星上是否曾有生命存在？答案有两种可能，是或不是，但根据目前掌握的证据判断，无法给出绝对的答复. 概率论是研究处理这类现象的数学理论. 本节介绍概率论中几个最基本的概念，包括样本空间、事件、概率测度、随机变量以及概率函数.

1. 样本空间和事件

在概率论中，**随机试验**指的是事先不能完全预知其结果的试验. 随机试验的所有可能结果组成该试验的**样本空间**，通常记为 Ω. 样本空间可以是离散的，也可以是连续的. 如无特殊说明，本书所论及的样本空间都将是离散的.

样本空间中的点，即随机试验的可能结果，称为**样本点**，或**原子事件**，记为 ω. 样本空间的子集称为**事件**，通常用大写字母表示：A，B，…. 如果随机试验的结果包含在一个事件之中，则称该事件发生了. 样本空间 Ω 本身也是一个事件，而且是一定会发生的**必然事件**. 空集也是一个事件，是**不可能事件**，通常记为 $\varnothing$. 事件之间可以进行交（$\cap$）、并（$\cup$）、差（$\setminus$）等各种集合运算. 若两事件 A 和 B 交空，即 $A \cap B = \varnothing$，则称它们为**互斥事件**，又称**不相容事件**. 两互斥事件不能同时发生. 若 A 和 B 互斥，且 $A \cup B = \Omega$，则称它们为**互补事件**.

例 1.1　考虑掷硬币试验，其结果有正反面两种可能，因此样本空间为 $\Omega = \{h, t\}$，其中 h 表示正面，t 表示反面，h 和 t 为两个互补的原子事件. □

例 1.2　考虑掷骰子试验，有 6 种可能的结果，样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$. 子集合 $\{1, 3, 5\}$ 表示的是“掷出结果为奇数”这一事件，其互补事件为 $\{2, 4, 6\}$，即“掷出结果为偶数”. □

例 1.3　考虑从香港科技大学的所有研究生中随机抽样的试验，其结果可能

是任一研究生，样本空间为 $\Omega=\{x \mid x$ 是科大研究生$\}$.〈研究生张三〉为一原子事件①.〈所有科大男研究生〉或〈所有年龄超过 25 岁的研究生〉是两个非原子事件. □

2. 概率测度

概率测度给样本空间中的每一事件 A 赋予一个数值 $P(A)\in[0,1]$，以度量该事件发生的可能性. 在数学上，它是一个从样本空间 Ω 的幂集 2^{Ω} 到区间 $[0,1]$ 的映射 $P:2^{\Omega}\to[0,1]$，且满足以下 3 个 **Kolmogorov 公理**（Kolmogorov，1933，1950）：

(1) $P(\Omega)=1$；

(2) $P(A)\geqslant 0,\ \forall A\in 2^{\Omega}$；

(3) $P(A\cup B)=P(A)+P(B),\forall A,B\in 2^{\Omega},A\cap B=\varnothing$.

$P(A)$ 称为事件 A 的**概率**. 上面这 3 个公理分别被称为概率测度的**规范性**、**非负性**和**有限可加性**. 规范性规定必然事件的概率为 1，非负性规定概率不能为负，有限可加性规定互斥事件的并集的概率等于它们各自概率的和. 从这 3 个公理出发，可以推出概率测度的诸多基本性质和定理.

例 1.4 例 1.1 中的样本空间的所有子集是：$\Omega,\varnothing,\{h\},\{t\}$. 设所掷为均匀硬币，则有 $P(\Omega)=1,P(\varnothing)=0,P(h)=P(t)=0.5$. 这里 $P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

例 1.5 在例 1.1 中，若所掷为不均匀硬币，且已知在 1000 次重复试验中正面出现 700 次，反面出现 300 次，可因此设 $P(\Omega)=1,P(\varnothing)=0,P(h)=0.7,P(t)=0.3$. 这里 $P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

例 1.6 在例 1.3 中，对 Ω 的任一子集 A，定义 $P(A)=\dfrac{|A|}{|\Omega|}$，其中 $|A|$ 是 A 中元素的个数，称为 A 的势. 那么，$P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

3. 随机变量

随机变量是定义在样本空间 Ω 上的函数，通常用大写字母表示，如 X,Y,Z. 随机变量的取值随试验的结果而定，通常用小写字母表示，如 x,y,z. 随机变量 X 的所有可能取值的集合称为它的**值域**，也称**状态空间**，记为 Ω_X. 随机变量可以是离散的，也可以是连续的. **离散随机变量**的状态空间是离散的，包含有限个或无穷可数个状态. **连续随机变量**的状态空间是连续的，包含无穷不可数个状态. 本书主要考虑离散随机变量.

① 假设只有一个叫张三的研究生.

例 1.7 在例 1.3 中，设 X 为“随机抽出的一个学生的性别”，则 X 是定义在“科大研究生”这个样本空间上的随机变量：$\Omega_X=\{m,f\}$，其中 m 表示男，f 表示女. □

例 1.8 同时掷两个质地均匀的硬币，其样本空间为 $\Omega=\{(h,h),(h,t),(t,h),(t,t)\}$. 对任一 $\omega\in\Omega,P(\{\omega\})=1/4$. 设 X 为“正面朝上的硬币个数”，那么 X 是定义在 Ω 上的一个随机变量：$X((h,h))=2,X((h,t))=1,X((t,h))=1,X((t,t))=0$，故 $\Omega_X=\{0,1,2\}$. □

4. 概率函数

设 X 为一随机变量，x 是它的一个取值. 在样本空间中，所有使 X 取值为 x 的原子事件组成一个事件，记之为 $\Omega_{X=x}=\{\omega\in\Omega \mid X(\omega)=x\}$，有时也简记为“$X=x$”. 注意，$\Omega_{X=x}$ 与 Ω_X 的含意完全不同，后者是随机变量 X 的状态空间，包括 X 的所有可能取值.

事件“$X=x$”的概率 $P(X=x)=P(\Omega_{X=x})$ 依赖于 X 的取值 x. 让 x 在 Ω_X 上变动，$P(X=x)$ 就成为 Ω_X 的一个取值于 $[0,1]$ 的函数，称之为随机变量 X 的**概率质量函数**，记为 $P(X)$. 根据概率测度的定义，有

$$P(X=x)\geqslant 0,\forall x\in\Omega_X;\ \sum_{x\in\Omega_X}P(X=x)=1.$$

为了记号上的方便，上面两式有时简记为

$$P(X)\geqslant 0;\quad \sum_{X}P(X)=1.$$

例 1.9 在例 1.7 中，设科大共有 500 名研究生，其中 400 名男生，100 名女生. 这里 $\Omega_X=\{m,f\},\Omega_{X=f}=$〈科大所有女研究生〉. $X=f$ 的概率为

$$P(X=f)=P(\Omega_{X=f})=P(\{\text{科大所有女研究生}\})=100/500=0.2.$$

X 的概率函数为

X	m	f
$P(X)$	0.8	0.2

□

例 1.10 在例 1.8 中，变量 X 的值域是 $\Omega_X=\{0,1,2\},\Omega_{X=0}=\{(t,t)\},\Omega_{X=1}=\{(t,h),(h,t)\},\Omega_{X=2}=\{(h,h)\}$. X 的概率函数为

X	0	1	2
$P(X)$	0.25	0.5	0.25

□

离散随机变量有概率质量函数. 与之对应，连续随机变量有概率密度函数. 由于本书主要关心离散随机变量，所以对概率密度函数不做详细介绍. 以后，我们有时会用“概率分布”一词来泛指概率质量函数、概率密度函数，或与它们等价的其它概念.

1.2 概率的解释

概率的解释主要有 5 种：古典解释、频率解释、主观解释、特性解释（Popper，1957）以及逻辑解释（Carnap，1950）. 本节主要介绍前 3 种解释，重点放在概率的主观解释上.

1.2.1 古典解释

概率的古典解释起源于 16 世纪数学家们对掷骰子等赌博活动的研究. 对一粒质地均匀的正方体骰子，投掷后其任何一面朝上的可能性相等，因此掷出每面的概率都应为 1/6. 一般地讲，如果事件 A 包含的样本数为 m，而样本空间的总样本数为 n，则事件 A 的概率应为

$$P(A)=\frac{\text{事件 } A \text{ 包含的样本数}}{\text{样本空间的总样本数}}=\frac{m}{n}. \tag{1.1}$$

用这种方法定义的概率称为**古典概率**. 古典概率的一个前提条件就是等可能性. 在实际应用中，这个前提一般很难满足，因此古典概率的应用范围很有限.

例 1.11 考虑如下掷 3 粒骰子的赌局：若结果之和为 {3，4，5，6，7，14，15，16，17，18} 中任一数字，则赌客胜，否则庄家胜. 3 粒骰子之和有从 3 ～ 18 共 16 种可能结果，其中赌客胜的有 10 种，所以表面看来似乎这是一个对赌客有利的赌局. 但是，由于结果不具有等可能性，因此不能简单认为赌客获胜的概率是 10/16. 实际上，这是一个对庄家有利的赌局，投掷 3 粒骰子有 216 种等可能结果：{(1，1，1)，(1，1，2)，…，(6，1，1)，…，(6，6，6)}，其中使赌客获胜的结果只有 69 种，而使庄家获胜的结果有 147 种. □

1.2.2 频率解释

给定一个质地不均匀的骰子，掷出 6 的概率为多大？古典解释无法处理这种情况，因为这时等可能性前提不成立. 为近似计算这一概率，人们通常进行多次重复试验，记下其中掷出 6 的次数，除之以总的试验次数，将结果作为掷出 6 的概率. 一般地讲，对于一个可在同样条件下重复进行的试验，如果事件 A 在所有 N 次试验中共发生了 M 次，则它的概率可以用其发生的频率来近似：$P(A)\approx M/N$. 这个近似的理论支持是**大数定律**：当 N 趋于无穷大时，频率几乎处处趋于概率. 即当 N 较大时，频率经常稳定地出现在概率附近，而当 N 越大时，越是更经常地稳定于概率，而且幅度也越小. 这就是概率的频率解释.

按照频率解释，概率只有当试验可以在同等条件下无限次重复时才有意义. 然而，实际中人们往往需要研究一些不可重复的事件发生的概率，例如总统竞选或体育比赛的结果. 频率解释对这些一次性事件无法处理. 在早期的人工智能研究中，概率的频率解释曾占据主导地位，这一度为概率论的应用造成了概念上的困难.

1.2.3 主观解释

主观解释又称贝叶斯解释，它认为概率即合理**信度**，反映的是个体的知识状态和主观信念. 在这种意义下的概率称为**主观概率**.

1. 主观概率的评估

相对于频率解释，主观解释的长处是它允许对一次性事件也进行概率评估. 例如：巴西队赢得下届世界杯足球赛冠军的概率是多大？频率解释认为此问题无意义，因为“下届世界杯决赛巴西队夺冠”不是一个可以重复的事件. 但是，主观解释仍可以根据各种先验知识给出一个主观概率评估. 主观概率的评估有许多方法，其中之一是下面的概率轮方法.

例 1.12（概率轮与概率评估） 设想有一质地均匀的概率轮（图 1.1），其上仅包含黑白两个连续区域，转动后指针停在任一位置的概率相等，因而其停在黑区的概率应等于黑区角度所占的百分比. 概率轮提供了一个进行主观概率评估的客观参照. 当评估“巴西队赢得下届世界杯足球赛冠军的概率”时，首先问如下问题：巴西队夺冠的可能性大，还是指针停在黑区的可能性大？如果认为巴西队夺冠更有可能，那么就设想一个更大的黑区，反之设想一个较小的黑区，再次问同样的问题. 如此反复，直到认为巴西队夺冠和指针停在黑区具有相同的可能性，然后测量黑区角度的大小，除以 360，即得到巴西队夺冠的概率. □

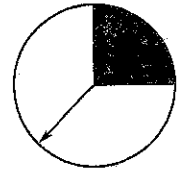


图 1.1 用于主观概率评估的概率轮

作为一种操作性强的手段，用概率轮进行主观概率评估在管理科学、心理学以及运筹学中被广泛使用，详见有关文献（Spetzler and von Holstein，1975）. 不难看出，这里存在一个评估精度的问题：概率值为 0.201 或 0.202，有什么根本的差别吗？两者差别如此之小，以至于人们往往很难断定自己的真实信度究竟是哪一个. 所幸的是，在贝叶斯网应用中，这往往不是一个大问题. 原因有三：首先，概率值的微小差别对决策的影响一般不大；其次，实际中往往会同时考虑多个事件的

概率，由于概率必须满足 Kolmogorov 公理，因此不同事件的概率之间存在一定的关系，而这些关系限制了主观概率的任意性；第三，在数据分析中，当数据量足够大时，主观概率的影响不大，这一点将在本小节最后详细论述.

2. 主观概率与 Kolmogorov 公理

为什么主观概率必须满足 Kolmogorov 公理呢？围绕这个问题，人们提出了几种理论和论证方法（Shafer and Pearl，1990）. 其中之一与赌博有关，它的基本思想是：如果一个人的主观概率不满足 Kolmogorov 公理，那么就可以构造一个赌局，使他认为合理而接受，但又将必输无疑，这样的赌局称为 Dutch book[①]. 对于理性个体，不应该有 Dutch book 存在，因此理性个体的主观概率必须满足 Kolmogorov 公理.

如果一个人对某事件 S 出现的主观概率为 P，$0 \leqslant P \leqslant 1$，则他最多愿意付 P 元去购买一个关于 S 的单位筹码，即如果 S 为真则可拿它兑换 1 元赔金的筹码. 同时，他也愿意以 P 元或更高的价格卖掉一个单位筹码. 下面的两个例子分别说明，如果一个人的主观概率不满足 Kolmogorov 第(1)和第(3)公理，那么就有针对他的 Dutch book 存在.

例 1.13 考虑 $\Omega = \{S, \neg S\}$. 假设一赌客的主观概率为 $P(S) = 0.51$，$P(\neg S) = 0.51$，从而总和 $P(\Omega) = P(S) + P(\neg S)$ 大于 1，违反了 Kolmogorov 第 (1) 公理. 因为 $P(S) = 0.51$，所以他愿意以 0.51 元买入一个赌 S 为真的单位筹码 C_S；同时因为 $P(\neg S) = 0.51$，所以他也愿意以 0.51 元买入一个赌 $\neg S$ 为真的单位筹码 $C_{\neg S}$. 构造如下的 Dutch book：让赌客以 1.02 元的价格同时买进 C_S 和 $C_{\neg S}$ 两个筹码. 对这个赌客来讲，此 Dutch book 是公平的. 但是一旦他接受，无论结果是 S 还是 $\neg S$，他都只能拿回 1 元，从而损失 0.02 元. □

例 1.14 考虑一个由 3 匹马 H_1，H_2，H_3 参加的跑马比赛. 假设有 8 种彩票 T_0，T_1，T_2，T_3，T_{12}，T_{13}，T_{23}，T_{123}，它们的赔金如下：

T_1：100 元 如果 H_1 赢	T_{12}：100 元 如果 H_1 或 H_2 赢
T_2：100 元 如果 H_2 赢	T_{13}：100 元 如果 H_1 或 H_3 赢
T_3：100 元 如果 H_3 赢	T_{23}：100 元 如果 H_2 或 H_3 赢
T_0：100 元 如果没有任何一匹马赢	T_{123}：100 元 如果任何一匹马赢

分别以 $P(H_1)$，$P(H_2)$，$P(H_1 \cup H_2)$ 记某人对 H_1 赢、H_2 赢、H_1 或 H_2 赢

① 在英式赌马比赛中，一个“book”是指为某人所接受的一套下注组合.

的主观概率估计，那么对他来说，彩票 T_1、T_2 和 T_{12} 的公平价格分别为 $P(H_1) \times 100$，$P(H_2) \times 100$ 和 $P(H_1 \cup H_2) \times 100$.

假设一赌客的概率评估如下：$P(H_1) = 0.3$，$P(H_2) = 0.4$，$P(H_1 \cup H_2) = 0.5$. 这里 $P(H_1) + P(H_2) \neq P(H_1 \cup H_2)$，违反了 Kolmogorov 第(3)公理. 一个针对此赌客的 Dutch book 如下：假设他手中开始时有彩票 T_{12}，从他手中以 50 元买走 T_{12}，再分别以 30 元和 40 元的价格将 T_1 和 T_2 卖给他. 对他来说，这个交易是公平的，但是他却蒙受了损失. 交易前后赌客手中彩票的价值变化如下：

	交易前	交易后
如果 H_1 赢	100 元	100 + 50 − 30 − 40 = 80 (元)
如果 H_2 赢	100 元	100 + 50 − 30 − 40 = 80 (元)
如果 H_3 赢	0 元	50 − 30 − 40 = − 20 (元)

无论哪匹马赢，该赌客都损失 20 元. □

3. 主观概率与贝叶斯网

贝叶斯网早期主要应用于专家系统. 在专家系统应用中，贝叶斯网的结构和参数是通过咨询专家而获得的，因此需要用类似于概率轮的方法进行概率评估，主观概率占有重要地位.

随着时间的推移，贝叶斯网越来越多地被用于分析数据，也就是要基于数据建立贝叶斯网模型. 这有两种情形：一是已知网络结构，对网络参数进行估计，称为参数学习；二是不知道网络结构，要通过分析数据，同时获得网络结构和网络参数，称为结构学习. 参数学习有两种方法——最大似然估计和贝叶斯估计. 最大似然估计完全基于数据，不需要先验概率. 贝叶斯估计则假定在考虑数据以前，网络参数服从某个先验分布. 这是先验的主观概率，它的影响随着数据量的增大而减小. 结构学习情形假设在考虑数据以前，不同结构的可能性相等，这也是先验的主观概率，它的影响也随着数据量的增大而减小. 所以，当有足够多的数据时，主观概率对数据分析的影响不大.

尽管概率的主观解释在贝叶斯网的实际应用中并不扮演非常重要的角色，但是在概念上，它对贝叶斯网却是至关重要的. 贝叶斯网所依赖的一个核心概念是条件独立，而概率的主观解释为直观理解条件概率和条件独立提供了一个自然的角度. 这一点的论证将在第 1.3.3 和 1.3.4 节中看到.

1.2.4 特性解释与逻辑解释

在特性解释中，均匀硬币“正面朝上”的概率为 1/2 是这个硬币的固有物理属性，与其是否投掷或投掷次数无关. 特性解释没有为概率提供可操作的运算方

法，因此很难应用于实际之中.

逻辑解释则认为概率是对知识状态的总结，是由从证据到假设的逻辑关系所决定的. 一旦相关的知识得到确定，则事件的可能性就已经被客观地确定下来，并应该能够通过逻辑分析来得到. 古典解释可以看做是逻辑解释的一个特例，它从等可能性的前提条件出发来计算概率. 同特性解释一样，逻辑解释的缺点在于它没能为概率提供一个可操作的运算方法.

1.3 多元概率分布

随机现象往往涉及多个随机因素，因而需要用多个随机变量来描述. 本节介绍多元概率的一些基本概念.

1.3.1 联合概率分布

我们知道，对单个随机变量 X，可以用概率函数 $P(X)$ 来描述它的各个状态的概率. 而对于多个随机变量 $X_1, \cdots, X_n$，则可以用**联合概率分布** $P(X_1, \cdots, X_n)$，简称**联合分布**来描述各变量所有可能的状态组合的概率. 它是一个定义在所有变量状态空间的笛卡儿乘积之上的函数：

$$P(X_1, \cdots, X_n): \otimes_{i=1}^{n} \Omega_{X_i} \to [0, 1],$$

其中所有函数值之和为 1，即

$$\sum_{X_1, \cdots, X_n} P(X_1, \cdots, X_n) = 1.$$

联合分布经常被表示为一张表，其中包含了 $\prod_{i=1}^{n} |\Omega_{X_i}|$ 个状态组合及其概率值. 如果所有变量都只取两个状态，则联合分布表共有 2^n 个项，刻画了变量之间的各种关系.

例 1.15 考虑香港市场上所有出租房屋. 从中随机抽取一间，考查其月租(R)和类型(T)这两个随机变量. 月租分为 4 等：{low(低于 2000 元)，medium (2000 ～ 6000 元)，upper medium (6000 ～ 12000 元)，high(高于 12000 元)}. 类型有 3 种：{public(公屋)，private(私家屋)，others(其它)}. 联合分布 $P(R, T)$ 如下：

R \ T	public	private	others
low	0.17	0.01	0.02
medium	0.44	0.03	0.01
upper medium	0.09	0.07	0.01
high	0	0.14	0.01

从表中可知，随机抽到中价公屋的可能性最大，为 44%. □

1.3.2 边缘概率分布

在例 1.15 中，由于有了联合分布 $P(R,T)$，所以可以回答这样的问题：随机抽取一间出租房屋为公屋的概率 $P(T=\text{public})$ 是多少？根据概率的有限可加性

$$\begin{aligned} P(T=\text{public}) =& P(T=\text{public}, R=\text{low}) \\ &+ P(T=\text{public}, R=\text{medium}) \\ &+ P(T=\text{public}, R=\text{upper medium}) \\ &+ P(T=\text{public}, R=\text{high}) \\ =& 0.7. \end{aligned}$$

同样地，可以计算 $P(T=\text{private})$ 和 $P(T=\text{others})$：

$$\begin{aligned} P(T=\text{private}) =& P(T=\text{private}, R=\text{low}) \\ &+ P(T=\text{private}, R=\text{medium}) \\ &+ P(T=\text{private}, R=\text{upper medium}) \\ &+ P(T=\text{private}, R=\text{high}) \\ =& 0.25, \\ P(T=\text{others}) =& P(T=\text{others}, R=\text{low}) \\ &+ P(T=\text{others}, R=\text{medium}) \\ &+ P(T=\text{others}, R=\text{upper medium}) \\ &+ P(T=\text{others}, R=\text{high}) \\ =& 0.05. \end{aligned}$$

为了简化记号，上面三式可分别缩写为

$$P(T=\text{public}) = \sum_R P(T=\text{public}, R),$$

$$P(T=\text{private}) = \sum_R P(T=\text{private}, R),$$

$$P(T=\text{others}) = \sum_R P(T=\text{others}, R).$$

这三个式子还可以进一步合并为下面一式：

$$P(T) = \sum_R P(T, R).$$

相对于联合分布 $P(R, T)$，$P(T)$ 称为**边缘分布**[①]. 下表同时给出了联合分布 $P(R, T)$ 和边缘分布 $P(T)$，$P(R)$：

① 这一术语来源于保险统计行业. 保险统计师通常把观察到的频率数据相加，并把结果写在保险统计报表的边缘，所以叫“边缘概率”.

R \ T	public	private	others	$P(R)$
low	0.17	0.01	0.02	0.20
medium	0.44	0.03	0.01	0.48
upper medium	0.09	0.07	0.01	0.17
high	0	0.14	0.01	0.15
$P(T)$	0.70	0.25	0.05	

记 $\mathbf{X}=\{X_1, \cdots, X_n\}$，$\mathbf{Y}$ 是 $\mathbf{X}$ 的真子集，即 $\mathbf{Y}\subset\mathbf{X}$，$\mathbf{Z}=\mathbf{X}\backslash\mathbf{Y}$. 则相对于 $P(\mathbf{X})$，$\mathbf{Y}$ 的边缘分布 $P(\mathbf{Y})$ 定义为

$$P(\mathbf{Y})=\sum_{\mathbf{Z}} P(X_1, \cdots, X_n). \tag{1.2}$$

从联合分布 $P(\mathbf{X})$ 到边缘分布 $P(\mathbf{Y})$ 的过程称为**边缘化**.

例 1.16 设有 3 个装有黑白两色球的口袋，第 1 个口袋黑白球各半，第 2 个口袋黑白球比例为 4∶1，第 3 个则全是黑球. 设随机变量 X, Y, Z 分别代表从这 3 个口袋随机抽出的球的颜色，其状态空间为 $\Omega_X=\Omega_Y=\Omega_Z=\{w, b\}$，其中 w 表示白，b 表示黑. 联合概率分布 $P(X,Y,Z)$ 如下：

X	Y	Z	$P(X,Y,Z)$
w	w	w	0
w	w	b	0.1
w	b	w	0
w	b	b	0.4
b	w	w	0
b	w	b	0.1
b	b	w	0
b	b	b	0.4

表中给出了 X, Y, Z 的所有 8 个可能的状态组合及其概率. 从表中可知抽球结果为 (w, w, b) 和 (b, b, b) 的概率一样大，都是 0.4；结果为 (w, w, b) 和 (b, w, b) 的概率也一样，都是 0.1；而其余所有满足 $Z=w$ 的状态组合的概率都为零，因为第 3 个袋子里没有白球. 这里边缘分布 $P(X)$ 为 $(0.5, 0.5)$，$P(Y)$ 为 $(0.2, 0.8)$，$P(Z)$ 为 $(0, 1)$. □

1.3.3 条件概率分布

条件概率与条件分布是用来刻画事件之间及变量之间关系的基本工具.

1. 条件概率

设 A, B 为两随机事件且 $P(B)>0$，事件 A 在给定事件 B 发生时的**条件概率**定义为

$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}. \tag{1.3}$$

直观上，$P(A \mid B)$ 是在已知 B 发生时，对 A 发生的信度；而 $P(A)$ 则是在不知道 B 是否发生时，对 A 发生的信度[①]. 由式 (1.3) 可得

$$P(A \cap B)=P(B)P(A \mid B). \tag{1.4}$$

这称为**概率的乘法定律**，其含意非常直观：对 A 和 B 同时发生的信度，等于对 B 发生的信度乘以已知 B 发生时对 A 发生的信度. 当然乘法定律也可以写为

$$P(A \cap B)=P(A)P(B \mid A). \tag{1.5}$$

例 1.17 在例 1.2 的掷骰子试验中，掷出 6 的概率为 1/6. 假定投掷后被告知“掷出的结果是偶数”，问此时对结果为 6 的信度是多少？设掷出 6 为事件 A，掷出结果为偶数为事件 B，则 $P(A)=1/6$，$P(B)=1/2$，$P(A \cap B)=1/6$. 所问的问题即是要计算如下条件概率：

$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}=\frac{1/6}{1/2}=\frac{1}{3}.$$

□

2. 条件分布

设 X 和 Y 是两随机变量，x 和 y 分别是它们的一个取值. 考虑事件 $X=x$ 在给定 $Y=y$ 时的条件概率为

$$P(X=x \mid Y=y)=\frac{P(X=x, Y=y)}{P(Y=y)}. \tag{1.6}$$

在式(1.6)中，固定 y，让 x 在 Ω_X 上变动，则得到一个在 Ω_X 上的函数. 这个函数称为**在给定 $Y=y$ 时变量 X 的条件概率分布**，记为 $P(X \mid Y=y)$. 用 $P(X \mid Y)$ 记 $\{P(X \mid Y=y) \mid y \in \Omega_Y\}$，即在 Y 取不同值时 X 的条件概率分布的集合. $P(X \mid Y)$ 称为**给定 Y 时变量 X 的条件概率分布**. 在式(1.6)中，让 x 和 y 在 Ω_X

① 这里用到的是概率的主观解释.

和 Ω_Y 上变动，则得到一组等式. 与 1.3.2 节类似，把这些等式缩写为

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}. \tag{1.7}$$

式 (1.7) 可视为是 $P(X \mid Y)$ 的直接定义.

更一般地，设 $\mathbf{X} = \{X_1, \cdots, X_n\}$ 和 $\mathbf{Y} = \{Y_1, \cdots, Y_m\}$ 为两个变量集合，$P(\mathbf{X},\mathbf{Y})$ 为 $\mathbf{X} \cup \mathbf{Y}$ 的联合概率分布，$P(\mathbf{Y})$ 为 $\mathbf{Y}$ 的边缘概率分布. 则给定 $\mathbf{Y}$ 时 $\mathbf{X}$ 的条件概率分布定义为

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{X},\mathbf{Y})}{P(\mathbf{Y})}. \tag{1.8}$$

例 1.18 在例 1.15 中问：随机抽取一间私家屋，其租金为 low 的概率多大? 这即是问给定 T =private 时 R =low 的条件概率 $P(R = \text{low} \mid T = \text{private})$. 按定义，有

$$P(R = \text{low} \mid T = \text{private}) = \frac{P(R = \text{low}, T = \text{private})}{P(T = \text{private})} = \frac{0.01}{0.25} = 0.04,$$

给定 T 时，变量 R 的条件分布 $P(R \mid T)$ 如下：

T \ R	low	medium	upper medium	high
public	$\frac{0.17}{0.7}$	$\frac{0.44}{0.7}$	$\frac{0.09}{0.7}$	$\frac{0}{0.7}$
private	$\frac{0.01}{0.25}$	$\frac{0.03}{0.25}$	$\frac{0.07}{0.25}$	$\frac{0.14}{0.25}$
others	$\frac{0.02}{0.05}$	$\frac{0.01}{0.05}$	$\frac{0.01}{0.05}$	$\frac{0.01}{0.05}$

表中第 1 行显示的是在给定 T =public 时，R 的条件概率分布，第 2 行是在给定 T = private 时，R 的条件概率分布，等等. 这里每行的数字之和为 1，即 $\sum_R P(R \mid T) = 1$. 这与例 1.15 中所列的联合分布 $P(R, T)$ 不同，那里表中所有数字之和为 1，即 $\sum_{R,T} P(R, T) = 1$. □

下面的例子涉及本节前面所介绍的 3 个主要概念，即联合分布、边缘分布和条件分布. 其目的是为读者建立这样的直观印象：几个随机变量的联合分布对应的是一张表，其中一些变量的边缘分布以及条件分布对应的也是表.

例 1.19 设有一袋积木，每块积木有 3 个属性：颜色、材料和形状. 设积木的颜色只能是红(r)或蓝(b)两种，材料只能是金属(m)或木头(w)，形状只能是正方体(6)或正四面体(4). 设 C, M, S 为 3 个随机变量，分别代表从袋中随机取出一块积木的颜色、材料和形状，则 $\Omega_C = \{r, b\}$, $\Omega_M = \{m, w\}$, $\Omega_S = \{6, 4\}$. 设联合概率分布 $P(C, M, S)$ 为

C	M	S	$P(C,M,S)$
r	m	6	0.10
r	m	4	0.10
r	w	6	0.25
r	w	4	0.05
b	m	6	0.15
b	m	4	0.10
b	w	6	0.20
b	w	4	0.05

那么，变量 C 和 M 的边缘分布 $P(C, M)$ 为

C	M	$P(C,M)$
r	m	0.20
r	w	0.30
b	m	0.25
b	w	0.25

条件分布 $P(C \mid M)$ 为

M \ C	r	b
m	$\frac{4}{9}$	$\frac{5}{9}$
w	$\frac{6}{11}$	$\frac{5}{11}$

□

3. 链规则

对两变量 X、Y 的联合分布 $P(X,Y)$，按照条件分布的定义，可得

$$P(X,Y) = P(X)P(Y \mid X). \tag{1.9}$$

将其推广到 n 个变量的联合分布 $P(X_1, X_2, \cdots, X_n)$，有

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2 \mid X_1) \cdots P(X_n \mid X_1, \cdots, X_{n-1}). \tag{1.10}$$

式(1.10)将一个联合分布分解为一系列条件分布的乘积，它称为链规则.

1.3.4 边缘独立与条件独立

1. 事件独立

设 A，B 为同一随机试验的两个不同事件. 我们称**事件 A 与 B 相互独立**，如果下式成立：

$$P(A \cap B) = P(A)P(B), \tag{1.11}$$

当 $P(B) > 0$ 时，由式 (1.11) 可得 $P(A) = P(A \mid B)$. $P(A \mid B)$ 是已知事件 B 发生时对 A 发生的信度，而 $P(A)$ 是在未知事件 B 是否发生时对 A 发生的信度. 所以 A 与 B 相互独立的直观含义是：对于事件 B 是否发生的了解不影响对事件 A 发生的信度. 当 $P(A) > 0$ 时，由式 (1.11) 可得 $P(B) = P(B \mid A)$，所以 A 与 B 相互独立意味着：对于事件 A 是否发生的了解也不影响对事件 B 发生的信度.

考虑 3 个事件 A，B 和 C，假定 $P(C) > 0$. 我们称**事件 A 与 B 在给定 C 时相互条件独立**，如果下式成立：

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C), \tag{1.12}$$

当 $P(B \cap C) > 0$ 时，由式 (1.12) 可得 $P(A \mid C) = P(A \mid B \cap C)$. $P(A \mid C)$ 是已知事件 C 发生时对事件 A 发生的信度，而 $P(A \mid B \cap C)$ 是已知事件 B 和 C 都已发生时对事件 A 发生的信度. 所以，事件 A 与 B 在给定 C 时相互条件独立的直观意义是：在已知事件 C 发生的前提下，对事件 B 是否发生的了解不会改变对事件 A 发生的信度；同样，对事件 A 是否发生的了解也不影响对事件 B 发生的信度.

2. 变量独立

两个随机变量 X 和 Y 称为**相互（边缘）独立**，记为 $X \perp Y$，如果下式成立：

$$P(X, Y) = P(X)P(Y), \tag{1.13}$$

考虑变量 Y 的某个取值 y，如果 $P(Y = y) > 0$，则由式 (1.13) 可得

$$P(X) = P(X \mid Y = y).$$

$P(X \mid Y = y)$ 是已知 $Y = y$ 时，变量 X 的概率(信度)分布，而 $P(X)$ 是未知 Y 的取值时 X 的概率(信度)分布. 所以，变量 X 与 Y 相互独立意味着：对变量 Y 的取值的了解不会改变变量 X 的概率(信度)分布；同样，对变量 X 的取值的了解也不会改变变量 Y 的概率(信度)分布.

更一般地，我们称随机变量 X_1，X_2，…，X_n 相互（边缘）独立，如果

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2)\cdots P(X_n).$$

例 1.20 在例 1.16 中，从 3 个袋子中抽球，所得球的颜色的联合概率分布

$P(X,Y,Z)$ 如下：

X	Y	Z	$P(X,Y,Z)$
w	w	w	0
w	w	b	0.1
w	b	w	0
w	b	b	0.4
b	w	w	0
b	w	b	0.1
b	b	w	0
b	b	b	0.4

而边缘分布 $P(X)$，$P(Y)$ 和 $P(Z)$ 则分别如下：

X	w	b
$P(X)$	0.5	0.5

Y	w	b
$P(Y)$	0.2	0.8

Z	w	b
$P(Z)$	0	1

容易验证 $P(X,Y,Z) = P(X)P(Y)P(Z)$，即 X,Y,Z 相互边缘独立. □

考虑 3 个随机变量 X，Y 和 Z，设 $P(Z = z) > 0, \forall z \in \Omega_Z$. 我们说 X 和 Y **在给定 Z 时相互条件独立**，记为 $X \perp Y \mid Z$. 如果下式成立：

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z). \tag{1.14}$$

设 y 和 z 分别是 Y 和 Z 的任意取值，且 $P(Y = y, Z = z) > 0$，由式 (1.14) 可得

$$P(X \mid Y = y, Z = z) = P(X \mid Z = z).$$

$P(X \mid Z = z)$ 是在已知 $Z = z$ 时，X 的概率（信度）分布，而 $P(X \mid Y = y, Z = z)$ 是在已知 $Y = y$ 以及 $Z = z$ 时，X 的概率（信度）分布. 因此，$X \perp Y \mid Z$ 的直观含意是：在已知 Z 的前提下，对 Y 的取值的了解不影响 X 的概率（信度）分布. 注意，这并不意味着在未知 Z 的取值时，X 和 Y 相互独立. $Y = y$ 有可能含有关于 X 的信息，只是所有这样的信息也都包含于 $Z = z$ 中，所以当已知 $Z = z$ 时，进一步了解到 $Y = y$ 并不增加关于 X 的信息. 当然，$X \perp Y \mid Z$ 也意味着，在已知 Z 的取值时，对 X 的取值的了解不影响 Y 的概率（信度）分布.

例 1.21 设有一装有两种硬币的口袋，其中一些是均匀硬币，掷出正面朝上的概率为 0.5；另一些为非均匀硬币，掷出正面朝上的概率为 0.8. 现从袋中随机取出一个硬币，抛掷若干次. 令 X_i 表示第 i 次抛掷硬币的结果，Y 表示该硬币

是否均匀. 这里，X_i 与 X_j $(i \neq j)$ 之间不是相互（边缘）独立的，因为如果掷了10次硬币，其中9次都是正面朝上，那么有理由相信这枚硬币是不均匀的，从而增大了下一次掷出正面朝上的信度. 所以 X_i 的值给了我们关于这枚硬币的一些信息，它有助于我们继续判断 X_j 的值.

另一方面，如果已经知道了 Y 的值，例如该硬币是不均匀的，那么不管前面的结果如何，以后每次掷硬币的结果为正面的概率都是0.8，我们将不能从前面的试验得到什么信息. 所以给定 Y 的值后，X_i 与 X_j 之间就是相互条件独立的. 本例中变量间的依赖关系可以用图1.2来表示：变量 Y 切断了变量 X_i 与变量 X_j 之间的“信息通道”. □

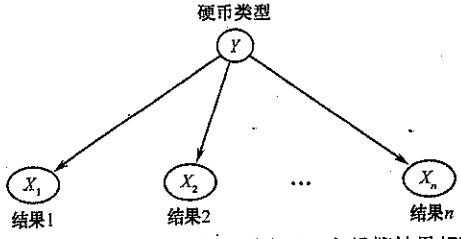


图1.2 条件独立示意：给定硬币类型，各投掷结果相互独立

命题1.1 考虑3个随机变量 X，Y 和 Z，设 $P(Z)>0$，下列条件相互等价：

(1) $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$；

(2) $P(X \mid Y, Z) = P(X \mid Z)$，当 $P(Y, Z)>0$；

(3) $P(X, Y \mid Z) = f(X, Z)g(Y, Z)$，$f$ 和 g 均为函数；

(4) $P(X \mid Y, Z) = f(X, Z)$，f 为一函数，当 $P(Y, Z)>0$；

(5) $P(X, Y, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$；

(6) $P(X, Y, Z) = \dfrac{P(X, Z)P(Y, Z)}{P(Z)}$；

(7) $P(X, Y, Z) = f(X, Z)g(Y, Z)$，$f$ 和 g 均为函数.

1.3.5 贝叶斯定理

先验概率和后验概率这两个概念是相对于某组证据而言的. 设 H 和 E 为两个随机变量，$H=h$ 为某一假设，$E=e$ 为一组证据. 在考虑证据 $E=e$ 之前，对事件 $H=h$ 的概率估计 $P(H=h)$ 称为**先验概率**. 而在考虑证据之后，对 $H=h$ 的概率估计 $P(H=h \mid E=e)$ 称为**后验概率**. **贝叶斯定理**描述了先验概率和后验概率之间的关系：

$$P(H=h \mid E=e) = \frac{P(H=h)\,P(E=e \mid H=h)}{P(E=e)}, \tag{1.15}$$

这又称为**贝叶斯规则**，或**贝叶斯公式**.

例1.22 假设有两个装满糖果的口袋 a 和 b. 袋 a 中有10块水果糖和30块巧克力糖. 袋 b 中水果糖和巧克力糖各有20块. 随机选择一个口袋，并从中随机取出一块糖果，发现是巧克力糖. 问这块糖是从袋 a 中抽出的可能性为多大?

用 C 表示取出的糖块，H 表示它来自的口袋，T 表示它的类型. 这里所关心的假设是 $H=a$，而证据是 $T=$“巧克力糖”. 在考虑证据之前，C 来自袋 a 和袋 b 的概率相同，都是0.5，所以先验概率是 $P(H=a)=0.5$. $P(T=$“巧克力糖”$\mid H=a)$ 是在已知 C 来自袋 a 的情况下它为巧克力的概率，即

$$P(T=\text{“巧克力糖”} \mid H=a) = 30/40 = 0.75.$$

由于总共有80块糖，其中的50块是巧克力，所以

$$P(T=\text{“巧克力糖”}) = 50/80 = 0.625.$$

利用贝叶斯公式，有

$$P(H=a \mid T=\text{“巧克力糖”}) = \frac{P(H=a)P(T=\text{“巧克力糖”} \mid H=a)}{P(T=\text{“巧克力糖”})} = \frac{0.5\times 0.75}{0.625} = 0.6.$$

这就是 $H=a$ 的后验概率. □

在贝叶斯定理中，$P(E=e \mid H=h)$ 称为 $H=h$ 的**似然度**，有时记为 $L(H=h \mid E=e)$. 贝叶斯定理之所以有用是因为似然度往往容易获得，而后验概率则不然. 在例1.22中，似然度 $P(T=$“巧克力糖”$\mid H=a)$ 由袋中不同糖果的组成所决定，是容易得到的；而后验概率 $P(H=a \mid T=$“巧克力糖”$)$ 则要通过一番推理计算才能得到. 下面再来看一个例子.

例1.23 设有一病人眼睛呈黄色，医生要判断他患乙肝的可能性. 这里的假设是病人患有乙肝 $D=t$，证据是病人眼睛呈黄色 $C=y$. 要直接判断 $P(D=t \mid C=y)$ 并不容易. 但根据对乙肝的了解，医生知道乙肝病人眼黄的概率是80%，即似然度 $P(C=y \mid D=t)=0.8$；而在医生的所有病人中，0.5% 的人患有乙肝，10%的人眼黄，即先验概率 $P(D=t)=0.005$，而 $P(C=y)=0.1$. 利用贝叶斯公式，有

$$P(D=t \mid C=y) = \frac{P(D=t)P(C=y \mid D=t)}{P(C=y)} = \frac{0.005\times 0.8}{0.1} = 0.04.$$

即病人患乙肝的概率是0.04，这是先验概率0.005的8倍. □

相对于证据 $E=e$，可以谈论一个事件的先验和后验概率，同时也可以谈论一个变量的先验和后验概率分布. 设 X 是一个所关心的变量，则有

$$P(X \mid E=e) = \frac{P(X)P(E=e \mid X)}{P(E=e)}. \tag{1.16}$$

这是贝叶斯定理的变量形式. 式中 $P(X)$ 是 X 的**先验分布**，$P(X \mid E=e)$ 是 X 的**后验分布**，$P(E=e \mid X)$ 称为 X 的**似然函数**，有时记为 $L(X \mid E=e)$. $P(E=e)$ 是一个**规一化常数**，它保证式子右边的函数是一个概率分布，其值由下式定义：

$$P(E=e) \doteq \sum_X P(X)\ P(E=e \mid X).$$

式 (1.16) 中，$P(E=e)$ 不依赖 X，所以有时把它写为如下形式：

$$P(X \mid E=e) \propto P(X)L(X \mid E=e), \tag{1.17}$$

即后验分布正比于先验分布和似然函数的乘积.

1.4 概率论与人工智能

贝叶斯网是为了处理人工智能研究中的不确定性问题而发展起来的. 人工智能兴起于 20 世纪 50 年代中期，它的目标是研究人类智能的机理，提供智能行为的计算模型，进而构造能具有智能行为的系统. 专家系统是人工智能的一个子领域，它的目标是将某一复杂领域的专家的知识和经验引入计算机系统，使得更多人能够借助计算机系统受惠于专家的经验.

不确定性是人工智能系统所面临的一个重要问题. 它来源于多个方面. 首先，智能系统对外部环境的观测往往是不完备的或是有误差的. 其次，推理涉及对复杂世界一定程度的抽象和简化，因此推理的前提往往因为例外情况而得不到完全满足. 在实际应用中，推理前提的例外很多，不能穷举，不确定性提供了一个总结各种例外情况的机制. 第三，多数复杂领域并没有一套完备的理论，从而推理必须在不确定中进行. 最后，有些客观规律本身就是统计的、随机的、非确定性的.

在六七十年代人工智能的早期研究中，基于规则的专家系统占据主导地位. 这类系统又称为产生式系统，它的知识库由多个 "IF-THEN" 语句构成，使用一阶逻辑和谓词演算进行符号推理. 由于产生式系统在早期人工智能研究中取得了较大成就，所以人们最先想到的是对它进行简单扩充以处理不确定性. 有些学者致力于发明各种新的逻辑，用非单调推理[①]的方式来处理推理前提的例外，这些逻辑包括：默认逻辑 (Reiter, 1987)、非单调模态逻辑 (McDermott and Doyle, 1987)、自认知逻辑 (Moore, 1987)、限定逻辑 (McCarthy, 1980) 等.

① 在高等数学中函数的单调性指的是如下特性：如果 $x \geqslant y$，那么 $f(x) \geqslant f(y)$. 相似地，一阶谓词演算也具有如下单调性：如果 A 是一个规则集合，则随着新规则的加入，它所能推出的结论集合也单调增大. 因此一阶逻辑通常被称为是单调逻辑，相应的推理叫单调推理. 可是，日常生活中人们的常识推理却是非单调的：我们通常直接 "跳入" 某一结论，而后随着新事实的发现再对原来以为为真的结论进行修正. 这种推理叫非单调推理 (Ginsberg, 1987).

另一些学者则试图对产生式系统进行简单修补，通过为推理规则附加一个数字来量化不确定性. 对这个附加数字的语义解释有多种，包括确定因子 (Shortliffe, 1976)、概率数 (Duda et al., 1976)，以及模糊集合论中的隶属度 (Zadeh, 1965) 等. 针对不同的解释，又有各种不同的演算规则. 这些方法在实际应用中都有各种各样的困难，主要原因是计算复杂度太高和推理结果的正确性不能保证. 有关详细讨论请读者参见有关文献 (Heckerman, 1985; Pearl, 1988; Shafer and Pearl, 1990).

概率论是研究随机性的数学. 用概率论处理不确定性的主要优点是能够保证推理结果的正确性. 但是，在 20 世纪 80 年代以前，人们普遍认为概率论不适用于人工智能的研究，这有几个原因. 首先，**频率学派**占支配地位，他们认为概率只有在试验可以在同等条件下无限次重复时才有意义，这就排除了对一次性事件谈论概率的可能性. 其次，人们还没有想到利用问题的结构对联合概率分布进行分解，所以使用概率的计算复杂度太高，难以实现. 第三，不少学者认为数值运算是计算机的专长，而符号推理才是人类智能的独有特点，因此符号逻辑，而非概率数值运算，才是研究人工智能的恰当框架. 另外，一些学者则往往倚重一些启发式方法和编程技巧来实现某种程度的智能行为.

自 80 年代以来，概率论重新引起了人们的兴趣，逐渐成为处理人工智能中不确定性的主流方法. 这有三方面的原因. 首先，到 70 年代末，越来越多的人开始认同概率的贝叶斯解释，将概率当做人的主观信念程度来看待，这使得在频率数据缺乏的情况下也可以使用概率 (Shafer and Pearl, 1990). 其次，**决策论** (Von Neumann and Morgenstern, 1947) 与人工智能的结合对概率论的复兴起到了巨大的推动作用. 决策论的一个核心法则是**最大期望效用**原则，基于决策论的智能系统被称为规范型系统. 与之相对，早期的人工智能系统侧重于描述和总结专家解决问题的逻辑推理过程或所使用的启发式规则，称为**描述型系统**. 两者的主要区别在于：①描述型系统是对专家建模，规范型系统是对问题领域建模；②描述型系统所使用的不确定性运算规则不能保证推理结果的正确性，规范型系统则使用概率运算规则和决策论，可以得到正确的结果；③描述型系统强调替代专家决策，规范型系统强调支持专家决策 (Jensen, 1996). 第三，在 80 年代初，有学者发现，利用问题的结构可以把联合概率分布进行分解，从而大大降低计算复杂度 (Pearl, 1982; Kim and Pearl, 1983; Pearl, 1986)，这对概率方法在人工智能研究中的崛起起到了关键性的作用.

1.5 信息论基础

信息论是建筑于概率论之上的研究信息传输和信息处理的数学理论. 它不仅

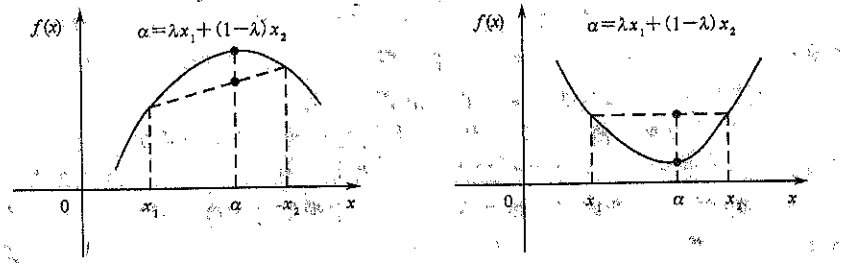
是信息技术的基础，还在诸如统计力学、机器学习等其它领域中起着重要作用. 本书的第三部分将会利用信息论的一些结果来对贝叶斯网学习算法进行分析. 同时，信息论也有助于加深读者对于“条件独立”这一概念的理解. 本节扼要介绍信息论的一些基本概念和结果.

1.5.1 Jensen 不等式

一个函数 f 在实数轴的某个区间 I 上被称为**凹函数**，如果 $\forall x_1, x_2 \in I$，有

$$f(\lambda x_1 + (1-\lambda)x_2) \geqslant \lambda f(x_1) + (1-\lambda) f(x_2), \forall \lambda \in [0,1]. \quad (1.18)$$

这一关系如图 1.3(a)所示. 若式 (1.18) 中的等号只在 $x_1 = x_2$ 时才成立，则称 f 在区间 I 上**严格凹**. 如果将式 (1.18) 中的不等号改变方向，就得到**凸函数**的定义，如图 1.3(b)所示[①]. 如果 f 是凹函数，则 $-f$ 是凸函数. 图 1.4 给出了信息论中几个常见的凹函数.



(a)凹函数　　(b)凸函数

图 1.3　凹函数和凸函数

定理 1.1 (Jensen 不等式)　设 f 为区间 I 上的凹函数，$p_i \in [0,1], i = 1, 2, \cdots, n$，且 $\sum_{i=1}^{n} p_i = 1$. 则对任何 $x_i \in I$，有

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \geqslant \sum_{i=1}^{n} p_i f(x_i). \quad (1.19)$$

若 f 严格凹，则式 (1.19) 的等号只在下列条件满足时才成立：若 $p_i \cdot p_j \neq 0$，则必有 $x_i = x_j$.

证明　用归纳法证明. 当 $n = 1$ 时，式 (1.19) 显然成立. 假设式 (1.19)

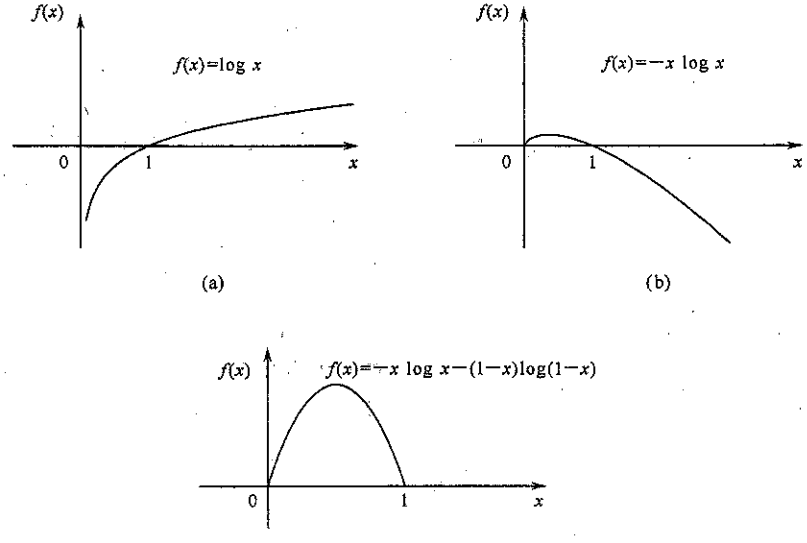
① 注意，凸函数和凹函数的数学曲线与汉字中“凸”和“凹”的直观形状刚好相反. 英文中凸函数是“convex”，对应开口向上的下凸曲线“∪”；凹函数是“concave”，对应开口向下的上凸曲线“∩”. 这一点经常容易引起混淆.

在 $n = k$ 时成立，证明它在 $n = k+1$ 时也成立，即

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} p_i x_i\right) &= f\left(\sum_{i=1}^{k} p_i x_i + p_{k+1} x_{k+1}\right) \\
&= f\left[(1-p_{k+1}) \frac{1}{1-p_{k+1}} \sum_{i=1}^{k} p_i x_i + p_{k+1} x_{k+1}\right] \text{(假设 } p_{k+1} \neq 1\text{)} \\
&\geqslant (1-p_{k+1}) f\left(\frac{1}{1-p_{k+1}} \sum_{i=1}^{k} p_i x_i\right) + p_{k+1} f(x_{k+1}) \quad \text{(根据凹函数定义)} \\
&= (1-p_{k+1}) f\left(\sum_{i=1}^{k} \frac{p_i}{1-p_{k+1}} x_i\right) + p_{k+1} f(x_{k+1}) \\
&\geqslant (1-p_{k+1}) \sum_{i=1}^{k} \frac{p_i}{1-p_{k+1}} f(x_i) + p_{k+1} f(x_{k+1}) \quad \text{(根据归纳假设)} \\
&= \sum_{i=1}^{k} p_i f(x_i) + p_{k+1} f(x_{k+1}) \\
&= \sum_{i=1}^{k+1} p_i f(x_i).
\end{aligned}$$

式 (1.19) 得证. 我们把 f 严格为凹的部分的证明作为练习留给读者. □

Jensen 不等式是凹函数的基本性质，在信息论中经常用到. 也有关于凸函



(a)　(b)

(c)

图 1.4　信息论中常见的凹函数

数的 Jensen 不等式，它可以通过将式（1.19）中的不等号换个方向而得到.

例 1.24 图 1.4(a)所示，对数函数 $f(x)=\log x$ 在区间 $(0,+\infty)$ 是凹函数. 根据 Jensen 不等式，有

$$\log\left(\sum_{i=1}^{n} p_i x_i\right) \geqslant \sum_{i=1}^{n} p_i \log x_i. \tag{1.20}$$

式（1.20）中，$\forall i, x_i>0$，$p_i \geqslant 0$，且 $\sum_{i=1}^{n} p_i = 1$. □

1.5.2 熵

一个离散随机变量 X 的**熵** $H(X)$ 的定义为

$$H(X) = \sum_{X} P(X)\log\frac{1}{P(X)} = -\sum_{X} P(X)\log P(X), \tag{1.21}$$

其中约定 $0\log\frac{1}{0}=0$. 式（1.21）的对数若以 2 为底，则熵的单位是**比特**；若以 e 为底，则其单位是**奈特**. 若无特殊说明，本书以后章节均采用比特为单位.

熵是对随机变量的不确定性的度量. 随机变量 X 的熵越大，说明它的不确定性也越大. 下面举两个例子来说明这一点.

例 1.25 考虑一个取值为 0 或 1 的随机变量 X，记 $p=P(X=1)$. 根据熵的定义，有

$$H(X) = -p\log p-(1-p)\log(1-p).$$

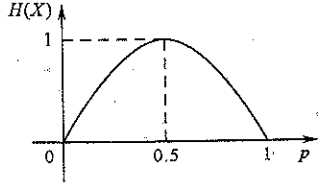


图 1.5 二值随机变量的熵，这里 $p=P(X=1)$

这个函数如图 1.5 所示. 当 $p=0$ 或 $p=1$ 时，我们肯定地知道 X 的取值，不确定性最小，$H(X)=0$. 当 $p=0.5$ 时，相当于投掷一个质地均匀的硬币，对 X 的取值的不确定性达到最大，此时 $H(X)=1$. □

例 1.26 记 X, Y 和 Z 分别为掷硬币、掷骰子以及从 54 张扑克牌中随意抽取一张的结果. 显然 X 的不确定性最小，Y 的不确定性居中，而 Z 的不确定性最大. 与之相应，这 3 个随机变量的熵之间也恰恰存在这样的关系，即

$$H(X)<H(Y)<H(Z);$$

$$H(X) = \frac{1}{2}\log 2+\frac{1}{2}\log 2=\log 2=1;$$

$$H(Y) = \underbrace{\frac{1}{6}\log 6+\cdots+\frac{1}{6}\log 6}_{6\text{个}}=\log 6;$$

$$H(Z) = \underbrace{\frac{1}{54}\log 54+\cdots+\frac{1}{54}\log 54}_{54\text{个}}=\log 54.$$

□

用 $|X|$ 来记变量 X 的取值个数，又称为变量的**势**.

命题 1.2（熵的基本性质）

(1) $H(X)\geqslant 0$；

(2) $H(X)\leqslant\log|X|$，等号成立当且仅当对 X 的所有取值 x 有 $P(X=x)=\frac{1}{|X|}$ ①.

证明 (1) 显然成立，因为对于 X 的任意取值 x，总有

$$-P(X=x)\log P(X=x)\geqslant 0.$$

对于 (2)，根据式（1.20），有

$$H(X) = \sum_{X} P(X)\log\frac{1}{P(X)} \leqslant \log\sum_{X} P(X)\frac{1}{P(X)} = \log|X|.$$

命题得证. □

1.5.3 联合熵、条件熵和互信息

联合熵是借助联合概率分布对熵的自然推广. 两个离散随机变量 X 和 Y 的联合熵的定义为

$$H(X,Y) = \sum_{X,Y} P(X,Y)\log\frac{1}{P(X,Y)} = -\sum_{X,Y} P(X,Y)\log P(X,Y). \tag{1.22}$$

条件熵是利用条件概率分布对熵的一个延伸. 随机变量 X 的熵是用它的概率分布 $P(X)$ 来定义的. 如果知道另一个随机变量 Y 的取值为 y，那么 X 的后验分布即为 $P(X\mid Y=y)$. 利用此条件分布可以定义给定 $Y=y$ 时 X 的条件熵为

$$H(X\mid Y=y) = \sum_{X} P(X\mid Y=y)\log\frac{1}{P(X\mid Y=y)}. \tag{1.23}$$

熵 $H(X)$ 度量的是随机变量 X 的不确定性，条件熵 $H(X\mid Y=y)$ 度量的则是已知 $Y=y$ 后，X 的不确定性.

在式(1.23)中，当 y 变化时，$H(X\mid Y=y)$ 也会发生改变. 由于知道 Y 的概率分布，因此可以计算观测 Y 后 X 的熵的期望值，即

① 这个性质经常被称为**最大熵原理**.

$$
\begin{aligned}
H(X \mid Y) &= \sum_{y \in \Omega_Y} P(Y=y)\, H(X \mid Y=y) \\
&= \sum_{y \in \Omega_Y} P(Y=y) \sum_{x} P(X \mid Y=y) \log \frac{1}{P(X \mid Y=y)} \\
&= \sum_{Y} \sum_{X} P(Y) P(X \mid Y) \log \frac{1}{P(X \mid Y)} \\
&= \sum_{X,Y} P(X, Y) \log \frac{1}{P(X \mid Y)}.
\end{aligned} \tag{1.24}
$$

$H(X \mid Y)$ 称为给定 Y 时 X 的条件熵.

注意 $H(X \mid Y)$ 与 $H(X \mid Y=y)$ 有所不同. 后者是在已知 Y 取某一特定值 y 时 X 的条件熵，或者说是在已知 $Y=y$ 后，X 剩余的不确定性. 而 $H(X \mid Y)$ 则是在未知 Y 的取值时，对观测到 Y 的取值后 X 剩余的不确定性的一个期望. 尤其值得注意的是，$H(X \mid Y=y)$ 可能比 $H(X)$ 大，即知道 Y 的具体取值 $Y=y$ 可能增大对 X 的不确定性；但读者在 1.5.5 节将看到，$H(X \mid Y)$ 永远不大于 $H(X)$，即平均来说，知道 Y 将不会增加 X 的不确定性.

例 1.27 设联合分布 $P(X,Y)$ 以及边缘分布 $P(X)$ 和 $P(Y)$ 如下：

	x_1	x_2	$P(Y)$
y_1	0	$\frac{3}{4}$	$\frac{3}{4}$
y_2	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$P(X)$	$\frac{1}{8}$	$\frac{7}{8}$	

从而可以得出

$$
\begin{aligned}
&H(X) = -\frac{1}{8}\log\frac{1}{8} - \frac{7}{8}\log\frac{7}{8} = 0.544; \\
&H(X \mid Y=y_1) = -0\log 0 - 1\log 1 = 0; \\
&H(X \mid Y=y_2) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1; \\
&H(X \mid Y) = \frac{3}{4} H(X \mid Y=y_1) + \frac{1}{4} H(X \mid Y=y_2) = 0.25.
\end{aligned}
$$

可以看到，观测到 $Y=y_1$ 使 X 的熵减小，观测到 $Y=y_2$ 使 X 的熵增大. 但平均来说，对 Y 的观测使 X 的熵减小. □

在观测到 Y 以前，X 的不确定性是 $H(X)$. 通过观测 Y，我们期望 X 的不确定性会变为 $H(X \mid Y)$. 因此 $H(X)$ 与 $H(X \mid Y)$ 之差

$$
I(X;Y) = H(X) - H(X \mid Y) \tag{1.25}
$$

就是对 Y 包含多少关于 X 的信息的一个度量，称之为 Y 关于 X 的信息. 很快我们将看到 $I(X;Y) = I(Y;X)$，因此它又称为 X 和 Y 之间的互信息. 以下定理给出联合熵、相对熵和互信息之间的重要关系.

定理 1.2 对任意两个离散随机变量 X 和 Y，有

$$
I(X;Y) = \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}, \tag{1.26}
$$

$$
I(X;Y) = I(Y;X), \tag{1.27}
$$

$$
H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y), \tag{1.28}
$$

$$
I(X;Y) + H(X,Y) = H(X) + H(Y). \tag{1.29}
$$

其中式 (1.28) 称为**熵的链规则**.

证明 (1) 对式 (1.26)，有

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= \sum_{X} P(X) \log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) \log \frac{1}{P(X \mid Y)} \\
&= \sum_{X,Y} P(X,Y) \log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) \log \frac{1}{P(X \mid Y)} \\
&= \sum_{X,Y} P(X,Y) \log \frac{P(X \mid Y)}{P(X)} \\
&= \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}.
\end{aligned}
$$

(2) 对式 (1.27)，这是第(1)步结果的显然推论.

(3) 对式 (1.28)，有

$$
\begin{aligned}
H(X,Y) &= -\sum_{X,Y} P(X,Y) \log P(X,Y) \\
&= -\sum_{X,Y} P(X,Y) \log P(X) - \sum_{X,Y} P(X,Y) \log P(Y \mid X) \\
&= -\sum_{X} P(X) \log P(X) - \sum_{X,Y} P(X,Y) \log P(Y \mid X) \\
&= H(X) + H(Y \mid X).
\end{aligned}
$$

同理可证 $H(X,Y) = H(Y) + H(X \mid Y)$.

(4) 对式 (1.29)，有

$$
\begin{aligned}
I(X;Y) + H(X,Y) &= (H(X) - H(X \mid Y)) + (H(Y) + H(X \mid Y)) \\
&= H(X) + H(Y).
\end{aligned}
$$

定理得证. □

联合熵、条件熵和互信息之间的关系可以用如图 1.6 所示的韦恩图来总结.

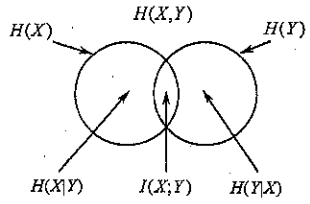


图 1.6 联合熵、条件熵以及互信息之间的关系

1.5.4 相对熵

对定义于随机变量 X 的状态空间 Ω_X 上的两个概率分布 $P(X)$ 和 $Q(X)$，可以用**相对熵**来度量它们之间的差异，即有

$$KL(P,Q)=\sum_X P(X)\log\frac{P(X)}{Q(X)}. \tag{1.30}$$

其中约定：$0\log\frac{0}{q}=0$；$p\log\frac{p}{0}=\infty,\forall p>0$. $KL(P,Q)$ 又被称为 $P(X)$ 和 $Q(X)$ 之间的 **Kullback-Leibler 距离**．但严格来讲它不是一个真正意义上的距离，因为 $KL(P,Q)\neq KL(Q,P)$.

定理 1.3（信息不等式） 设 $P(X)$ 和 $Q(X)$ 为定义在某个变量 X 的状态空间 Ω_X 上的两个概率分布，则有

$$KL(P,Q)\geqslant 0, \tag{1.31}$$

其中，当且仅当 P 与 Q 相同，即 $P(X=x)=Q(X=x)$，$\forall x\in\Omega_X$ 时等号成立.

证明

$$\begin{aligned}\sum_X P(X)\log\frac{P(X)}{Q(X)}&=-\sum_X P(X)\log\frac{Q(X)}{P(X)}\\&\geqslant-\log\sum_X P(X)\frac{Q(X)}{P(X)}\text{(根据式(1.20))}\\&=-\log\sum_X Q(X)\\&=-\log 1=0.\end{aligned}$$

定理得证. □

以下推论在本书的第三部分将会被多次用到.

推论 1.1 对于满足 $\sum_X f(X)>0$ 的非负函数 $f(X)$，定义概率分布 $P^*(X)$ 为

$$P^*(X)=\frac{f(X)}{\sum_X f(X)}.$$

那么对于任意其它的概率分布 $P(X)$，则有

$$\sum_X f(X)\log P^*(X)\geqslant\sum_X f(X)\log P(X),$$

其中当且仅当 P^* 与 P 相同时等号成立.

证明 根据定理 1.3，有

$$KL(P^*,P)=\sum_X P^*(X)\log\frac{P^*(X)}{P(X)}\geqslant 0.$$

因此有

$$\sum_X P^*(X)\log P^*(X)\geqslant\sum_X P^*(X)\log P(X),$$

即

$$\sum_X\frac{f(X)}{\sum_X f(X)}\log P^*(X)\geqslant\sum_X\frac{f(X)}{\sum_X f(X)}\log P(X).$$

从而有

$$\sum_X f(X)\log P^*(X)\geqslant\sum_X f(X)\log P(X).$$

推论得证. □

1.5.5 互信息与变量独立

本节指出互信息与变量独立之间的两个关系．首先有以下定理.

定理 1.4 对任意两个离散随机变量 X 和 Y，有

(1) $I(X;Y)\geqslant 0$;

(2) $H(X\mid Y)\leqslant H(X)$.

上面两式当且仅当 X 与 Y 相互独立时等号成立.

证明 由式 (1.26) 可得

$$I(X;Y)=KL(P(X,Y),P(X)P(Y)), \tag{1.32}$$

即 $I(X;Y)$ 是分布于 $P(X,Y)$ 和 $P(X)P(Y)$ 之间的相对熵．根据信息不等式，$I(X;Y)\geqslant 0$，当且仅当 $P(X,Y)=P(X)P(Y)$ 时等号成立．亦即 $I(X;Y)=0$ 当且仅当 X 与 Y 相互独立．由于 $I(X;Y)=H(X)-H(X\mid Y)$，所以 $H(X\mid Y)\leqslant H(X)$，而且 $H(X\mid Y)=H(X)$ 当且仅当 X 与 Y 相互独立．定理得证. □

定理 1.4 从信息论角度为“边缘独立”这一概念提供了一个直观解释，即两个随机变量相互独立当且仅当它们之间的互信息为零.

接下来考虑 3 个变量 X, Y 和 Z 之间的条件独立关系. 条件熵 $H(X \mid Z)$ 是给定 Z 时 X 剩余的不确定性. 如果进一步再给定 Y, X 剩余的不确定性变为 $H(X \mid Z, Y)$. 因此这两者之差即是给定 Z 时观测 Y 的取值会带来的关于 X 的信息量, 即

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Z, Y), \tag{1.33}$$

称为给定 Z 时 Y 关于 X 的信息. 容易证明 $I(X;Y \mid Z) = I(Y;X \mid Z)$, 于是 $I(X;Y \mid Z)$ 也称为给定 Z 时 X 和 Y 之间的**条件互信息**.

定理 1.5 对任意 3 个离散随机变量 X, Y 和 Z, 有

(1) $I(X;Y \mid Z) \geqslant 0$;

(2) $H(X \mid Y,Z) \leqslant H(X \mid Z)$.

上面两式当且仅当 $X \perp Y \mid Z$ 时等号成立.

证明 从条件互信息的定义出发, 有

$$\begin{aligned}
I(X; Y \mid Z) &= H(X \mid Z) - H(X \mid Y,Z) \\
&= \sum_{X,Z} P(X, Z) \log \frac{1}{P(X \mid Z)} - \sum_{X,Y,Z} P(X, Y, Z) \log \frac{1}{P(X \mid Y, Z)} \\
&= \sum_{X,Y,Z} P(X, Y, Z) \log \frac{1}{P(X \mid Z)} - \sum_{X,Y,Z} P(X, Y, Z) \log \frac{1}{P(X \mid Y, Z)} \\
&= \sum_{X,Y,Z} P(X, Y, Z) \log \frac{P(X \mid Y, Z)}{P(X \mid Z)} \\
&= \sum_{Z} P(Z) \sum_{X,Y} P(X, Y \mid Z) \log \frac{P(X, Y \mid Z)}{P(X \mid Z) P(Y \mid Z)} \\
&= \sum_{Z} P(Z) KL(P(X, Y \mid Z), P(X \mid Z) P(Y \mid Z)) \\
&\geqslant 0. \text{(根据定理 1.3)}
\end{aligned}$$

这样就同时证明了上面两式(1)和(2), 而且其中当且仅当 $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$, 即 $X \perp Y \mid Z$ 时等号成立. □

定理 1.5 的意义在于, 它从信息论角度为随机变量之间的“条件独立”这一概念提供了一个直观解释, 即给定 Z, 两个随机变量 X 和 Y 相互条件独立, 当且仅当它们的条件互信息为零. 或者说, Y 关于 X 的信息已全部包含在 Z 中, 从而观测到 Z 后, 再对 Y 进行的观测不会带来关于 X 的更多信息. 另一方面, 如果 X 和 Y 在给定 Z 时相互不独立, 则 $H(X \mid Z,Y) < H(X \mid Z)$, 即在已知 Z 的基础上对 Y 的进一步观测将会带来关于 X 的新信息, 从而降低 X 的不确定性.