普通高等教育"十一五"国家级规划教材

信息论基础

邹永魁 宋立新 编著

科学出版社
www.sciencep.com

普通高等教育"十一五"国家级规划教材

信息论基础

邹永魁　宋立新　编著

科学出版社

北　京

内 容 简 介

本书是为高等学校信息与计算科学专业本科生编写的教材. 内容包括概率论的基础知识, 香农提出的有关信息量化的基本概念、方法和定理, 以及信源的基本编码理论和信道的基本编码理论.

本书可作为高等学校信息与计算科学专业以及信息类专业的教材, 也可作为相关课程的教学参考书.

图书在版编目(CIP)数据

信息论基础/邹永魁, 宋立新编著. —北京: 科学出版社, 2010

普通高等教育“十一五”国家级规划教材

ISBN 978-7-03-027049-8

Ⅰ. 信… Ⅱ. ①邹… ②宋… Ⅲ. 信息论-高等学校-教材 Ⅳ. G201

中国版本图书馆 CIP 数据核字 (2010) 第 047905 号

责任编辑: 王 静 房 阳 / 责任校对: 朱光光
责任印制: 张克忠 / 封面设计: 陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

明辉印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2010 年 4 月第 一 版 开本: B5(720×1000)
2010 年 4 月第一次印刷 印张: 8
印数: 1—4 000 字数: 162 000

定价: 19.00 元

(如有印装质量问题, 我社负责调换)

前　言

信息科学是一门涉及面极广的、有着广泛应用的科学. 其任务是研究信息的定义、性质、传输、处理、控制的基本原理和方法, 也是现代信息科学和技术的一门基础理论. 它的理论基础是从通信科学发展起来的信息论. 香农提出的信息论是一种统计意义上的信息理论. 这一理论对通信技术的发展产生了持久和深刻的影响. 因此, 目前各高等院校的信息类专业 (包括信息与计算科学专业) 都把信息论作为本科生的一门重要的专业基础课. 同时, 由于信息论的思想和方法已广泛地渗透到许多其他的学科, 信息论的许多结果也有相当的普遍意义, 因此信息论在许多其他领域 (如计算机科学、系统科学、统计学、经济学甚至社会学) 都获得了成功的应用.

本书系作者根据多年教学实践经验编著而成. 书中主要介绍了香农信息论的基本概念、基本方法和主要结果, 并用基本的概率论工具阐述香农信息论的基本思想和基本方法.

全书共分 5 章. 第 1 章是绪论. 第 2 章介绍了概率统计的基础知识, 即使没有学习过概率统计的同学通过学习该章的内容后, 也能够继续学习本书的后续内容. 第 3 章介绍了信息量化的基本理论 —— 熵及其性质. 为了强调熵的相对性, 我们从事件的互信息入手引入有关信息的量化概念, 并由此导出事件自信息的概念, 从而给出熵的严格定义. 在介绍每一个概念的时候, 我们都希望通过分析具体的例子, 帮助读者理解相应的概念. 在熵的概念基础上, 我们着重介绍了平稳离散信源的信源熵的定义和计算理论. 第 4 章介绍了信源的编码理论和方法, 香农提出的典型列集合的概念在编码理论中起着至关重要的作用. 我们用信息论的理论和方法阐述了离散无记忆信源的等长编码理论和不等长编码理论, 特别详细地介绍了最佳不等长编码的理论和方法以及香农编码的基本方法, 并从对策论的角度对比了这两种方法的实用有效性; 该章还介绍了平稳有记忆信源 (马尔可夫信源) 的编码定理. 第 5 章介绍了离散无记忆信道的容量和编码定理. 我们首先从一般的意义上给出了信道容量的定义, 这一概念的提出是信息传输理论中的一个里程碑, 它极大地推动了信息理论的发展. 然后我们介绍了几种简单易行的信道容量的计算方法. 在该章的最后, 我们介绍了信道编码理论的一个重要概念 —— 联合典型列对, 并简介了信道编码的基本思想.

感谢吉林大学数学学院信息与计算科学专业的部分同学, 他们阅读了本书的初稿, 提出了许多有价值的参考意见.

本书虽经实际授课使用和多次修改, 疏漏及不足仍在所难免, 敬请读者批评指正.

作 者

2009 年 11 月

目　　录

第 1 章　绪　　论

信息论是应用近代概率统计方法研究信息传输、交换、储存和处理的一门学科, 也是源于通信实践发展起来的一门新兴的应用科学.

信息是系统传输、交换、储存和处理的对象, 信息载荷在语言、文字、数据、图像等消息之中. 在信息论中, 信息和消息是紧密相关的两个不同概念. 同样一个消息, 如一张当日报纸, 不同的人从中获得的信息是不一样的; 同样的天气预报, 如"明天有雨", 对于干旱地区和雨量充沛地区来说, 其信息含量也不一样; 一张纸写上几个字成为一封家书, 对于收信者来说是家书抵万金, 但对别人而言就可能是废纸一张.

因此, 信息是一种奇妙的东西, 它是有别于物质和能量的一种存在. 信息的本质和它的定义是当前科学界乃至哲学界热衷讨论的课题.

信息的度量是信息论研究的基本课题. 从目前的研究来看, 要对通常意义下的信息给出一个统一的度量是困难的. 至今最为成功也是最为普及的信息度量是由信息论创始人香农 (Shannon) 提出的, 是建立在概率模型基础上的信息度量, 他把信息定义为 "用来消除不确定性的东西", 用概率的某种函数来描述不确定性是自然的, 所以香农用

$$I(A) = -\log_a P(A) \tag{1.0.1}$$

来度量事件发生所提供的自信息, 这个定义与人们的直觉经验相吻合. 当不特别强调对数的底 a 具体是什么时, 本书中将省略 a 的书写.

如果一个随机试验有 N 个可能值, 若它们出现的概率分别为 $p_1, p_2, \cdots, p_n$, 则这些事件自信息出现的平均值

$$H = -\sum_{i=1}^{N} p_i \log p_i \tag{1.0.2}$$

作为这个随机试验或随机消息所提供的平均消息也是合理的, H 也称为熵, 这是借用统计物理学中的一个名词. 在物理学中, 熵是描述系统的不规则性或不确定性程度的一个物理量.

信息论所研究的通信系统基本模型如图 1.0.1 所示.

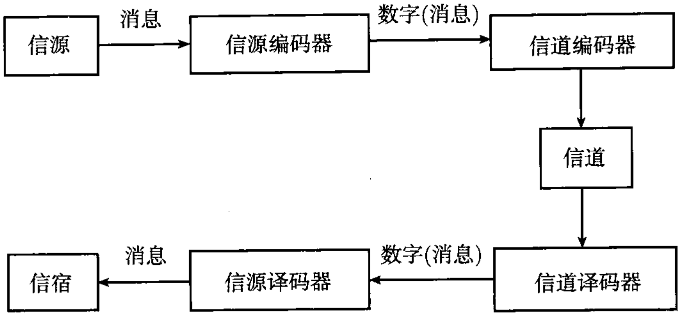

图 1.0.1　通信系统基本模型

信源是产生消息 (或消息序列) 的源. 消息通常是符号序列或时间函数. 消息取值服从一定的统计规律, 所以信源的数学模型可以是一个离散的随机序列或连续的随机过程.

例 1.0.1　*信源* $=\{a,b,\cdots,z,\#\}$, *消息为语言文字, 即字母的序列. 信源的统计规律为各字母在语言文字中出现的概率不同.*

信源编码器把信源产生的消息变换成数字序列. 在不允许编码失真的情况下, 信源编码器的目的是在保证其输出数字序列能无错误地恢复输入信息序列的前提下, 降低输出数字序列的速率, 也就是保证在不失真的条件下, 对输入消息序列进行压缩. 或者在允许失真的情况下, 信源编码的目的是对给定信源, 在保证消息平均失真不超过某个给定允许值 D 的条件下, 尽量降低输出数字序列的速率.

信道在实际通信系统中是指传输信号的媒介或通道, 如架空明线、电缆、电离层、人造卫星等. 在信息论的模型中也把发送端和接受端的调制解调器等归入信道, 在信道的输入、输出模型中, 根据噪声和干扰的统计特征, 用输入、输出的条件概率 (或称转移概率) 来描述信道特征.

信道编码器把信源编码输出的数字序列变换成适合于信道传输的, 由信道入口符号组成的序列. 信道编码器最主要的作用是要对其输出序列提供保护, 以抵抗信道噪声和干扰.

信道译码器和信源译码器分别是信道编码和信源编码的反变换, 信宿是消息的接受者.

第 2 章　概率统计的基础知识

2.1　概 率 空 间

掷一粒骰子, 观察出现的点数. 毋庸置疑, 共有 6 种可能结果, 即出现的结果必为 1~6, 可是事先却不能确定将会出现哪一个点数. 一般来讲, 把能够事先明确一次试验的所有可能结果, 但却不能具体确定将会出现哪一种结果的实验称为随机试验, 简称为试验. 在随机试验中, 每一个可能出现的结果称为样本点, 记为 ω; 而样本点的全体称为样本空间, 记为 Ω. 刚才所举的骰子的例子中, 试验的样本空间 $\Omega=\{1,2,3,4,5,6\}$.

掷骰子, 根据不同的目的和不同的情况, 有时希望它出现偶数点或奇数点, 有时希望它出现不小于 5 的点数. 这类所关心的事件, 称为随机事件, 简称为事件. 从数学的角度来看, 事件也可以描述成样本空间的子集. 为了方便起见, 在本书的以后部分, 将事件和子集当成同一名词使用. 刚才例子中两个随机事件分别相当于子集 $A_1=\{2,4,6\}$, $A_2=\{1,3,5\}$, $A_3=\{5,6\}$.

在一次随机试验中, 观察到了一个样本点 $\omega\in\Omega$. 如果 ω 属于子集 A, 即 $\omega\in A$, 那么就称发生了事件 A. 例如, 在掷骰子的例子中, 如果骰子朝上的一面数字出现了 $\omega=5$, 则有 $5\in A_2$, $5\in A_3$, 因而称发生了事件 A_2 和 A_3.

对于任意事件 A 来讲, 其对立事件 $\bar{A}$ 是指当且仅当事件 A 没有发生时发生的事件. 按集合论的表示方法, 记为

$$\bar{A}=\{\omega\in\Omega|\omega\notin A\}.$$

对于任意两个事件 A_1 与 A_2, 其和事件 $A_1\cup A_2$ 是指事件 A_1 与 A_2 中至少有一个发生的事件, 记为

$$A_1\cup A_2=\{\omega\in\Omega|\omega\in A_1\text{或 }\omega\in A_2\}.$$

对于任意两个事件 A_1 与 A_2, 其积事件 $A_1\cap A_2$ 是指事件 A_1 与 A_2 同时发生

的事件, 记为

$$A_1 \cap A_2 = \{\omega \in \Omega | \omega \in A_1 \text{且 } \omega \in A_2\}.$$

特别地, 当 $A_1 \cap A_2 = \varnothing$ 的时候, 称事件 A_1 与事件 A_2 互不相容.

上面所讲的和与积的定义还可以推广到三个及三个以上事件的情况. 如果存在三个及三个以上事件, 并且其中任意两个事件互不相容, 那么就称这些事件互不相容.

一般来讲, 所关心的事件将随着目的和场合的不同而不同, 就是在同一样本空间 Ω 里, 各种各样的事件群都可能成为被考察的对象. 因此, 可以给出如下 σ 域的概念, 它刻画的是按照一定规则构成的事件群.

定义 2.1.1　σ 域 (σ 代数), 记作 $\mathfrak{F}$, 是由 Ω 的一些子集构成的集类且满足

(1) $\Omega \in \mathfrak{F}$;

(2) 若 $A \in \mathfrak{F}$, 则 $\bar{A} \in \mathfrak{F}$;

(3) 若 $A_n \in \mathfrak{F}(n = 1, 2, \cdots)$, 则 $\bigcup\limits_{n=1}^{\infty} A_n \in \mathfrak{F}$.

若 $\mathfrak{F}$ 为 σ 域, 则由 (1), (2) 得 $\varnothing \in \mathfrak{F}$. 若 $A_n \in \mathfrak{F}(n = 1, 2, \cdots)$, 则

$$\bigcap_{n=1}^{+\infty} A_n = \overline{\bigcup_{n=1}^{\infty} \bar{A}_n} \in \mathfrak{F}. \tag{2.1.1}$$

若 $\mathfrak{F}$ 是样本空间 Ω 的一些子集构成的一个 σ 域, 则称它为事件域, $\mathfrak{F}$ 中的元素称为事件, Ω 称为必然事件, $\varnothing$ 称为不可能事件.

单个样本点构成的集合并不一定总是事件. 如在掷骰子试验中, 可以取 $\mathfrak{F} = \{\{1, 3, 5\}, \{2, 4, 6\}, \varnothing, \Omega\}$.

性质 2.1.1　若给定 Ω 的一个非空集类 ξ, 则必存在 Ω 中唯一的 σ 域 $m(\xi)$, 它是 Ω 中所有包含 ξ 的 σ 域中最小的, 也就是说, 如果 $\mathfrak{G}$ 也是包含 ξ 的一个 σ 域, 则必定成立 $m(\xi) \subset \mathfrak{G}$.

例 2.1.1　一维 Borel 点集, 称由一切形如 $[a, b)$ 的有界左闭右开区间构成的集类所产生的 σ 域为一维 Borel σ 域, 记为 $\mathfrak{B}_1$.

性质 2.1.2 若 x, y 为任意实数, 则如下集合都属于 $\mathfrak{B}_1$:

$$\{x\}=\bigcap_{n=1}^{\infty}\left[x, x+\frac{1}{n}\right),$$

$$(x,y)=[x,y)\cap\overline{\{x\}},$$

$$[x,y]=[x,y)\cup\{y\},$$

$$(x,y]=(x,y)\cup\{y\}.$$

因此, $\mathfrak{B}_1$ 中包含一切开区间、闭区间、单点集、可数集以及由它们经过可数次交、并运算得出的集合.

确定好所关心的作为事件集合的 σ 域 $\mathfrak{F}$ 以后, 再去考察包含在 $\mathfrak{F}$ 中的种种事件所发生的概率. 习惯上, 将事件 A 发生的概率记为 $P(A)$.

定义 2.1.2 定义在事件域 $\mathfrak{F}$ 上的一个集合函数 P 称为概率, 如果它满足如下三个条件:

(1) $P(A)\geqslant 0$ 对一切 $A\in\mathfrak{F}$ 都成立;

(2) $P(\Omega)=1$;

(3) 若 $A_i\in\mathfrak{F}\ (i=1,2,\cdots)$ 且两两互不相容, 则

$$P\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}P(A_i).$$

在定义 2.1.2 中, 性质 (3) 称为概率的可数可加性或完全可加性.

称三元体 $(\Omega,\mathfrak{F},P)$ 为概率空间, 其中, Ω 为样本空间, $\mathfrak{F}$ 为事件域, P 为概率.

性质 2.1.3 (概率的基本公式)

$$P(\varnothing)=0,$$

$$P(\bar{A})=1-P(A),$$

$$P\left(\bigcup_{i}A_i\right)=1-P\left(\bigcap_{i}\bar{A}_i\right),$$

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)\ \text{(加法公式)}.$$

性质 2.1.4 (概率的连续性) 对于无限序列 $A_1, A_2, A_3, \cdots$, 如果存在单调包含关系 $A_1 \subset A_2 \subset A_3 \subset \cdots$. 令 $A = \bigcup_{i=1}^{\infty} A_i$, 则

$$P(A) = \lim_{i \to \infty} P(A_i). \tag{2.1.2}$$

如果存在单调包含关系 $A_1 \supset A_2 \supset A_3 \supset \cdots$, 令 $A = \bigcap_{i=1}^{\infty} A_i$, 则

$$P(A) = \lim_{i \to \infty} P(A_i). \tag{2.1.3}$$

2.2 条件概率

例 2.2.1 在一次抽奖会上, 抽选箱里有 3 个红球和 7 个白球. 摇一次抽选箱, 就会掉出一个球来. 如果这个球是红的就中奖, 白的就没中奖. 现在假定掉出来的球不再放回抽选箱中. 这时最初摇奖的人中奖的概率为 3/10. 问:

(1) 在第一次摇奖的人中奖的情况下, 第二次摇奖的人中奖的概率是多少?

(2) 在第一次摇奖的人没中奖的情况下, 第二次摇奖的人中奖的概率是多少?

解 (1) 第一次摇奖的人中奖以后, 抽选箱中还剩下 2 个红球和 7 个白球, 因而接着出现红球的概率为

$$\frac{2}{2+7} = \frac{2}{9}\left(< \frac{3}{10}\right).$$

(2) 若第一次摇奖的人没中奖, 抽选箱中还剩下 3 个红球和 6 个白球, 因而接着出现红球的概率为

$$\frac{3}{3+6} = \frac{3}{9}\left(> \frac{3}{10}\right).$$ ■

例 2.2.1 说明最初的实验结果一般会影响第二次试验结果的概率. 如果假定最初的试验结果发生了事件 B, 那么在第二次试验中事件 A 发生的概率就可记为 $P(A|B)$, 并称它为在事件 B 发生的条件下事件 A 发生的条件概率.

定义 2.2.1 设 $(\Omega, \mathfrak{F}, P)$ 是一个概率空间, $B \in \mathfrak{F}$ 且 $P(B) > 0$, 则对任意事件 $A \in \mathfrak{F}$, 定义

$$P(A|B) = \frac{P(AB)}{P(B)}, \tag{2.2.1}$$

称 $P(A|B)$ 为在事件 B 发生的条件下事件 A 发生的条件概率.

固定事件 B, 将 $P(A|B)$ 看成 A 的函数, 可以证明它也是定义在样本空间 Ω 的 σ 域 $\mathfrak{F}$ 上的一个概率函数, 因此称之为条件概率.

例 2.2.2 先后投掷两个均匀的硬币, 可能的结果为

$$(\text{正}, \text{正}), \quad (\text{正}, \text{反}), \quad (\text{反}, \text{正}), \quad (\text{反}, \text{反}).$$

A 表示投掷的结果中有一个是正的, 一个是反的这一事件, 则 $P(A)=2/4$, B 表示投掷的结果有一个是正面的这一事件, 则 $P(B)=3/4$,

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{P(A)}{P(B)}=\frac{2}{3},$$

则在已知投掷的结果中有一个是正面的条件下, 投出一正、一反的概率为 $2/3$.

由条件概率的定义立即可以得到如下的乘法公式:

性质 2.2.1 (乘法公式)

$$P(A\cap B)=P(B)P(A|B)=P(A)P(B|A). \tag{2.2.2}$$

例 2.2.3 在肝癌普查中发现某地区的自然人群中, 在 100000 人内平均有 40 人患原发性肝癌, 有 34 人甲胎球蛋白含量高, 有 32 人既患有原发性肝癌又出现甲胎球蛋白含量高, 从这个地区的居民中任意选出一人, 若他患有原发性肝癌则记为 C, 若甲胎球蛋白含量高记为 D, 则

$$P(C)=0.0004,$$

$$P(D)=0.00034,$$

$$P(CD)=0.00032.$$

计算条件概率为

$$P(D|C)=\frac{P(DC)}{P(C)}=0.8,$$

$$P(C|D)=\frac{P(CD)}{P(D)}=0.9412.$$

由以上计算可得患原发性肝癌的人中有 80% 的人出现甲胎球蛋白含量高的症状, 而甲胎球蛋白含量的测定有利于发现原发性肝癌患者. 若出现甲胎球蛋白高含量这一现象, 则有 94% 以上的概率对患原发性肝癌的人作出正确诊断.

性质 2.2.2 (全概率公式)　设事件 $A_1, A_2, \cdots, A_n, \cdots$ 是样本空间 Ω 的一个分割, 即 $A_i(i=1,2,\cdots)$ 两两互不相容 $(A_i \bigcap A_j = \varnothing)$ 且 $\sum\limits_{i=1}^{\infty} A_i = \Omega$, 则 $\forall B \in \mathfrak{F}$,

$$P(B) = \sum_{i=1}^{\infty} P(A_i)P(B|A_i). \tag{2.2.3}$$

证明　$B = \sum\limits_{i=1}^{\infty} A_i B$, 因为 A_i 互不相容, 则 $A_i B$ 也互不相容, 从而

$$\begin{aligned} P(B) &= \sum_{i=1}^{\infty} P(A_i B) \\ &= \sum_{i=1}^{\infty} P(A_i)P(B|A_i). \end{aligned} \tag{2.2.4}$$

■

例 2.2.4　送检的两批灯管在运输中各打碎一只, 若每批 10 只, 而第一批灯管中有一个次品, 第二批灯管中有两个次品, 现在从剩下的灯管中任取一只, 问抽得次品的概率是多少?

解　以 B 表示抽的次品这一事件, 以 A_1, A_2, A_3, A_4 分别记前后两批中打碎者为 (次, 次), (次, 好), (好, 次) 和 (好, 好) 这些事件, 则其概率分别为

$$\begin{aligned} P(A_1) &= \frac{2}{100}, \\ P(A_2) &= \frac{8}{100}, \\ P(A_3) &= \frac{18}{100}, \\ P(A_4) &= \frac{72}{100}. \end{aligned}$$

显然, 在事件 A_i 发生的条件下, 事件 B 发生的条件概率容易求得为

$$P(B|A_1)=\frac{1}{18},$$

$$P(B|A_2)=\frac{2}{18},$$

$$P(B|A_3)=\frac{2}{18},$$

$$P(B|A_4)=\frac{3}{18}.$$

因此, 利用全概率公式可求得

$$P(B)=\sum_{i=1}^{4}P(A_i)P(B|A_i)=15\%. \qquad \blacksquare$$

性质 2.2.3(贝叶斯公式) 若 $A_1, A_2, \cdots, A_n, \cdots$ 两两互不相容, 事件 B 满足 $B=\sum_{i=1}^{\infty}BA_i$, 则有

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{\infty}P(A_j)P(B|A_j)}. \tag{2.2.5}$$

证明 因为 $B=\sum_{i=1}^{\infty}BA_i$, 由乘法公式得

$$P(BA_i)=P(A_i)P(B|A_i),$$

再由全概率公式得

$$\begin{aligned}P(A_i|B)&=\frac{P(A_iB)}{P(B)}\\&=\frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{\infty}P(A_j)P(B|A_j)}.\end{aligned} \qquad \blacksquare$$

注 2.2.1 假定 $A_1, A_2, \cdots$ 是导致试验结果 B 的“原因”, $P(A_i)$ 称为先验概率, 它反映了各种“原因”发生的可能性大小, 这种先验概率一般是以往经验的总结, 在一次具体的试验前已经知道. 现在试验之后产生了事件 B, 这个信息将有利于探讨引起这一事件发生的“原因”, 条件概率 $P(A_i|B)$ 称为后验概率, 它反映了试验之后对这种“原因”发生的可能性大小的新认识.