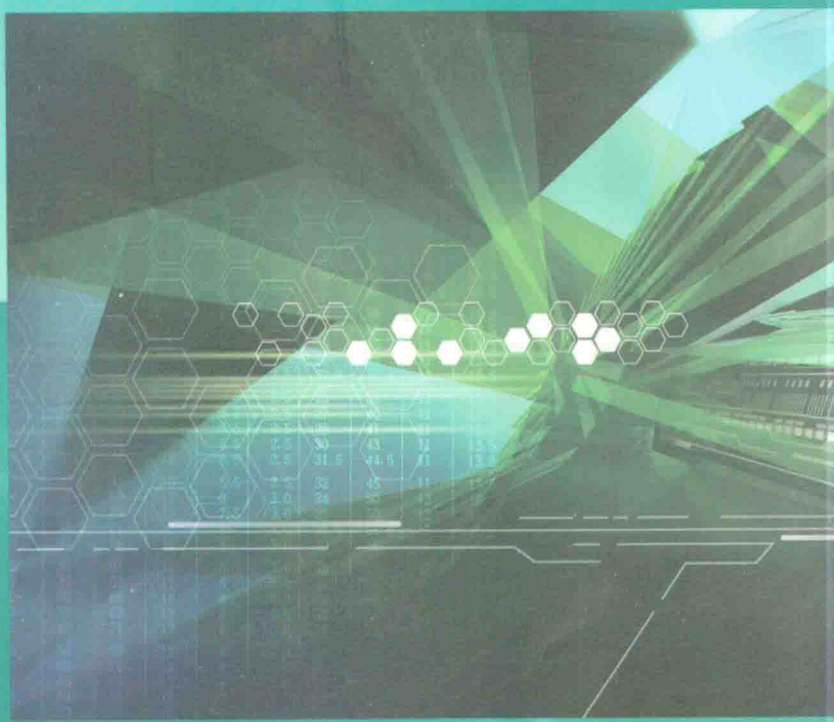
普通高等学校信息与计算科学专业系列丛书

信息论与编码理论

主　编　辛小龙
副主编　王　伟　付玉龙



高等教育出版社

普通高等学校信息与计算科学专业

信息论与编码理论

Xinxilun yu Bianma Lilun

主　编　辛小龙
副主编　王　伟　付玉龙

高等教育出版社·北京

内容提要

本书从信息科学的基本概念和基本方法入手，系统而又全面地介绍了信息论和编码理论的基本概念和理论，浅显易懂，简明易教。

全书共九章，内容包括绪论、离散信源及其信息度量、无失真信源编码、离散信道及其信道编码定理、限失真信源编码和率失真函数、连续信源的信息度量、线性码、循环码以及密码学基础。除第一章外各章后还附有习题。

本书可作为信息与计算科学、数学与应用数学、统计学、计算机科学、通信工程等专业的本科教材，也可作为相关专业研究生的学习参考书，还可供有关工程技术人员参考。

图书在版编目(CIP)数据

信息论与编码理论．辛小龙主编．--北京：高等教育出版社，2014.11

ISBN 978-7-04-041210-9

Ⅰ．①信… Ⅱ．①辛… Ⅲ．①信息论-高等学校-教材 ②信源编码-高等学校-教材 Ⅳ．①TN911.2

中国版本图书馆 CIP 数据核字(2014)第 232423 号

策划编辑 张长虹	责任编辑 张长虹	特约编辑 马兆海	封面设计 姜 磊
版式设计 王 莹	插图绘制 宗小梅	责任校对 孟 玲	责任印制 田 甜

出版发行	高等教育出版社	咨询电话	400-810-0598
社 址	北京市西城区德外大街 4 号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	三河市吉祥印务有限公司	网上订购	http://www.landraco.com
开 本	787mm ×960mm 1/16		http://www.landraco.com.cn
印 张	10	版 次	2014年11月第 1 版
字 数	180 千字	印 次	2014年11月第 1 次印刷
购书热线	010-58581118	定 价	17.00 元

物 料 号 41210-00

前　言

信息论是人们在长期通信工程的实践中，由通信技术与概率论、随机过程和数理统计相结合而逐步发展起来的一门科学。通常，人们公认信息论的奠基人是美国科学家香农(C. E. Shannon)，他在1948年发表了著名的论文《通信的数学理论》，为信息论奠定了理论基础。半个多世纪以来，以通信理论为核心的经典信息论，包括编码理论和密码理论，正以信息技术为物化手段，向高精尖方向迅猛发展，并以神奇般的力量把人类社会推向信息时代。随着信息理论的迅猛发展和信息概念的不断深化，信息论所涉及内容早已超越了狭义的通信范畴，进入到了信息科学这一更广阔、更新的领域。

本教材是在原《信息科学基础——信息论、编码及密码理论》的基础上形成的，作者结合了近几年的教学实践经验，对其做了部分修改和完善，增加了一些典型的习题，扩充了相应的内容。本教材着重介绍香农信息理论的基本理论、基本分析方法和主要结论，浅显易懂，简明易教。全书共九章。第一章通过一对特殊的数学模型——二元对称信源和二元对称信道，介绍信息理论的核心思想。第二章主要针对离散信源，介绍信息度量，以及信息熵、条件熵、联合熵、相对熵和互信息等概念，讨论它们的一些主要性质。第三章介绍无失真信源编码，应用渐近等分性证明了信源编码定理，同时介绍一些变长编码方法。第四章针对离散信源，介绍信道编码定理，引入信道容量的概念并讨论了信道容量的计算。第五章讨论限失真信源编码，引入率失真函数的概念，讨论了率失真函数的计算问题。第六章讨论连续信源的信息度量，介绍了连续信源的率失真函数、高斯信道及其信道容量等概念。第七章介绍线性码的相关理论。第八章介绍循环码的编码理论。第九章介绍密码学的基础，包括古典及现代密码学及RSA公钥密码体制等内容。另外，为了方便教师更好地组织教学，作者为本书的习题编写了解答，教师可以通过电子邮件 zhangchh@ hep. com. cn 索取。

本教材受西北大学重点课程项目资助，在此对西北大学教务处的支持表示感谢！教材由西北大学辛小龙教授任主编，西安石油大学王伟老师和西安电子科技大学付玉龙老师任副主编，在编写过程中参阅了大量的教科书、专著和文献，在此一并对作者表示感谢！

由于水平所限，书中不妥之处在所难免，敬请读者批评指正。

编　者

2014年8月21日于西安

目　录

第一章 绪 论

§1.1 序 言

当今的时代是一个信息的时代，信息处理技术的不断进步极大地影响了我们的生活，使我们的生活质量得到很大提高. 本课程将介绍信息科学的基础理论和基本方法，课程将基于一个通讯系统的抽象数学模型进行展开，课程的数学基础为概率论. 整个课程可分为信息基础理论和编码理论两部分组成.

“信息”是当代使用频率很高的一个概念，也是很难说清楚的一个概念. 据不完全统计，信息的定义有 100 多种，它们都从不同的侧面、不同的层次揭示了信息的某些特征和性质，但至今仍没有统一的、能为各界普遍认同的定义.

Claude Elwood Shannon(C. E. 香农)给出了信息的比较全面的定义. 他于 1916 年 4 月 30 日出生于美国密西根州. 1940 年于麻省理工学院获得博士学位. 1941 年 Shannon 加入了 AT&T 电话公司并在贝尔实验室工作到 1972 年. 1948 年发表了著名文章《通信的数学理论》，奠定了信息论的基础.

美国数学家 C. E. 香农的这篇奠基性论文，于 1948 年发表在《贝尔系统技术杂志》第 27 卷上. 原文共分五章. 香农在这篇论文中把通信的数学理论建立在概率论的基础上，把通信的基本问题归结为通信的一方能以一定的概率复现另一方发出的消息，并针对这一基本问题对信息作了定量描述. 香农在这篇论文中还精确地定义了信源、信道、信宿、编码、译码等概念，建立了通信系统的数学模型，并得出了信源编码定理和信道编码定理等重要结果. 这篇论文的发表标志一门新的学科——信息论的诞生.

著名的信息论和编码学者 Slepian 这样描述 Shannon 的这篇著作：

在 21 世纪或许再没有人的工作能比香农的这篇论文《通信的数学理论》，更深刻地影响人们对通信的理解.

这篇论文一经发表，立刻被世界范围的通信工程师和数学家共同采纳. 论文中的思想不断被传承、扩充，并被新的思想所补充. 这个分支已成长为科学领域的一个严密的、令人振奋的学科.

2001 年 2 月 26 日，Shannon 于马萨诸塞州辞世. 著名信息论和编码学者 Dr. Richard Blahut 在 Shannon 塑像落成典礼上这样评价 Shannon：“在我看来，两三百年之后，当人们回过头来看我们这个时代的时候，他们可能不会记得谁曾是美

国的总统,他们也不会记得谁曾是影星或摇滚歌星,但是仍然会知晓 Shannon 的名字,学校里仍然会讲授信息论."

§1.2 香农文章的序言

(Introduction of A Mathematical Theory of Communication)

近年来,一般通信理论的进展主要集中在信噪比的研究.各种调谐方法的最新研究成果,例如 PCM 和 PPM 这些为信噪比而变换带宽的方法,增强了人们对于一般通信系统理论研究的兴趣.在本文中,我们将扩充一般通信理论,引入许多新的因素,特别是信道的噪声效应,以及噪声对原始信息的统计结构和信宿特性的影响.

通信中的基本问题就是精确地或近似地在某一点再生另一点选择的信息.信息常常是有意义的;即:它们意味着联系一些现实或概念系统.通信的这些语义方面的特性与工程问题互不相干.它的意义体现在实际信息是从一组可能的信息中选择出来的一个.每一个系统必须被设计为对应任何一个可能的选择,而不能局限于实际被选的那一个,因为在设计时它是未知的.

如果在一个集合中,信息个数是有限的,则这个数或它的一个单调函数就可以看作是,当一个信息由这个集合所选出时的信息度量,这里假设全部的选择是等可能的.正如 Hartley 所指出的,最自然的选择是对数函数.尽管当我们考虑信息的统计特性的影响和信息的连续分布时,要对这个定义进行必要的扩充,然而我们在所有情况都将用本质上的对数测度.

对数测度更方便是基于以下原因:

1. 它在实践上更有用.工程中的重要参数,如时间、带宽、数字的分程传递等,趋向于随可能的数目线性改变.例如,给一组继电器加一个继电器,将会加倍可能情况的数目.它给这个数的以 2 为底的对数加 1,加倍时间大致得到可能信息数目的平方,或加倍其对数.

2. 它作为一个合适的测度,接近我们的直觉感观.如果我们直觉地用共同标准线性比较测量实体,这个测度有很强的相关性.例如,人们的感觉是,两张穿孔卡片与一张穿孔卡片相比,有两倍的信息储藏能力,并且两个通道与一个相比,有两倍的信息传输能力.

3. 它在数学方面更合适.许多极限运算在对数方式下很简单,但是在普通数字形式下却需要繁琐的重述.

对数底的选择对应于信息测量的单位.如果以 2 为底,产生的单位叫二进位数,或叫比特.一个拥有两个稳定位置的设备,像一个继电器或一个双稳态多谐振荡器,可以存储 1 比特信息.N 个如此的装置能存储 N 比特信息,因为可能的

状态的总数是 2^N，而 $\log_2 2^N = N$. 如果以 10 为底产生的单位成为十进制数，因为 $\log_2 M = \log_{10} M / \log_{10} 2 = 3.32 \log_{10} M$，所以，1 个十进制单位大约为 3.3 个比特.

一个台式计算机的数字轮有十个稳定的位置，从而拥有存储十进制数的能力. 在涉及积分和微分的分析工作中，以 e 为底是很方便的. 以此为底的信息结果将被叫做自然对数.

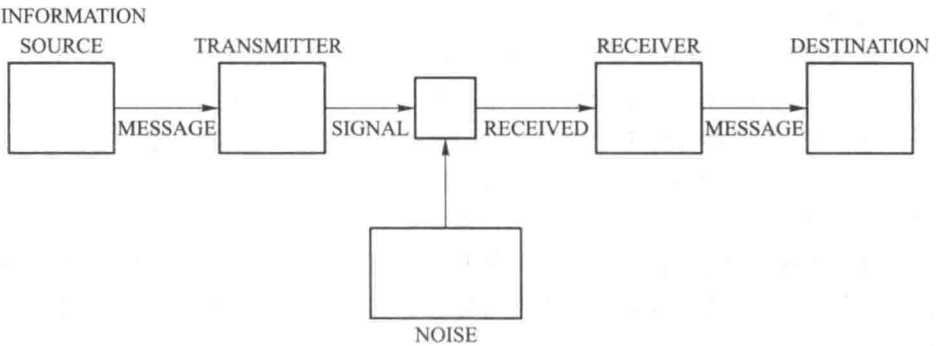


图 1.2.1 Schematic diagram of a general communication system

1. 信源. 它产生单个消息或一个消息列，这些消息可被送往终端. 信息有不同的类型：(a) 电传打字系统中的电报中的字母序列；(b) 无线电话中的单一时间函数 $f(t)$；(c) 在黑白电视机中时间和其余变量的函数——这里信息可以被看作两个空间坐标和时间的函数 $f(x,y,t)$，即在点 (x,y) 和时刻 t 的光强度；(d) 两个或更多的时间的函数，设为 $f(t), g(t), h(t)$—这是三维声音传播的情形；(e) 多变量的几个函数—在彩色电视机中信息由三个函数 $f(x,y,t), g(x,y,t), h(x,y,t)$ 表示，这些函数定义在一个三维闭区域，也可以将它们看作一个向量场的三个分量，黑白电视机也有类似情况；(f) 各种组合的情况也可能出现，例如在电视中有联合的音频信道.

2. 传送器(编码器). 它能对信息进行某种方式的操作，以产生适合信道传算的信号. 在电话中这种操作仅仅是将噪声压力转变为电流. 在电信技术中，我们将传送的信息变换为一系列的点、线和空间的编码操作. 在一个多元 PCM 系统中，不同的语音函数必须被取样、压缩、量化和编码，最终合适的结合，来构造信号. 声音传播机系统、电视和频率调制器是将复杂操作用于信息以获得信号的其余几个例子.

3. 信道. 它只是将信号从发送者传递到接受者的媒介. 它可以是电线、电缆、无线电波、光束等等. 在传送期间，或在某个终端，信号可能被噪音干扰. 这就是图 1.2.1 所标出的噪声源，被噪声源所干扰的发送信号产生了接收信号.

4. 接收器(译码器). 它通常完成发送者所作的操作的反操作，以从接收信号来重建消息.

5. 信宿. 它是对信息感兴趣的人或事物.

我们想要考虑通信系统的某些一般问题. 这首先需要描述由物理模型抽象出来的数学系统的相关性质. 我们可以将通信系统大致分为三类:离散的、连续的和混合的. 离散系统是指不论消息还是信号,都是一列离散符号. 一个典型的例子是:在电信技术中的消息就是字母和符号的序列,即点,莫尔斯电码和空间构成的序列.

连续型系统是指消息和信号都被看作为一个连续函数的系统,例如,无线电通信或电视机. 混合系统是指既有离散变量、又有连续变量的系统. 例如 PCM 语音传输系统. 我们将首先考虑离散系统,因为这种情形不仅能应用于通信系统,而且也能应用于计算机理论、电话机和其余领域的设计. 它也是研究连续和混合型系统的基础.

离散无噪系统:离散无噪信道电传和电报是用于传递信息的两个离散信道的简单例子. 一般的离散系统是指将有限符号集的一个选择的序列由一个地方传送到另一个地方.

假定符号 S_i 中的每一个都有 t_i 秒的逗留(不同的符号,时间可以不同). 不需要全部可能的序列都能被系统所传送;只有某些序列被允许,这些序列就是信道可能的信号.

第二章　离散信源及其信息度量

§2.1　自　信　息

什么是“信息”？信息一词在我国由来已久，泛指音讯和消息，但不同词典中对“信息”一词有不同的解释，且已不是作为科学名词或技术术语来定义，所以含义模糊又难于捉摸，但人人都感觉到信息的存在.

所谓“信源”，是指消息的来源，如信源输出的消息是以取值离散的符号形式出现，其不同符号数可以是有限个，或可列无限个，我们称其为离散信源. 如信源输出的消息的取值是连续的，可取不可列无限多个值，我们称其为连续信源. 本章首先讨论离散信源.

通常用随机变量 X 表示一个离散信源，X 的可能取值，即信源可能输出的不同符号用集合 $\mathcal{X}$ 表示. 如掷硬币这个随机试验看作一个信源的话，其取值集合为 $\mathcal{X}=\{$正，反$\}$，掷骰子的结果可用集合 $\mathcal{X}=\{1,2,3,4,5,6\}$ 表示.

当信源发出某个信号 $x_0\in\mathcal{X}$ 后，它提供了多少信息呢？即要解决信息的度量问题，我们把它称为 x_0 的自信息，记为 $I(x_0)$. 它是信号 x_0 的不确定性的一种度量，而 x_0 的不确定性即是它发生的可能性，这可以用 x_0 发生的概率 $p(x_0)$ 的大小来描述，因此 $I(x_0)$ 应当是概率 $p(x_0)$ 的一个函数. 那么 $I(x_0)$ 是一个什么样的函数呢？它应该满足哪些性质呢？

首先 $x\in\mathcal{X}$ 的概率越大，其发生的可能性越大，不确定性越小，$I(x_0)$ 应当越小，因此 $I(x_0)$ 应当是概率 $p(x)$ 的单调减函数；其次，如果信源连续独立地发出 2 个信号 x,y，即它们的联合分布 $P(x,y)=P(x)P(y)$，则 x,y 的自信息应是它们各自信息量之和，即 $I(x,y)=I(x)+I(y)$. 于是我们得到自信息应该满足的 5 条公理：

(i) 非负性：$I(x)\geqslant 0$；

(ii) 如果 $p(x)=0$，则 $I(x)\to\infty$；

(iii) 如果 $p(x)=1$，则 $I(x)=0$；

(iv) 严格单调性：如果 $p(x)>p(y)$，则 $I(x)<I(y)$；

(v) 如果 $p(x,y)=p(x)p(y)$，则 $I(x,y)=I(x)+I(y)$.

根据这 5 条公理我们可以得到自信息量表示的唯一性定理.

定理 2.1.1 若自信息 $I(x)$ 满足上述 5 条公理，则

$$I(x) = C\log\frac{1}{p(x)},$$

其中 C 为常数.

我们只需先证明以下引理.

引理 2.1.1 如果实函数 $f(x)$ $(1 \leqslant x < \infty)$ 满足以下条件：

(i) $f(x) \geqslant 0$；

(ii) $f(x)$ 是严格单调增函数，即 $x < y \Rightarrow f(x) < f(y)$；

(iii) $f(x \cdot y) = f(x) + f(y)$，

则 $f(x) = C\log x$.

证明 反复使用(iii)，对任意自然数 k，我们有

$$f(x^k) = f(x \cdot x^{k-1}) = f(x) + f(x^{k-1}) = \cdots = kf(x), \tag{2.1.1}$$

从而 $f(1) = 0$. 进而由于(i)和(ii)，对于任意 $x > 1$，$f(x) > 0$，对于任意大于 1 的 x, y 与任意自然数 k，总可以找到非负整数 n，使

$$0 < y^n \leqslant x^k < y^{n+1},$$

取对数并除以 $k\log y$ 得

$$\frac{n}{k} \leqslant \frac{\log x}{\log y} < \frac{n+1}{k}, \tag{2.1.2}$$

另一方面，由于(2.1.1)及条件(ii)可得

$$nf(y) \leqslant kf(x) < (n+1)f(y)$$

或

$$\frac{n}{k} \leqslant \frac{f(x)}{f(y)} < \frac{n+1}{k}, \tag{2.1.3}$$

由于(2.1.2)，(2.1.3)我们有

$$\left|\frac{f(x)}{f(y)} - \frac{\log x}{\log y}\right| \leqslant \frac{1}{k},$$

当 $k \to \infty$ 时，

$$\frac{f(x)}{f(y)} = \frac{\log x}{\log y}.$$

因此

$$\frac{f(x)}{\log x}=\frac{f(y)}{\log y}=C$$

或

$$f(x)=C\log x.$$

为证明定理 2.1.1，只需对 $f\left(\frac{1}{p(x)}\right)=I(p(x))$ 应用引理即可.

定义 2.1.1 设 $x\in X$ 有概率 $p(x)$，则 x 的自信息定义为 $I(x)=\log\frac{1}{p(x)}$.

自信息有两个含义：

1. 当事件发生前，表示该事件发生的不确定性；
2. 当事件发生后，表示该事件所提供的信息量.

自信息量的单位取决于对数所取的底，若以 2 为底，单位为比特，以 e 为底，单位为奈特(nat)，以 10 为底，单位为哈特(hartley)，通常取比特(bit)为单位.

§2.2 熵、联合熵、条件熵

自信息反映了信源发出的每个信号的信息量，那么对整个信源来说，其每个信号的平均信息量是多少？我们把这个信息量称为熵. 换句话说，如果用随机变量代表一个信源，则熵就是它的平均不确定性的度量.

设 X 是取值于离散字母集 X 的随机变量(X 也称状态集)，其概率分布函数为 $p(x)=P_r\{X=x\}$，$x\in X$，我们用 $p(x)$ 和 $p(y)$ 分别表示随机变量 X 和 Y 的概率分布函数，有时为明确区别起见，用 $p_X(x)$ 和 $p_Y(y)$ 表示，用 $X\sim p(x)$ 表示 X 服从分布 $p(x)$.

定义 2.2.1 离散随机变量 X 的熵定义为

$$H(X)=-\sum_{x\in X}p(x)\log p(x).$$

我们也用 $H(p)$ 表示这个熵，有时也称它为概率分布 p 的熵，其中对数以 2 为底时，熵的单位为比特(bit)，若对数以 e 为底，则熵的单位为奈特(nat)，若对数以 10 为底，则熵的单位为哈特(hartley). 注意熵只是概率分布 p 的函数，与 X 取什么值并无关系，用 E 表示数学期望，E_p 表示关于分布 p 的数学期望，即

$$E_p[g(X)]=\sum_{x\in X}g(x)p(x),$$

则熵可表示为**随机变量** $\log\frac{1}{p(X)}$ **的数学期望**，即

$$H(X)=E_p\left[\log\frac{1}{p(X)}\right].$$

可见熵是**自信息的概率加权平均值**. 熵有以下一些性质.

引理 2.2.1 $H(X)\geqslant 0$,且等号成立的充要条件是 X 有退化分布.

证明 因 $0\leqslant p(x)\leqslant 1$,从而 $-p(x)\log p(x)\geqslant 0$,由 $H(X)$ 定义即得 $H(X)\geqslant 0$,其中等号成立的充要条件为 $p(x)=0$ 或 $p(x)=1$. 由概率分布的定义 $p(x)\geqslant 0$,$\sum\limits_x p(x)=1$ 知,只有一个 x_0 使 $p(x_0)=1$,而对其他 $x\in\mathcal{X}-\{x_0\}$,$p(x)=0$,即 p 为退化分布.

定义 2.2.2 设一对随机变量 (X,Y) 的联合分布为

$$p(x,y)=P_r\{X=x,Y=y\},x\in\mathcal{X},y\in Y,$$

则定义 (X,Y) 的联合熵 $H(X,Y)$ 为

$$H(X,Y)=-\sum_{x\in\mathcal{X}}\sum_{y\in Y}p(x,y)\log p(x,y),$$

或写成数学期望形式:

$$H(X,Y)=-E[\log p(X,Y)].$$

联合熵的概念可进一步推广到 n 维随机变量.

定义 2.2.3 设 n 维随机变量对 $(X_1,X_2,\cdots,X_n)$ 的联合分布为

$$p(x_1,x_2,\cdots,x_n)=P_r\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\},x_1\in\mathcal{X}_1,x_2\in\mathcal{X}_2,\cdots,x_n\in\mathcal{X}_n,$$

则定义联合熵为

$$H(X_1,X_2,\cdots,X_n)=-\sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2}\cdots\sum_{x_n\in\mathcal{X}_n}p(x_1,x_2,\cdots,x_n)\log p(x_1,x_2,\cdots,x_n).$$

以下我们再定义给定一个随机变量条件下另一个随机变量的条件熵.

定义 2.2.4 设随机变量对 (X,Y) 有联合分布 $p(x,y)$,用

$$p(y\mid x)=P_r\{Y=y\mid X=x\},x\in\mathcal{X},y\in Y$$

表示条件概率分布,则给定 $X=x$ 条件下 Y 的熵定义为

$$H(Y\mid X=x)=-\sum_{y\in Y}p(y\mid x)\log p(y\mid x).$$

而给定随机变量 X 条件下 Y 的熵记为 $H(Y\mid X)$,它是 $H(Y\mid X=x)$ 关于 X 的平均值,即

$$H(Y\mid X)=\sum_{x\in\mathcal{X}}p(x)H(Y\mid X=x)$$

$$= -\sum_{x\in X} p(x)\sum_{y\in Y} p(y\mid x)\log p(y\mid x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x)$$

$$= -E[\log p(Y\mid X)].$$

以下的定理给出了随机变量对联合熵与单个随机变量的熵和两变量的条件熵的关系.

定理 2.2.1(链法则) $H(X,Y)=H(X)+H(Y|X)$.

证明 $H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y\mid x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x)$$

$$= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x)$$

$$= H(X) + H(Y\mid X).$$

注 1 由于 X 和 Y 的对称性,易知

$$H(X,Y)=H(Y)+H(X\mid Y)=H(Y,X).$$

注 2 类似于 $H(Y\mid X)$,我们可定义 $H(Y,Z\mid X)$ 如下:

$$H(Y,Z\mid X) = -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log p(y,z\mid x).$$

定义 $H(Z\mid Y,X)$ 如下:

$$H(Z\mid Y,X) = -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log p(z\mid x,y).$$

还可以进一步推广到任意多个随机变量的情形,即类似可定义

$$H(Y\mid X_1,X_2,\cdots,X_n)\text{等}.$$

链法则也可进一步推广到多维情形,为简化记号,今后记

$$X^n=(X_1,X_2,\cdots,X_n),x^n=(x_1,x_2,\cdots,x_n).$$

定理 2.2.2(熵的链法则) 设 $X_1,X_2,\cdots,X_n$ 的联合分布为 $p(x_1,x_2,\cdots,x_n)$,则

$$H(X_1,X_2,\cdots,X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1},\cdots,X_1).$$

证明 只需重复应用定理 2.2.1 中两个随机变量熵的链法则.

$$\begin{aligned}
H(X_1,X_2) &= H(X_1) + H(X_2 \mid X_1),\\
H(X_1,X_2,X_3) &= H(X_1) + H(X_2,X_3 \mid X_1)\\
&= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2,X_1),\\
&\cdots\cdots\\
H(X^n) &= H(X_1,X_2,\cdots,X_n)\\
&= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_{n-1},\cdots,X_1)\\
&= \sum_{i=1}^{n} H(X_i \mid X_{i-1},\cdots,X_1).
\end{aligned}$$

注 也可利用

$$\begin{aligned}
p(x^n) &= p(x_1,x_2,\cdots,x_n)\\
&= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2,x_1)\cdots p(x_n \mid x_{n-1},\cdots,x_1)
\end{aligned}$$

两边取负对数 $-\log$ 后再对联合分布 $p(x_1,x_2,\cdots,x_n)$ 取期望.

§2.3 相对熵和互信息

随机变量的熵是随机变量不确定性的度量,它是描述一个随机变量平均所需信息量的度量.本节中我们介绍两个相关的概念:相对熵和互信息.

相对熵是两个概率分布差异的一种度量,而互信息则是一个随机变量包含的关于另一个随机变量的信息量的度量.

定义 2.3.1 定义在同一个字母集合 $\mathcal{X}$ 上的两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵定义为

$$\begin{aligned}
D(p \parallel q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\\
&= E_p\left[\log\frac{p(x)}{q(x)}\right].
\end{aligned}$$

在上述定义中,按惯例规定 $0\cdot\log\dfrac{0}{q}=0$,$p\cdot\log\dfrac{p}{0}=\infty$,一般地说,$D(p\parallel q)\neq D(q\parallel p)$,以下的定理表明,相对熵 $D(p\parallel q)$ 总是非负的.