中国矿业大学教材建设工程资助教材

Xinxilun YU Bianma Jianming Jiaocheng

信息论与编码简明教程

陈兴同　主编

中国矿业大学出版社
China University of Mining and Technology Press

信息论与编码简明教程

主　编　陈兴同

中国矿业大学出版社

内 容 提 要

本教材涵盖了经典信息论所研究的内容，重点对离散信源、离散信道的信息处理进行分析研究，包括信息度量问题，编码的存在性问题，以及常见编码的基本方法. 对连续信源、连续信道的信息处理只作简单介绍，比较适合作为信息与计算科学专业以及相关专业的信息论基础教材.

图书在版编目(CIP)数据

信息论与编码简明教程/陈兴同主编. —徐州：
中国矿业大学出版社，2016.7

ISBN 978-7-5646-2941-0

Ⅰ. ①信… Ⅱ. ①陈… 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(2015)第 297000 号

书　　名　信息论与编码简明教程
主　　编　陈兴同
责任编辑　张　岩　耿东锋
出版发行　中国矿业大学出版社有限责任公司
（江苏省徐州市解放南路　邮编 221008）
营销热线　(0516)83885307　83884995
出版服务　(0516)83885767　83884920
网　　址　http://www.cumtp.com　**E-mail**:cumtpvip@cumtp.com
印　　刷　徐州中矿大印发科技有限公司
开　　本　787×1092　1/16　**印张** 10.75　**字数** 268 千字
版次印次　2016 年 7 月第 1 版　2016 年 7 月第 1 次印刷
定　　价　28.00 元

前 言

“信息”这个词现在已经成为当今社会上最流行的词汇之一，也是当前热门的研究对象.我们避开绕人的定义，可以将“信息”看成是物质属性的表示，获得了信息，人们就可以了解或掌握物质的运动状态及运动方式，消除对物质运动进行判断时产生的不确定性.信息包含一定的意义，得用某种方式来度量才能知道获得信息的多少，也得用某种方式来表示才能进行交流、传播、保存.在“信息论”这门课程中我们将舍弃信息的具体意义，只学习信息在通信过程中的度量与表示问题.

什么是通信？顾名思义，“通信”就是在某处准确地或尽可能准确地恢复另一处所发送的消息，从中获得有用的信息，以实现信息的传播或储存.通信过程可以描述为：从信源发出的消息(其中包含有用的信息)经过编码后转变成信号通过信道传送给信宿即接收者，信宿通过译码恢复成信源所发出的消息，从而获得所需信息.这样的通信过程可用图1来描述.通信过程中的信息论所讨论的内容可以分为三个层次：最高层(第一层)：信息；中间层(第二层)：消息——用于表示信息，是信息的载体；最低层(第三层)：信号——消息的载体，仅存在于信道或存储介质中.

图1 通信系统简化图

衡量通信系统的性能有三个常用指标：

传输有效性：单位时间内传送信息的多少，涉及数据压缩问题.

传输可靠性：传输过程中造成的差错尽可能少，涉及数据纠错问题.

传输安全性：传输的信息在通信过程中不应泄露，涉及数据加密问题.

这三个方面涉及的内容已经成为信息论研究的基本内容，包括信息度量、信道容量、压缩编码、传输编码、信号分析、保密通信、密码分析等.

“信息论”是从1940年后，在仙农(C. E. Shannon)工作的基础上，根据通信的实践与发展逐渐形成的一门应用数学学科，它专门研究信息的有效处理、可

靠传输中规律性.可划分成经典信息论即仙农信息论与广义信息论.仙农信息论基于概率模型,成功地定义了信息度量,很好地解决了信息处理过程中压缩编码问题、信息传输中的纠错编码问题.仙农认为任何信息源例如文本、图像、影视、声音等数据都包含着可以量化的信息,可以比较它们包含信息的多少.仙农是通信工程师,他在第二次世界大战期间的通信工程实践中,将遇到的问题上升到理论的高度,发表了有关"通信中的数学理论"等论文,开创了信息研究的新纪元,不仅解决了一些通信工程中的重大理论问题,同时新的理论研究成果又对通信工程起到了巨大的指导作用.现在许多新的信息技术和通信技术仍然是以仙农的经典信息论原理为基础发展起来的,所以我们要好好学习这门课.

从20世纪四五十年代开始以苏联为首的一批概率论专家专注于信息论的基础理论研究,而以美国为首的一批科学家及工程师专注于信息的有效处理与可靠传输问题的研究,两方面互相推动,涌现了许多好的理论结果与应用成果.

经典信息论所研究的内容包括:信息度量、信源编码方法与存在定理、信道编码方法与存在定理、纠错编码、限失真数据压缩编码方法与存在定理等.通过建立通信过程中的数学模型,以概率论、随机过程等数学工具进行研究.

根据信息与计算科学专业的基本教学要求,结合作者多年的教学实践,编写了这本教材.基本想法是:第一,覆盖仙农信息论的基本内容,主要是仙农的三大编码存在性定理,弱化了证明过程中使用的典型序列法及联合典型序列法;第二,在每个编码定理的原则下,重点介绍几种基本编码算法,并讨论编码算法的性能指标,不追求很全面的编码算法,更多的编码算法及特性可以留给同学们去搜集自学;第三,增加同学们对信息处理的感性认识,结合比较流行的MATLAB软件,设计了部分内容的实验.

由于时间仓促,作者水平有限,不足或谬误之处在所难免,请读者朋友批评指正.

作者

2015年8月

目　录

1 随机变量及其信息度量

信息论中采用的数学模型基本上都是概率模型，而离散型概率又是现代数字通信中常用的概率模型. 本章重点讨论离散随机变量的信息度量，最后简单介绍连续型随机变量的信息度量.

1.1 离散随机变量及其分布

离散型随机变量是指取值空间为可数离散集合(无限集合或有限集合)的随机变量.

1.1.1 一维分布律

设离散型随机变量 X 取值空间为 $\mathscr{X}=\{x_1,x_2,x_3,\cdots\}$，它的分布律是指每个取值的概率

$$p_i = p(x_i) = P\{X = x_i\}, i = 1,2,\cdots, \tag{1-1}$$

并且符合要求

$$p_i \geqslant 0, i = 1,2,\cdots, \sum_i p_i = 1. \tag{1-2}$$

如果 $\mathscr{X}$ 是一个有限集 $\mathscr{X}=\{x_1,x_2,x_3,\cdots,x_N\}$，则它的分布律也可以用两行表格或矩阵来表示，第一行表示取值，第二行表示取每个值的概率.

$$\begin{array}{c|cccc} X & x_1 & x_2 & \cdots & x_N \\ \hline p_i & p_1 & p_2 & \cdots & p_N \end{array} \quad \text{或} \quad X \sim \begin{bmatrix} x_1 & x_2 & \cdots & x_M \\ p_1 & p_2 & \cdots & p_N \end{bmatrix}. \tag{1-3}$$

另外，分布律(1-1)也定义了一元离散函数. 它类似于连续型随机变量概率密度函数，不妨称为离散型随机变量 X 的概率函数，记作：

$$X \sim p(x), x \in \mathscr{X}. \tag{1-4}$$

多个随机变量的概率函数将用下标来区分，比如 $p_X(x)$，$p_Y(x)$.

描述离散随机变量分布情况也可以用分布函数，它的定义为：

$$F(x) = P\{X \leqslant x\} = \sum_{x_i \leqslant x} p_i, x \in \mathbb{R}. \tag{1-5}$$

其实就是离散随机变量 X 取值不大于 x 的概率，它是一元阶梯形函数，分断点 x_i 处的跃度就是概率 p_i.

最后，如果行向量 $\boldsymbol{p}=(p_1,p_2\cdots)$ 满足条件式(1-2)，就称它是一个概率分布向量.

1.1.2 联合分布律

设离散型随机变量 X,Y 的取值空间分别为 $\mathscr{X}=\{x_1,x_2,x_3,\cdots\}$，$\mathscr{Y}=\{y_1,y_2,y_3,\cdots\}$，它

们的联合分布律是

$$p_{ij} = p(x_i, y_j) = P\{X = x_i, Y = y_j\}, x_i \in \mathscr{X}, y_j \in \mathscr{Y}. \tag{1-6}$$

如果 X,Y 的取值空间 $\mathscr{X},\mathscr{Y}$ 是有限集，则联合分布律也可以用二维表或矩阵

$X\backslash Y$	y_1	y_2	$\cdots$	y_M
x_1	p_{11}	p_{12}	$\cdots$	p_{1M}
x_2	p_{21}	p_{22}	$\cdots$	p_{2M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_N	p_{N1}	p_{N2}	$\cdots$	p_{NM}

或 $$(X,Y)\sim\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NM} \end{pmatrix} \tag{1-7}$$

表示. 另外，分布律(1-6)也定义了一个二元离散函数

$$p(x,y)=P\{X=x,Y=y\},(x,y)\in\mathscr{X}\times\mathscr{Y},$$

不妨称它为随机向量(X,Y)的联合概率函数，记作

$$(X,Y) \sim p(x,y), x \in \mathscr{X}, y \in \mathscr{Y}. \tag{1-8}$$

当有多个联合分布时，用下标来区分，比如：$p_{X,Y}(x,y)$，$p_{U,V}(x,y)$.

也可以用二元分布函数描述联合分布：

$$F(x,y) = P\{X \leqslant x, Y \leqslant y\} = \sum_{x_i \leqslant x, y_i \leqslant y} p_{ij}, (x,y) \in \mathbb{R}^2.$$

它是一个二元阶梯形函数，每个阶梯处的跃度就是概率 p_{ij}.

对于 n 维离散随机向量$(X_1,X_2,\cdots,X_n)$的联合分布，可以用 n 维联合分布律或 n 元分布函数来描述.

$$p(x_1,x_2,\cdots,x_n)=P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\},x_i\in\mathscr{X}_i,i=1,2,\cdots,n$$

$$F(x_1,x_2,\cdots,x_n)=P\{X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n\},x_i\in\mathbb{R},i=1,2,\cdots,n,$$

其中 $\mathscr{X}_i$ 是第 i 个随机变量的取值空间，当然对 n 个有限离散随机变量的联合分布律也可以用 n 维数组来描述.

1.1.3 边缘分布律

设两个离散型随机变量 X,Y 的联合分布律为式(1-8)，则可以求出 X,Y 各自的分布律 $p_X(x)$，$p_Y(y)$，称为边缘分布律，它们的求法如下：

$$p_X(x) = \sum_y p(x,y), p_Y(y) = \sum_x p(x,y). \tag{1-9}$$

对于 n 维离散随机向量$(X_1,X_2,\cdots,X_n)$，它的每一个分量 X_i，任意两个分量 X_i,X_j，任意三个分量 X_i,X_j,X_k，任意 $n-1$ 个分量 $X_{i_1},X_{i_2},\cdots,X_{i_{n-1}}$ 也分别是随机变量或随机向量，也都有确定的分布律，称为这个 n 维随机向量的边缘分布. 它们可以从 n 维联合分布律获得：

$$p_{X_i}(x_i) = \sum_{x_j \in \mathscr{X}_j, j \neq i} p(x_1,x_2,\cdots,x_n).$$

$$p_{X_i,X_j}(x_i,x_j) = \sum_{\substack{x_l \in \mathscr{X}_l, l \neq i \\ x_m \in \mathscr{X}_m, m \neq j}} p(x_1,x_2,\cdots,x_n).$$

$$F_{X_1}(x_1)=F(x_1,+\infty,\cdots,+\infty).$$

$$F_{X_i,X_j}(x_i,x_j)=F(+\infty,\cdots,+\infty,x_i,+\infty,\cdots,+\infty,x_j,+\infty,\cdots,+\infty).$$

1.1.4 条件分布律

若两个离散型随机变量 X,Y 有联合分布律(1-8),并且边缘分布 $p_X(x)>0$,则可以定义条件概率

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

如果取 $X=x_i,Y=y_j$,则将条件概率记为 $p(y_j|x_i)$或 $p_{j|i}$. 每给随机变量 X 的一个取值 x_i 就能得到一个关于 Y 的条件分布律

$$p(y_1\mid x_i),p(y_2\mid x_i),p(y_3\mid x_i),\cdots. \tag{1-10}$$

如果 X,Y 的取值空间 $\mathscr{X},\mathscr{Y}$ 是有限集,就可以将每个条件分布律作为行构成矩阵

$$\begin{bmatrix} p(y_1|x_1) & p(y_2|x_1) & \cdots & p(y_M|x_1)\\ p(y_1|x_2) & p(y_2|x_2) & \cdots & p(y_M|x_2)\\ \vdots & \vdots & & \vdots\\ p(y_1|x_N) & p(y_2|x_N) & \cdots & (y_M|x_N) \end{bmatrix} \text{或} \begin{bmatrix} p_{1|1} & p_{2|1} & \cdots & p_{M|1}\\ p_{1|2} & p_{2|2} & \cdots & p_{M|2}\\ \vdots & \vdots & \ddots & \vdots\\ p_{1|N} & p_{2|N} & \cdots & p_{M|N} \end{bmatrix} \tag{1-11}$$

称为条件分布矩阵,它可以作为信道传输特性的概率模型,同时也定义了一个在集合 $\mathscr{X}\times\mathscr{Y}$ 上的二元离散函数 $p(y|x)$,称为条件概率函数.

也可以定义用随机向量作为条件的条件概率.

$$\begin{aligned} p(y_j\mid x_1,x_2,\cdots,x_n)&=P\{Y=y_i\mid X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}\\ &=\frac{P\{X_1=x_1,\cdots,X_n=x_n,Y=y_j\}}{P\{X_1=x_1,\cdots,X_n=x_n\}}\\ &=\frac{p(x_1,x_2,\cdots,x_n,y_i)}{p(x_1,x_2,\cdots,x_n)}, \end{aligned}$$

其中 $y_j\in\mathscr{Y}_j,x_i\in\mathscr{X}_i,i=1,2,\cdots,n.$

另外还可以定义更一般的条件概率,比如:

$$\begin{aligned} p(y_1,y_2,\cdots,y_m\mid x_1,x_2,\cdots,x_n)&=P\{Y_1=y_1,\cdots,Y_m=y_m\mid X_1=x_1,\cdots,X_n=x_n\}\\ &=\frac{P\{Y_1=y_1,\cdots,Y_m=y_m,X_1=x_1,\cdots,X_n=x_n\}}{P\{X_1=x_1,\cdots,X_n=x_n\}}\\ &=\frac{p(x_1,\cdots,x_n,y_1,\cdots,y_m)}{p(x_1,x_2,\cdots,x_n)}, \end{aligned}$$

其中 $y_j\in\mathscr{Y}_j,x_i\in\mathscr{X}_i,i=1,2,\cdots,N;j=1,2,\cdots,M.$

1.1.5 数学期望

离散型随机变量的数学期望定义为所取值的概率平均值. 这里同时给出随机变量函数 $Y=f(X)$及 $Z=f(X_1,X_2,\cdots,X_n)$的数学期望公式:

$$E(X)=\sum_{x\in\mathscr{X}}xp(x),$$

$$E(Y)=E[f(X)]=\sum_{x\in\mathscr{X}}f(x)p(x),$$

$$E(Z)=E[f(X_1,X_2,\cdots,X_n)]=\sum_{x_1\in\mathscr{X}_1}\cdots\sum_{x_n\in\mathscr{X}_n}f(x_1,\cdots,x_n)p(x_1,\cdots,x_n).$$

6 条件数学期望

条件数学期望是指条件分布的数学期望. 在事件$\{X=x_i\}$发生的条件下随机变量Y具有条件分布(1-10),当然可以求数学期望:

$$E(Y \mid x_i) = E(Y \mid X = x_i) = \sum_{y \in \mathcal{Y}} yp(y \mid x_i), x_i \in \mathcal{X}.$$

由于条件分布矩阵(1-11)每一行都是一个条件分布,故每一行都可以求数学期望,可以求到一系列条件期望值$E(Y|x_1), E(Y|x_2), \cdots$,因此条件数学期望$E(Y|x_i)$可以看作是在取值空间$\mathcal{X}$上的函数$E(Y|x), x \in \mathcal{X}$.

类似地可以定义其他条件数学期望,比如在更多条件下的条件数学期望:

$$E(Y \mid x_1, x_2, \cdots, x_n) = E(Y \mid X_1 = x_1, \cdots, X_n = x_n) = \sum_{y \in \mathcal{Y}} yp(y \mid x_1, x_2, \cdots, x_n).$$

1.2 凸 函 数

为了研究信息熵的性质,这节简单介绍凸函数概念及一些概率不等式.

1.2.1 一元凸函数

一元凸或凹函数是指在某区间上图像向上或向下弯曲的一元函数.

定义 1.2.1 设$y=f(x), x \in (a,b) \subset \mathbb{R}$是一元函数.

(1) 如果$\forall x_1, x_2 \in (a,b), \lambda \in [0,1]$总有

$$f(\lambda x_1 + (1-\lambda)x_2) \leqslant \lambda f(x_1) + (1-\lambda) f(x_2),$$

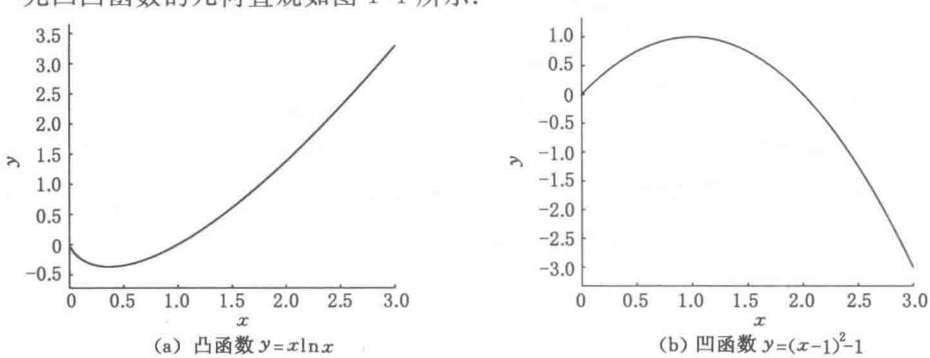
则称$f(x)$为区间(a,b)上的凸函数;如果$-f(x)$为凸函数,则称$f(x)$为区间(a,b)上的凹函数.

(2) 如果$\forall x_1, x_2 \in (a,b), x_1 \neq x_2, \lambda \in (0,1)$总有

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2),$$

则称$f(x)$为区间(a,b)上的严格凸函数;如果$-f(x)$为严格凸函数,则称$f(x)$为区间(a,b)上的严格凹函数.

一元凹凸函数的几何直观如图 1-1 所示.



(a) 凸函数$y=x\ln x$　　(b) 凹函数$y=(x-1)^2-1$

图 1-1 凸函数图像直观

常见的凸函数有：

$$f(x)=x^2, f(x)=e^x, f(x)=x\log x(x>0).$$

注 1.2.1 如果 $f(x)=ax+b$，则它既是凸函数也是凹函数，或者说是一个不凸不凹函数，可以根据实际需要规定它的凹凸性.

定理 1.2.1 凸(凹)函数的判定定理：

(1) 设函数 $f(x)$在区间(a,b)上二阶连续可微，如果对任何 $x\in(a,b)$有 $f''(x)\geqslant 0$，则 $f(x)$是区间(a,b)上的凸函数.

(2) 设函数 $f(x)$在区间(a,b)上二阶连续可微，如果对任何 $x\in(a,b)$有 $f''(x)>0$，则 $f(x)$是区间(a,b)上的严格凸函数.

(3) 设函数 $f(x)$在区间(a,b)上二阶连续可微，如果对任何 $x\in(a,b)$有 $f''(x)\leqslant 0$，则 $f(x)$是区间(a,b)上的凹函数.

(4) 设函数 $f(x)$在区间(a,b)上二阶连续可微，如果对任何 $x\in(a,b)$有 $f''(x)<0$，则 $f(x)$是区间(a,b)上的严格凹函数.

1.2.2 多元凸函数

定义 1.2.2 设 $\mathcal{G}$ 为 n 维欧氏空间$\mathbb{R}^n$中的一个非空集合，如果对于 $\mathcal{G}$ 中任意两个点 $\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}$，连接它们的线段仍属于 $\mathcal{G}$，即 $\forall\lambda\in[0,1]$有 $\lambda\boldsymbol{x}^{(1)}+(1-\lambda)\boldsymbol{x}^{(2)}\in\mathcal{G}$，则称 $\mathcal{G}$ 为 $\mathbb{R}^n$ 中的一个凸集合. 如果它又是闭集，则称为闭凸集.

平面上的常见的凸集合有直线、三角形区域、圆域等. 而 n 维欧氏空间中常见的凸集合有

(1) 超平面：$H=\{\boldsymbol{x}\in\mathbb{R}^n\mid \boldsymbol{p}^{\mathrm{T}}\boldsymbol{x}=\alpha\}$，$\boldsymbol{p}\in\mathbb{R}^n$，$\alpha\in\mathbb{R}$.

(2) 半空间：$H^-=\{\boldsymbol{x}\in\mathbb{R}^n\mid \boldsymbol{p}^{\mathrm{T}}\boldsymbol{x}\leqslant\alpha\}$，$\boldsymbol{p}\in\mathbb{R}^n$，$\alpha\in\mathbb{R}$.

(3) 射线：$L=\{\boldsymbol{x}\in\mathbb{R}^n\mid \boldsymbol{x}=\boldsymbol{x}^{(0)}+\lambda\boldsymbol{d},\lambda\geqslant 0\}$，$\boldsymbol{x}^{(0)},\boldsymbol{d}\in\mathbb{R}^n$.

定义 1.2.3 设 $\mathcal{G}$ 为 $\mathbb{R}^n$ 中的非空凸集合，$f(\boldsymbol{x})=f(x_1,\cdots,x_n)$是定义在 $\mathcal{G}$ 上的 n 元实函数.

(1) 若对 $\forall\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}\in\mathcal{G}$，$\forall\lambda\in[0,1]$有 $f(\lambda\boldsymbol{x}^{(1)}+(1-\lambda)\boldsymbol{x}^{(2)})\leqslant\lambda f(\boldsymbol{x}^{(1)})+(1-\lambda)f(\boldsymbol{x}^{(2)})$，则称 $f(\boldsymbol{x})$为 $\mathcal{G}$ 上的 n 元凸函数.

(2) 若 $-f(\boldsymbol{x})$为 $\mathcal{G}$ 上的 n 元凸函数，则称 $f(\boldsymbol{x})$为 $\mathcal{G}$ 上的 n 元凹函数.

(3) 若对 $\forall\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}\in\mathcal{G},\boldsymbol{x}^{(1)}\neq\boldsymbol{x}^{(2)}$，$\forall\lambda\in(0,1)$有 $f(\lambda\boldsymbol{x}^{(1)}+(1-\lambda)\boldsymbol{x}^{(2)})<\lambda f(\boldsymbol{x}^{(1)})+(1-\lambda)f(\boldsymbol{x}^{(2)})$，则称 $f(\boldsymbol{x})$为 $\mathcal{G}$ 上的 n 元严格凸函数.

(4) 若 $-f(\boldsymbol{x})$为 $\mathcal{G}$ 上的 n 元严格凸函数，则称 $f(\boldsymbol{x})$为 $\mathcal{G}$ 上的 n 元严格凹函数.

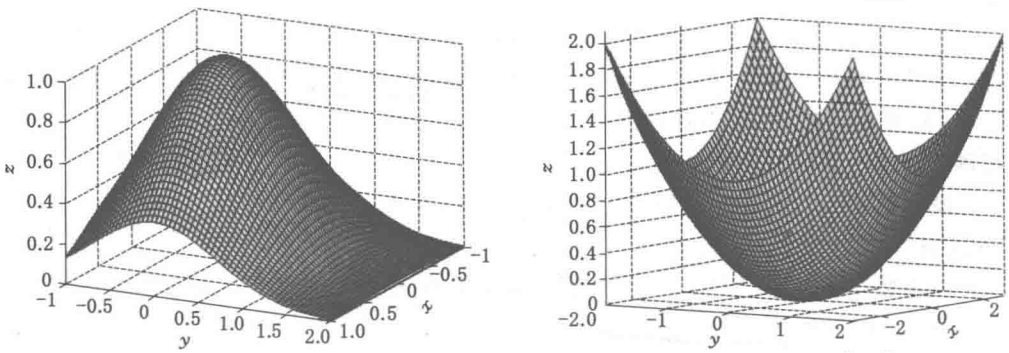
注 1.2.2 超平面函数 $f(\boldsymbol{x})=a_1x_1+a_2x_2+\cdots+a_nx_n$ 是不凸不凹函数，可根据需要规定它的凹凸性.

二元凸函数的几何直观见图 1-2.

凸(凹)函数一般和区域有关，在不同区域上，同一函数的凹凸性也可能不同，如图 1-3 所示.

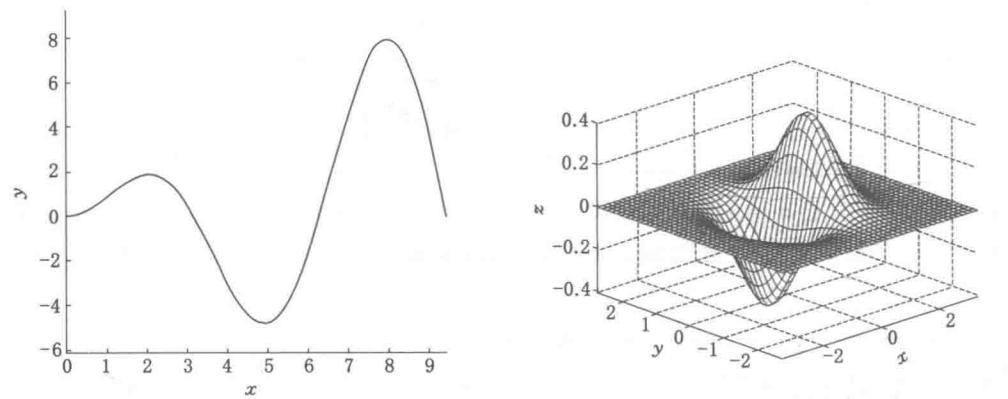
1.2.3 常用不等式

利用凸函数可以建立许多重要的不等式，而各种形式的琴生(Jensen)不等式就是常用



(a) 凹函数 $z=e^{-x^2-y^2}$　　(b) 凸函数 $z=\frac{x^2}{4}+\frac{y^2}{9}$

图 1-2　二维凸函数图像直观



(a) 函数 $y=x\sin x$　　(b) 函数 $z=x\cdot e^{-x^2-y^2}$

图 1-3　非凸函数图像直观

工具.

定理 1.2.2(Jensen 不等式)

(1) 设 $f(x)$是区间(a,b)上的凸函数，$x_1,x_2,\cdots,x_n\in(a,b)$，并且 $\lambda_1,\lambda_2,\cdots,\lambda_n$构成概率分布律，那么

$$f(\lambda_1x_1+\lambda_2x_2+\cdots+\lambda_nx_n)\leqslant\lambda_1f(x_1)+\lambda_2f(x_2)+\cdots+\lambda_nf(x_n).$$

(2) 设有限离散型随机变量 X 的取值空间是 $\mathscr{X}$，如果 $f(x)$是区间 $D\supseteq\mathscr{X}$ 上的一个凸函数，则 $f[E(X)]\leqslant E[f(X)]$.

(3) 设 $f(x)$是区间(a,b)上的严格凸(凹)函数，$x_1,x_2,\cdots,x_n\in(a,b)$，并且 $\lambda_1,\lambda_2,\cdots,\lambda_n$构成概率分布律. 若

$$f(\lambda_1x_1+\lambda_2x_2+\cdots+\lambda_nx_n)=\lambda_1f(x_1)+\lambda_2f(x_2)+\cdots+\lambda_nf(x_n),$$

则或者 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 构成退化分布，或者 $x_1=x_2=\cdots=x_n$.

定理 1.2.3(概率分布不等式)　如果向量$(p_1,p_2,\cdots,p_n)$，$(q_1,q_2,\cdots,q_n)$ 是两个概率分布，且 $q_i>0$，$i=1,2,\cdots,n$，则有不等式

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geqslant 0 \text{ 或} \sum_{i=1}^{n} p_i \log p_i \geqslant \sum_{i=1}^{n} p_i \log q_i \tag{1-12}$$

并且等号成立当且仅当两个概率分布相同即 $p_i=q_i, i=1,2,\cdots,n$.

注 1.2.3 在此定理中可能会遇到 $0\log\frac{0}{q_i}$, $0\log 0$, 故规定 $0\log\frac{0}{q_i}=0\log 0=0$.

证明 易证明函数 $f(x)=x\log x=x\ln x\log \mathrm{e}, x\geqslant 0$ 是严格凸函数, 如图 1-1(a)所示. 现在取

$$x_i=\frac{p_i}{q_i}, i=1,2,\cdots,n,$$

由 Jensen 不等式得

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{n} q_i \frac{p_i}{q_i} \log \frac{p_i}{q_i} = \sum_{i=1}^{n} q_i f(x_i) \geqslant f\left(\sum_{i=1}^{n} q_i x_i\right)$$

$$= f\left(\sum_{i=1}^{n} p_i\right) = f(1) = 0.$$

又因为 $(q_1,q_2,\cdots,q_n)$ 不是退化分布, 根据定理 1.2.2(3), 等号成立当且仅当 $x_1=x_2=\cdots=x_n$, 这就证明了 $p_i=q_i, i=1,2,\cdots,n$.

定理 1.2.4(对数和不等式) 设 $(a_1,a_2,\cdots,a_n)$ 是非负向量, 而 $(b_1,b_2,\cdots,b_n)$ 是正向量, 则

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geqslant \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}, \tag{1-13}$$

其中等号成立当且仅当这两个向量成比例.

证明 由于函数 $f(x)=x\log x, x\geqslant 0$ 是严格凸函数, 取

$$\lambda_i = \frac{b_i}{\sum_{j=1}^{n} b_i}, x_i = \frac{a_i}{b_i},$$

则 $\lambda_i>0, \sum_{i=1}^{n}\lambda_i = 1, x_i \geqslant 0$, 故由 Jensen 不等式

$$\sum_{i=1}^{n} \lambda_i f(x_i) \geqslant f\left(\sum_{i=1}^{n} \lambda_i x_i\right),$$

得到

$$\sum_{i=1}^{n} \frac{b_i}{\sum_{j=1}^{n} b_i} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geqslant \sum_{i=1}^{n} \left(\frac{b_i}{\sum_{j=1}^{n} b_i} \frac{a_i}{b_i}\right) \log \sum_{i=1}^{n} \left(\frac{b_i}{\sum_{j=1}^{n} b_i} \frac{a_i}{b_i}\right),$$

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{n} b_j} \log \frac{a_i}{b_i} \geqslant \sum_{i=1}^{n} \left(\frac{a_i}{\sum_{j=1}^{n} b_j}\right) \log \sum_{i=1}^{n} \left(\frac{a_i}{\sum_{j=1}^{n} b_j}\right).$$

两边同乘常量 $\sum_{j=1}^{n} b_j$ 得结果. 由于函数 $f(x)$ 是严格凸函数, 根据定理 1.2.2(3), 等号成立当且仅当所有 x_i 全相等, 这就证明了等号成立的充要条件.

注 1.2.4 在此定理中 $b_1,b_2,\cdots,b_n$ 也可以是非负向量，但应当 $\sum_{i=1}^{n}b_i>0$，此时不等式仍然成立，只是可能会遇到 $0\log\frac{0}{q}$，$p\log\frac{p}{0}$，故规定 $0\log\frac{0}{q}=0$，$p\log\frac{p}{0}=+\infty$.

1.3 随机事件的自信息量

现在来考虑随机事件的信息度量问题.

1.3.1 自信息

随机事件 A 发生与否具有不确定性，概率 $p=P(A)$ 是度量这种不确定性的一种数值. 因为概率小的事件不容易发生，要推测它何时发生比较困难，通常需要更多的信息；另一方面，概率小的事件一旦发生，一般会造成更大的轰动，产生强大的新闻效应，它提供的信息更多，因此可以说小概率事件包含的信息量大. 大概率事件因为较容易发生，要推测它何时发生需要的信息量少，它发生时产生的轰动效应也小，故可以说大概率事件包含的信息量少.

我们常用 $I(A)$ 或 $I(p)$ 来表示事件 A 包含的信息量大小，称为事件 A 的自信息量. 它应当具有什么表达式呢？根据上面分析，它应当是概率的函数 $I(A)=f(p)$，并且应当满足：

(1) 是概率 p 的单调递减函数；

(2) 具有可加性，即当两事件 A,B 独立时，$I(AB)=I(A)+I(B)$；

(3) 是非负函数；

(4) 当概率 $P(A)\to 0$ 时，$I(A)\to+\infty$；

(5) 当 $P(A)=1$ 时 $I(A)=0$.

满足这些性质的函数 $I(A)$ 可以由下面引理确定.

引理 1.3.1 如果非负函数 $f(x)$，$x\geqslant 1$ 满足下面两个条件：

(1) 当 $1\leqslant x<y$ 时 $f(x)<f(y)$，

(2) 当 $x,y\geqslant 1$ 时 $f(xy)=f(x)+f(y)$，

则这个函数必有表达式 $f(x)=C\log x$，其中常数 $C>0$.

证明留作练习.

根据这个引理可得：自信息量应当是概率的对数函数.

定义 1.3.1 随机事件 A 的自信息量 $I(A)$ 定义为

$$I(A)=\log\frac{1}{P(A)}.$$

基本事件 $\{X=x\}$ 的自信息量也记为 $I(x)$：

$$I(x)=\log\frac{1}{p(x)}=-\log p(x). \tag{1-14}$$

其中 $p(x)=p\{X=x\}>0$. 显然 $I(x)$，$x\in\mathscr{X}$ 是一个离散函数，故可以确定随机变量 X 的函数 $I(X)=-\log p(X)$.

1.3.2 自信息量的单位

自信息量的单位与对数底的选择有关，对不同的底用不同的单位.

(1) 用以 2 为底的对数时，$I(A)$的单位为比特，符号为 bit，常常用于工程上.

(2) 用以 e 为底的对数时，$I(A)$的单位为奈特，符号为 nat，常常用于理论推导上.

(3) 用以 10 为底的对数时，$I(A)$ 的单位为哈特，符号为 hat.

(4) 用其他大于 1 的正整数 d 为底的对数时，$I(A)$的单位为“d 进制信息单位”.

各个单位之间换算可用对数的换底公式求得：

(1) 1 nat$=\log_2 \mathrm{e}\approx 1.442\ 695$ bits；

(2) 1 hat$=\log_2 10\approx 3.321\ 928$ bits；

(3) 1 hat$=\ln 10\approx 2.302\ 585$ nats.

1.3.3 联合自信息

设两个随机事件$\{X=x\}$与$\{Y=y\}$同时发生的概率(不妨称为联合概率)$p(x,y)>0$，则这两个事件的联合自信息量定义为

$$I(x,y)=\log\frac{1}{p(x,y)}=-\log p(x,y). \tag{1-15}$$

可以推广到多个随机事件的情况，称

$$I(x_1,x_2,\cdots,x_n)=\log\frac{1}{p(x_1,x_2,\cdots,x_n)},x_i\in\mathscr{X}_i,i=1,2,\cdots,n \tag{1-16}$$

为随机事件$\{X_1=x_1\},\{X_2=x_2\},\cdots,\{X_n=x_n\}$的联合自信息量，其中 $p(x_1,x_2,\cdots,x_n)$为这 n 个随机事件的联合概率.

显然联合自信息量(1-15)定义了二元离散函数 $I(x,y),(x,y)\in\mathscr{X}\times\mathscr{Y}$，故可以确定随机变量 X,Y 的函数 $I(X,Y)$，类似地也有函数 $I(X_1,X_2,\cdots,X_n)$.

1.3.4 条件自信息

如果以某个事件发生为先决条件，还可以定义条件自信息量. 设在事件$\{X=x\}$发生条件下事件$\{Y=y\}$条件概率 $p(y|x)>0$，则称

$$I(y|x)=\log\frac{1}{p(y|x)}=-\log p(y|x) \tag{1-17}$$

为在事件$\{X=x\}$发生的条件下事件$\{Y=y\}$的条件自信息量.

可以推广到多个条件情况，称

$$I(y|x_1,x_2,\cdots,x_n)=\log\frac{1}{p(y|x_1,x_2,\cdots,x_n)} \tag{1-18}$$

为在事件$\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$发生的条件下事件$\{Y=y\}$的条件自信息量. 更一般地，称

$$I(y_1,y_2,\cdots,y_m|x_1,x_2,\cdots,x_n)=\log\frac{1}{p(y_1,y_2,\cdots,y_m|x_1,x_2,\cdots,x_n)}$$

为在事件$\{X_1=x_1,\cdots,X_n=x_n\}$发生的条件下事件$\{Y_1=y_1,\cdots,Y_m=y_m\}$的条件自信息量.

条件自信息用于刻画条件事件$\{X=x\}$的发生对事件$\{Y=y\}$包含信息量的影响，若事件$\{X=x\}$发生对事件$\{Y=y\}$的发生有利，则条件自信息量$I(y|x)$会比$I(y)$减少，否则会增大.

1.3.5 互自信息

设两个随机事件$\{X=x\}$与$\{Y=y\}$发生的概率$p_X(x)>0$，$p_Y(y)>0$，并且联合概率$p(x,y)>0$，则这两个随机事件的互自信息量定义为

$$I(x;y)=\log\frac{p(x,y)}{p_X(x)p_Y(y)}. \tag{1-19}$$

注意到互自信息的表达式(1-19)还可以写成

$$I(x;y)=\log p(x|y)-\log p_X(x)=I(x)-I(x|y).$$

故互自信息表示在已知事件$\{Y=y\}$发生的条件下事件$\{X=x\}$的自信息量的减少量；另外(1-19)也可以写成

$$I(x;y)=\log p(y|x)-\log p_Y(y)=I(y)-I(y|x).$$

故互自信息也表示在已知事件$\{X=x\}$发生的条件下事件$\{Y=y\}$的自信息量的减少量. 这两个量是相等的，从而可以认为互自信息量$I(x,y)$是两随机事件包含的公共信息量.

由于$I(x;y),(x,y)\in\mathscr{X}\times\mathscr{Y}$是二元离散函数，故它可以确定随机变量$X,Y$的二元函数

$$g(X,Y)=\log\frac{p(X,Y)}{p_X(X)p_Y(Y)}.$$

可以推广到多个随机事件的情况，称

$$I(x_1,\cdots,x_n;y_1,\cdots,y_m)=\log\frac{p(x_1,\cdots,x_n,y_1,\cdots,y_m)}{p_X(x_1,\cdots,x_n)p_Y(y_1,\cdots,y_m)}$$

为$\{X_1=x_1,\cdots,X_n=x_n\}$与$\{Y_1=y_1,\cdots,Y_m=y_m\}$的互自信息量，其中$p_X(x_1,x_2,\cdots,x_n)$为随机事件$\{X_1=x_1,\cdots,X_n=x_n\}$的联合概率，$p_Y(y_1,y_2,\cdots,y_m)$为随机事件$\{Y_1=y_1,\cdots,Y_m=y_m\}$的联合概率，$p(x_1,x_2,\cdots,x_n,y_1,y_2,\cdots,y_n)$为随机事件$\{X_1=x_1,\cdots,X_n=x_n,Y_1=y_1,\cdots,Y_m=y_m\}$的联合概率.

根据条件概率

$$p(x,y|z)=\frac{p(x,y,x)}{p_Z(z)},p(x|z)=\frac{p(x,z)}{p_Z(z)},q(y|z)=\frac{p(y,z)}{p_Z(z)}$$

也可以定义条件互自信息量

$$I(x;y|z)=\log\frac{p(x,y|z)}{p(x|z)q(y|z)}, \tag{1-20}$$

以及一般形式

$$I(x_1,\cdots,x_n;y_1,\cdots,y_m|z_1,\cdots,z_k)$$
$$=\log\frac{p(x_1,\cdots,x_n,y_1,\cdots,y_m|z_1,\cdots,z_k)}{p(x_1,\cdots,x_n|z_1,\cdots,z_k)q(y_1,\cdots,y_m|z_1,\cdots,z_k)}.$$

例题 1.3.1 如果随机变量(X,Y)具有分布律$p(x,y)$：