云南大学数学与统计学院
上机实践报告

课程名称： 信息论基础实验	**年级：** 2015 级	**上机实践成绩：**
指导教师： 陆正福	**姓名：** 刘鹏	
上机实践名称： 信息论中常用函数的图形绘制	**学号：** 20151910042	**上机实践日期：** 2017-10-26
上机实践编号： No.01	**组号：**	**上机实践时间：** 8:11

一、实验目的

1. 熟悉信息论中常用函数的图形，为后继学习奠定直观认识基础。
2. 熟悉实验所用编程平台

二、实验内容

绘制信息论中常用函数的图形

(1) $y = \ln(x)$
(2) $y = \ln(x) - x + 1$
(3) $y = x\ln(x)$
(4) $y = \frac{\ln(x)}{x}$
(5) $y = H(\mathrm{x}) = -x\ln(\mathrm{x}) - (1-x)\ln(1-x)$
(6) $D(p||q)(\text{given } q)$
(7) $D(p||p)$ (given p)
(8) $I(X;Y)\big(\text{given p}(y|x)\big)$
(9) $I(X;Y)\big(\text{given p}(x)\big)$

三、实验平台

Windows 10 1709 Enterprise 中文版；
Python 3.6.0；
Wing IDE Professional 6.0.5-1 集成开发环境。

四、实验记录与实验结果分析

把这一部分函数分为两部分，第一部分绘制比较简单的前四个图像。

题 1

绘制如下函数的图像：

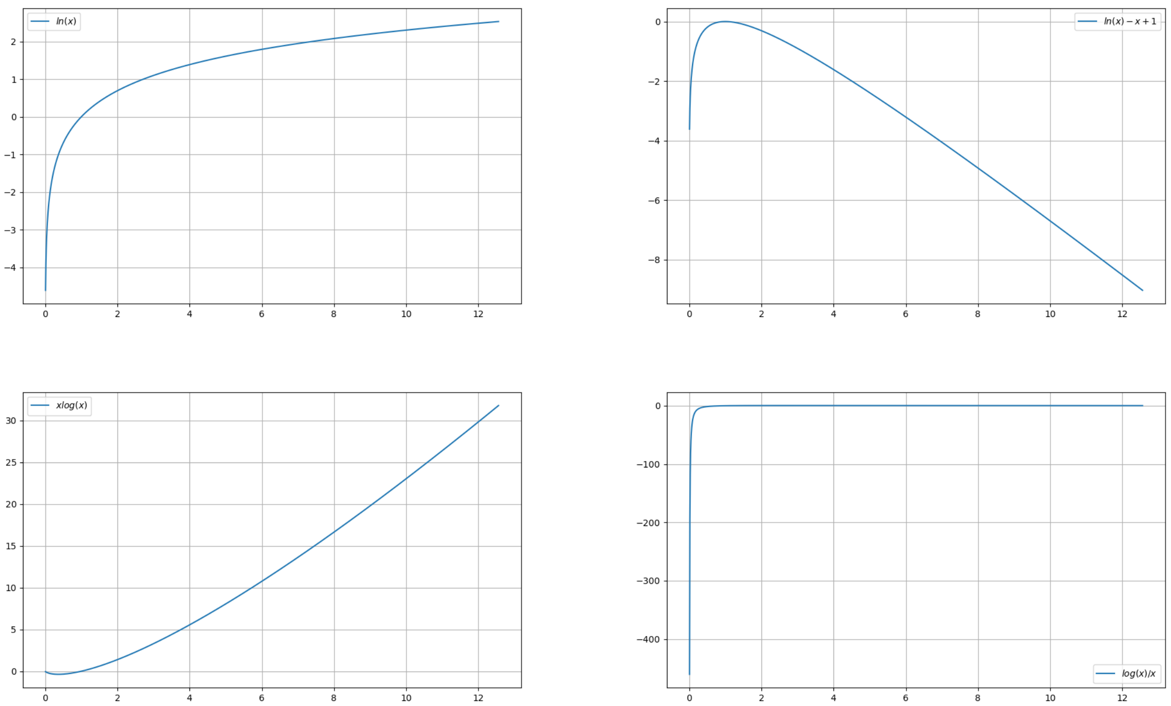
(1) $\mathrm{y} = \ln(\mathrm{x})$
(2) $\mathrm{y} = \ln(\mathrm{x}) - \mathrm{x} + 1$
(3) $\mathrm{y} = \mathrm{x}\ln(\mathrm{x})$
(4) $\mathrm{y} = \frac{\ln(\mathrm{x})}{\mathrm{x}}$

程序代码：

```python
"""filename: 1.1 Plot.py"""

import matplotlib.pyplot as pl
import numpy as np
```

```

x = np.arange(0, 4*np.pi, 0.01)
y = np.log(x)
pl.figure(figsize=(10,6))
pl.plot(x, y,label="$ln(x)$")

pl.grid()
pl.legend()
pl.show()

"""--------------------------------------"""

x = np.arange(0, 4*np.pi, 0.01)
y = np.log(x) - x + 1
pl.figure(figsize=(10,6))
pl.plot(x, y,label="$ln(x) - x + 1$")

pl.grid()
pl.legend()
pl.show()

"""--------------------------------------"""

x = np.arange(0, 4*np.pi, 0.01)
y = x * np.log(x)
pl.figure(figsize=(10,6))
pl.plot(x, y,label="$xlog(x)$")

pl.grid()
pl.legend()
pl.show()

"""--------------------------------------"""

x = np.arange(0, 4*np.pi, 0.01)
y = np.log(x) / x
pl.figure(figsize=(10,6))
pl.plot(x, y,label="$log(x)/x$")

pl.grid()
pl.legend()
pl.show()

"""--------------------------------------"""

x = np.arange(0, 4*np.pi, 0.01)
y = np.log(x) / x
pl.figure(figsize=(10,6))
pl.plot(x, y,label="$log(x)/x$")

pl.grid()
pl.legend()
pl.show()
```

输出结果：



代码分析：

利用 Python3 的 numpy 公开库，进行数据的生成，利用 matplotlib 的 pyplot 库进行函数图像绘制。

题 2

绘制如下函数的图像：

(5) $y = H(x) = -x\ln(x) - (1-x)\ln(1-x)$
(6) $D(p||q)(\text{given } q)$
(7) $D(p||p)\ (\text{given } p)$
(8) $I(X;Y)\big(\text{given } p(y|x)\big)$
(9) $I(X;Y)\big(\text{given } p(x)\big)$

分析：

对于(6)，很难给出一个图像，首先需要定义一个分布，但是在这之后同样不能作图像，因为函数是满足一一对应的映射。

五、教材翻译

Preface

＊前言

What is information? How do we quantify or measure the amount of information that is present in a file of data, or a string of text? How do we encode the information so that it can be stored efficiently, or transmitted reliably?

什么是信息？我们如何量化信息？如果可以量化，那如何测量数据、文本字符串中的信息？我们怎样编码信息才能在存储、传输过程中更加高效可靠？

The main concepts and principles of information theory were developed by Claude E. Shannon in the 1940s. Yet only now, and thanks to the emergence of the information age and digital communication, are the ideas of information theory being looked at again in a new light. Because of information theory and the results arising from coding theory we now know how to quantify information, how we can efficiently encode it and how reliably we can transmit it.

＊信息理论的主要概念和原理是由 Claude E. Shannon 在 20 世纪 40 年代开拓的。但是一直到现在，由于信息时代和数字通信的出现，信息理论的观念才又一次以新的视角重新审视。由于信息理论和编码理论的结果，我们现在知道如何量化信息，如何高效地编码信息以及如何可靠地传输信息。

This book introduces the main concepts behind how we model information sources and channels, how we code sources for efficient storage and transmission, and the fundamentals of coding theory and applications to state-of-the-art error correcting and error detecting codes.

＊本书介绍了信源建模和信道建模的主要概念，如何为高效存储和传输进行信源编码，以及编码理论和应用的优点，以及最新的纠错码和检错码。

This textbook has been written for upper level undergraduate students and graduate students in mathematics, engineering and computer science. Most of the material presented in this text was developed over many years at The University of Western Australia in the unit Information Theory and Coding 314, which was a core unit for students majoring in Communications and Electrical and Electronic Engineering, and was a unit offered to students enrolled in the Master of Engineering by Coursework and Dissertation in the Intelligent Information Processing Systems course.

＊本教材是为数学、工程和计算机科学专业的高年级本科生和研究生编写的。本书提供的大部分材料是西澳大利亚大学信息理论和 Coding 314 单元多年开发的，后者是通信和电子电气工程专业的核心单元，也是一个提供给在智能信息处理系统课程和论文工程硕士课程的学生的学习单元。

The number of books on the market dealing with information theory and coding has been on the rise over the past five years. However, very few, if any, of these books have been able to cover the fundamentals of the theory without losing the reader in the complex mathematical abstractions. And fewer books are able to provide the important theoretical framework when discussing the algorithms and implementation details of modern coding systems. This book does not abandon the theoretical foundations of information and coding theory and presents working algorithms and implementations which can be used to fabricate and design real systems. The main emphasis is on the underlying concepts that govern information theory and the necessary mathematical background that describe modern coding systems. One of the strengths of the book are the many worked examples that appear throughout the book that allow the reader to immediately understand the concept being explained, or the algorithm being described. These are backed up by fairly comprehensive exercise sets at the end of each chapter (including exercises identified by an * which are more advanced or challenging).

＊市场上处理信息理论和编码的书籍在过去五年中一直在增多。然而，很少有这些书能够在覆盖基本原理的同事，而不会让读者在复杂的数学问题上被弄晕。在讨论现代编码系统的算法和实现细节时，只有很少的书籍能够提供重要的理论框架。本书不会放弃信息和编码理论的理论基础，并提出可用于制造和设计真实系统的工作算法和实现。本书的主要重点是支配信息理论的基本概念和描述现代编码系统的必要数学背景。本书的优点之一是书中出现了很多可以让读者立即理解所阐述的概念和算法的例子。这些在每章末尾都有相当全面的练习套件（包括更先进或具有挑战性的练习）。

The material in the book has been selected for completeness and to present a balanced coverage. There is discussion of cascading of information channels and additivity of information which is rarely found in modern texts. Arithmetic coding is fully explained with both worked examples for encoding and decoding. The connection between coding of extensions and Markov modelling is clearly established (this is usually not apparent in other textbooks). Three complete chapters are devoted to block codes for error detection and correction. A large part of these chapters deals with an exposition of the concepts from abstract algebra that underpin the design of these codes. We decided that this material should form part of the main text (rather than be relegated to an appendix) to emphasize the importance of understanding the mathematics of these and other advanced coding strategies.

＊为呈现完整而均衡的覆盖面，书中的材料已经被选定。信道叠加与信息可叠加在本书中有讨论，但这在现代文献中很少见到。算法编码完全用编码和解码的两个例子来解释。扩展码和马尔可夫模型之间的联系清楚地建立起来（这在其他教科书中通常是很不容易见到的）。三个完整的章节专门用于错误检测和纠正的分组代码。这些章节的很大一部分是对抽象代数概念的阐述，这些概念来自这些代码的最初设计。我们决定把这些内容作为主要内容的一部分（而不是放在附录中），以强调理解这些和其他先进

编码策略的数学的重要性。

Chapter 1 introduces the concepts of entropy and information sources and explains how information sources are modelled. In Chapter 2 this analysis is extended to information channels where the concept of mutual information is introduced and channel capacity is discussed. Chapter 3 covers source coding for efficient storage and transmission with an introduction to the theory and main concepts, a discussion of Shannon's Noiseless Coding Theorem and details of the Huffman and arithmetic coding algorithms. Chapter 4 provides the basic principles behind the various compression algorithms including run-length coding and dictionary coders. Chapter 5 introduces the fundamental principles of channel coding, the importance of the Hamming distance in the analysis and design of codes and a statement of what Shannon's Fundamental Coding Theorem tells us we can do with channel codes. Chapter 6 introduces the algebraic concepts of groups, rings, fields and linear spaces over the binary field and introduces binary block codes. Chapter 7 provides the details of the theory of rings of polynomials and cyclic codes and describes how to analyze and design various linear cyclic codes including Hamming codes, Cyclic Redundancy Codes and Reed-Muller codes. Chapter 8 deals with burst-correcting codes and describes the design of Fire codes, BCH codes and Reed-Solomon codes. Chapter 9 completes the discussion on channel coding by describing the convolutional encoder, decoding of convolutional codes, trellis modulation and Turbo codes.
*第一章介绍熵和信源的概念，并解释信源建模。在第二章中，将这种分析扩展到了信息通道，并在其中塔伦了互信息和信道容量。第 3 章通过阐述 Shannon 的无噪声编码定理以及霍夫曼编码和算术编码算法的细节，介绍了可以有效存储和传输的信源编码，并介绍了理论和主要概念。第 4 章提供了各种压缩算法的基本原理，包括游程编码和字典编码器。第 5 章介绍了信道编码的基本原理、Hamming 距离在代码分析和设计中的重要性，以及 Shannon 基本编码定理下的信道编码可操作性。第 6 章介绍了二进制字段上的群，环，域和线性空间等代数概念，并介绍了二进制分组码。第 7 章详细介绍了多项式环和循环码理论，并介绍了如何分析和设计各种线性循环码，包括汉明码，循环冗余码和 Reed-Muller 码。第 8 章涉及到突发校正码，并描述了 Fire 码，BCH 码和 Reed-Solomon 码的设计。第 9 章通过描述卷积编码器，卷积码解码，网格调制和 Turbo 码完成信道编码的讨论。

This book can be used as a textbook for a one semester undergraduate course in information theory and source coding (all of Chapters 1 to 4), a one semester graduate course in coding theory (all of Chapters 5 to 9) or as part of a one semester undergraduate course in communications systems covering information theory and coding (selected material from Chapters 1, 2, 3, 5, 6 and 7).
*本书可作为信息理论与信源编码（全部第一至四章）一学期本科课程的教材，编码理论（全部第五章至第九章）的一学期研究生课程或作为通信系统中的一学期本科课程，涵盖信息论和编码（第 1，2，3，5，6 和 7 章的选材）。

We would like to thank Sean Davey and Nishith Arora for their help with the LATEX formatting of the manuscript. We would also like to thank Ken Rosen for his review of our draft manuscript and his many helpful suggestions and Sunil Nair from CRC Press for encouraging us to write this book in the first place! Our examples on arithmetic coding were greatly facilitated by the use of the conversion calculator (which is one of the few that can handle fractions!) made available by www.math.com.
*我们要感谢 Sean Davey 和 Nishith Arora 在稿件格式上的帮助。我们还要感谢肯·罗森对我们的稿子和他的许多有益的建议的评论，以及 CRC 出版社的 Sunil Nair，鼓励我们把这本书写在第一位！我们使用转换计算器（这是少数几个可以处理分数的计算器之一）使得我们在算术编码方面的例子变得非常容易。

The manuscript was written in L_AT^EX and we are indebted to the open source software community for developing such a powerful text processing environment. We are especially grateful to the developers of LyX (www.lyx.org) for making writing the document that much more enjoyable and to the makers of xfig (www.xfig.org) for providing such an easy-to-use drawing package.
*手稿是用 L_AT^EX 编写的，我们感谢开源软件社区开发如此强大的文本处理环境。我们特别感谢 LyX（www.lyx.org）的开发人员编写更令人愉快的文档，并感谢 xfig（www.xfig.org）的制作人提供这样一个易于使用的绘图软件包。

Roberto Togneri
Chris deSilva

Chapter 1 Entropy and Information

＊熵与信息

1.1 structure

＊1.1 节 结构

Structure is a concept of which we all have an intuitive understanding. However, it is not easy to articulate that understanding and give a precise definition of what structure is. We might try to explain structure in terms of such things as regularity, predictability, symmetry and permanence. We might also try to describe what structure is not, using terms such as featureless, random, chaotic, transient and aleatory.

＊结构是我们都有一个直观的理解的概念。然而，阐明这种理解并明确结构的定义是不容易的。我们可以尝试用规律性，可预测性，对称性和持久性等方面来解释结构。我们也可以尝试使用诸如无特征，随机，混沌，瞬态和偶然性等术语来描述什么不是结构。

Part of the problem of trying to define structure is that there are many different kinds of behavior and phenomena which might be described as structured, and finding a definition that covers all of them is very difficult.

＊因为有许多不同类型的行为和现象可能被描述为结构化的，所以找到一个涵盖所有这些结构的定义是非常困难的。

Consider the distribution of the stars in the night sky. Overall, it would appear that this distribution is random, without any structure. Yet people have found patterns in the stars and imposed a structure on the distribution by naming constellations.

＊考虑夜空中星星的分布。总体来看，这种分布似乎是随机的，没有任何结构。然而，人们已经在恒星中找到了模式，并且通过命名星座而在分布上施加了一个结构。

Again, consider what would happen if you took the pixels on the screen of your computer when it was showing a complicated and colorful scene and strung them out in a single row. The distribution of colors in this single row of pixels would appear to be quite arbitrary, yet the complicated pattern of the two-dimensional array of pixels would still be there.

＊再次考虑一下，如果在计算机屏幕上显示复杂多彩的场景并将其排成一行时，会发生什么情况。这个单行像素中的颜色分布看起来是相当随意的，但是二维像素阵列的复杂模式仍然存在。

These two examples illustrate the point that we must distinguish between the presence of structure and our perception of structure. In the case of the constellations, the structure is imposed by our brains. In the case of the picture on our computer screen, we can only see the pattern if the pixels are arranged in a certain way.

＊这两个例子说明我们必须区分结构的存在和我们对结构的认识。在星座的情况下，结构是由我们的大脑强加的。在我们电脑屏幕上的图片的情况下，只有像素以某种方式排列，我们才能看到图案。

Structure relates to the way in which things are put together, the way in which the parts make up the whole. Yet there is a difference between the structure of, say, a bridge and that of a piece of music. The parts of the Golden Gate Bridge or the Sydney Harbor Bridge are solid and fixed in relation to one another. Seeing one part of the bridge gives you a good idea of what the rest of it looks like.

＊结构涉及事物的组合方式，各部分构成整体的方式。然而，桥梁的结构和音乐的结构是有区别的。金门大桥或悉尼海港大桥的各个部分是牢固的，相互固定的。看到桥的一部分，可以让你很好地了解它的其余部分。

The structure of pieces of music is quite different. The notes of a melody can be arranged according to the whim or the genius of the composer. Having heard part of the melody you cannot be sure of what the next note is going to be, leave alone any other part of the melody. In fact, pieces of music often have a complicated, multi-layered structure, which is not obvious to the casual listener.

＊不同的音乐的结构是完全不同的。旋律的音符可以根据作曲家的奇思妙想或天才来安排。听完部分旋律，你不能确定下一个音符是什么，只留下旋律的其他部分。实际上，音乐片断往往有一个复杂的，多层次的结构，这对于偶然的听众来说是不明显的。

In this book, we are going to be concerned with things that have structure. The kinds of structure we will be concerned with will be like the structure of pieces of music. They will not be fixed and obvious.

＊在这本书中，我们将关注具有结构的东西。我们所关心的结构将会像音乐的结构一样。他们不会是固定的，明显的。

1.2 Structure in Randomness

＊1.2 节 随机结构

Structure may be present in phenomena that appear to be random. When it is present, it makes the phenomena more predictable. Nevertheless, the fact that randomness is present means that we have to talk about the phenomena in terms of probabilities.

＊结构可能存在于看似随机的现象中。当它出现时，它使现象更加可预测。然而，随机性是具有提前性的这一事实意味着我们必须从概率的角度来谈论随机现象。

Let us consider a very simple example of how structure can make a random phenomenon more predictable. Suppose we have a fair die. The probability of any face coming up when the die is thrown is 1/6. In this case, it is not possible to predict which face will come up more than one-sixth of the time, on average.

＊让我们考虑一个结构如何使随机现象更可预测的非常简单的例子。假设我们有一个均匀的骰子。骰子任何一面落地的概率是 1/6。在这种情况下，平均来说，任何一个面出现的概率都不会大于六分之一，我们也不会做出这种预测。

On the other hand, if we have a die that has been biased, this introduces some structure into the situation. Suppose that the biasing has the effect of making the probability of the face with six spots coming up 55/100, the probability of the face with one spot coming up 5/100 and the probability of any other face coming up 1/10. Then the prediction that the face with six spots will come up will be right more than half the time, on average.

＊另一方面，如果我们有一个不平均的骰子，这就引入了一些结构。假设偏倚的作用是让 6 点那一面着地的概率达到 55/100，1 点的那一面的概率达到 5/100，其他的概率达为 1/10。那么平均说来，6 点的面落地的概率就会超过一半

Another example of structure in randomness that facilitates prediction arises from phenomena that are correlated. If we have information about one of the phenomena, we can make predictions about the other. For example, we know that the IQ of identical twins is highly correlated. In general, we cannot make any reliable prediction about the IQ of one of a pair of twins. But if we know the IQ of one twin, we can make a reliable prediction of the IQ of the other.

＊有助于预测的随机结构的另一个例子来自相关现象。如果我们有关于其中一种现象的信息，我们可以预测另一种现象。例如，我们知道同卵双胞胎的智商高度相关。一般来说，我们不能对一对双胞胎之一的智商做出任何可靠的预测。但是如果我们知道一个双胞胎的智商，我们可以对另一个的智商做出可靠的预测。

In order to talk about structure in randomness in quantitative terms, we need to use probability theory.

＊为了从定量的角度来谈论随机性的结构，我们需要使用概率论。

1.3 First Concepts of Probability Theory

1.3 节 概率论的第一个概念

To describe a phenomenon in terms of probability theory, we need to define a set of outcomes, which is called the sample space. For the present, we will restrict consideration to sample spaces which are finite sets.

*为了用概率论来描述一个现象，我们需要定义一组结果，这就是所谓的样本空间。目前，我们仅限有限的样本空间。

DEFINITION 1.1 Probability Distribution A probability distribution *on a sample space* $\mathrm{S}=\{\mathrm{s}_1,\mathrm{s}_2,\cdots,\mathrm{s}_\mathrm{N}\}$ *is a function* P *that assigns a probability to each outcome in the sample space.* P *is a map from* S *to the unit interval,* $P: S\rightarrow[0,\ 1]$*, which must satisfy* $\sum_{i=1}^{N}P(s_i)=1$.

定义 1.1 随机分布 随机分布是指定义在样本空间$\mathrm{S}=\{\mathrm{s}_1,\mathrm{s}_2,\cdots,\mathrm{s}_\mathrm{N}\}$上的一个函数 P，对于 S 中的任何一个元素，都可以通过 P 找到一个对应。P 是一个由 S 到区间$[0,\ 1]$的一个映射，而且必须满足$\sum_{i=1}^{N}P(s_i)=1$。

DEFINITION 1.2 Events Events *are subsets of the sample space.*

定义 1.2 事件 事件是样本空间的子集。

We can extend a probability distribution P from S to the set of all subsets of S, which we denote by $\mathcal{P}(S)$, by setting $P(E)=\sum_{s\in E}P(s)$ for any $E\in\mathcal{P}(S)$. Note that $P(\emptyset)=0$.

*我们可以把定义在 S 上的概率分布 P 拓展到 S 的所有子集$\mathcal{P}(S)$上面，令$P(E)=\sum_{s\in E}P(s)$，其中$E\in\mathcal{P}(S)$，并且让$P(\emptyset)=0$。

If E and F are events and $E\cap F=\emptyset$ then $P(E\cup F)=P(E)+P(F)$.

*如果 E 与 F 是事件，而且两者之交为空，那么两个事件的并的分布等于两个时间的分布之和。

DEFINITION 1.3 Expected Value
If $S=\{s_1,\ s_2,\ \cdots,\ s_N\}$ *is a sample space with probability distribution P, and* $f: S\rightarrow V$ *is a function from the sample space to a vector space V, the expected value of f is* $f=\sum_{i=1}^{N}P(s_i)f(s_i)$

定义 1.3 期望值 如果$S=\{s_1,\ s_2,\ \cdots,\ s_N\}$是一个样本空间，$P$ 是 S 上的一个分布，$f: S\rightarrow V$是一个从样本空间到向量空间的映射，那么映射的期望值是$f=\sum_{i=1}^{N}P(s_i)f(s_i)$。

NOTE We will often have equations that involve summation over the elements of a finite set. In the equations above, the set has been $S=\{s_1,\ s_2,\ \cdots,\ s_N\}$ and the summation has been denoted by $\sum_{i=1}^{N}$. In other places in the text we will denote such summations simply by $\sum_{s\in S}$.

*我们将会经常用到有限集合上面元素和的等式。在上面的等式中，$S=\{s_1,\ s_2,\ \cdots,\ s_N\}$，和式可以通过符号$\sum_{i=1}^{N}$来表示，简记为$\sum_{s\in S}$。

1.4 Surprise and Entropy

*1.4 节 震惊度和熵

In everyday life, events can surprise us. Usually, the more unlikely or unexpected an event is, the more surprising it is. We can quantify this idea using a probability distribution.

*在日常生活中，事件会让我们感到吃惊。通常情况下，事件越不可能或不可预料，就越令人惊讶。我们可以用概率分布来量化这个想法。

DEFINITION 1.4 Surprise

If E is an event in a sample space S, we define the surprise of E to be $s(E)=-\log(P(E))=\log(1/P(E))$.

*定义 1.4 如果 E 是样本空间 S 里面的一个事件， E 的震惊度$s(E)=-\log(P(E))=\log(1/P(E))$。

Events for which $P(E)$ = 1, which are certain to occur, have zero surprise, as we would expect, and events that are impossible, that is, for which $P(E)$ = 0, have infinite surprise.

*$P(E)$ = 1 的事件肯定会发生，如我们所期望的那样，没有任何意外的事件发生，而 $P(E)$ = 0 的事件是无法预料的。

Defining the surprise as the negative logarithm of the probability not only gives us the appropriate limiting values as the probability tends to 0 or 1, it also makes surprise additive. If several independent events occur in succession, the total surprise they generate is the sum of their individual surprises.

*将震惊度定义为概率的负对数不仅给了我们适当的限制值，因为概率趋向于 0 或 1，它也使得我们可以添加。如果有几个独立的事件发生，那么他们产生的总震惊度就是各自震惊度的总和。

DEFINITION 1.5 Entropy *We can restrict the surprise to the sample space and consider it to be a function from the sample space to the real numbers. The expected value of the surprise is the entropy of the probability distribution.*

If the sample space is $S=\{s_1,\ s_2,\ \cdots,\ s_N\}$, with probability P, the entropy of the probability distribution is given by

$$H(P)=-\sum_{i=1}^{N}P(s_i)\log(P(s_i)) \tag{1.1}$$

*定义 1.5 熵我们可以将震惊度函数的取值限制在样本空间上，并认为它是从样本空间到实数的函数。震惊度的期望就是概率分布的熵。

如果样本空间是$S=\{s_1,\ s_2,\ \cdots,\ s_N\}$，分布为 P，那么分布的熵就是$H(P)=-\sum_{i=1}^{N}P(s_i)\log(P(s_i))$ (1.1)

The concept of entropy was introduced into thermodynamics in the nineteenth century. It was considered to be a measure of the extent to which a system was disordered. The tendency of systems to become more disordered over time is described by the Second Law of Thermodynamics, which states that the entropy of a system cannot spontaneously decrease. In the 1940’s, Shannon [6] introduced the concept into communications theory and founded the subject of information theory. It was then realized that entropy is a property of any stochastic system and the concept is now used widely in many fields. Today, information theory (as described in books such as [1], [2], [3]) is still principally concerned with communications systems, but there are widespread applications in statistics, information processing and computing (see [2], [4], [5]).

*熵的概念被引入了十九世纪的热力学。它被认为是衡量一个系统混乱的程度的量。随着时间的推移，系统变得更加混乱的趋势被描述为热力学第二定律，它指出一个系统的熵不能自然地减小。在 20 世纪 40 年代，香农将这个概念引入到通信理论中，并建立了信息论的主题。然后认识到熵是任何随机系统的一个属性，这个概念已经在当今许多领域被广泛使用。今天，信息理论（如[1]，[2]，[3]等书中所述）仍主要涉及通信系统，但在统计，信息处理和计算中有广泛的应用（见[2]，[4]，[5]）。

Let us consider some examples of probability distributions and see how the entropy is related to predictability. First, let us note the form of the function $s(p)=-p\log(p)$ where $0<p\leqslant1$ and $\log$ denotes the logarithm to base 2. (The actual base does not matter, but we shall be using base 2 throughout the rest of this book, so we may as well start here.) The graph of this function is shown in Figure 1.1.

*让我们考虑一些概率分布的例子，看看熵是如何与可预测性相关的。记$s(p)=-p\log(p)$，其中$0<p\leqslant1$，log 符号表示取以 2 为底的对数。（具体以谁为底并不重要，但是本书的其他部分用 2，所以这里也用 2）.这个函数的图像如 1.1 所示。

Note that $-p\log(p)$ approaches 0 as p tends to 0 and also as p tends to 1. This means that outcomes that are almost certain to occur and outcomes that are unlikely to occur both contribute little to the entropy. Outcomes whose probability is close to 0.4 make a comparatively large contribution to the entropy.

*当 p 趋于 0 或者 1 的时候，$-p\log(p)$趋于 0。这意味着几乎可以确定的结果以及不太可能发生的结果对熵的贡献都不大。其概率接近于 0.4 的结果对熵的贡献相对较大。

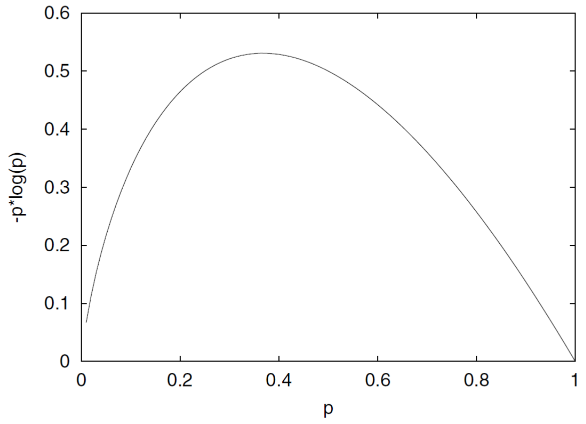


FIGURE 1.1
The graph of $-p\log(p)$.

Roughly speaking, a system whose entropy is *E* is about as unpredictable as a system with 2^E equally likely outcomes.

* 大体说来，熵为 E 的系统，与结果数目为2^E的均匀分布的不可预测性相同。

1.5 Units of Entropy

＊1.5 节 熵的单位

The units in which entropy is measured depend on the base of the logarithms used to calculate it. If we use logarithms to the base 2, then the unit is the bit. If we use natural logarithms (base e), the entropy is measured in natural units, sometimes referred to as nits. Converting between the different units is simple.

＊熵的测量单位取决于用来计算熵的基数。如果我们使用底数2的对数，那么单位就是这个位数。如果我们使用自然对数（e），则熵以自然单位（有时称为尼特）进行度量。不同单位之间的转换很简单。

PROPOSITION 1.1

If H_e is the entropy of a probability distribution measured using natural logarithms, and H_r is the entropy of the same probability distribution measured using logarithms to the base r, then

$$H_r = \frac{H_e}{\ln(r)}. \tag{1.2}$$

＊1.1 如果H_e是用自然对数进行计算得到的熵，而H_r是同分布下，用 r 为对数的底数进行计算得到的熵，那么

$$H_r = \frac{H_e}{\ln(r)}. \quad (1.2)$$

1.6 The Minimum and Maximum Values of Entropy

⋆1.6 节 熵的最大值与最小值

If we have a sample space S with N elements, and probability distribution P on S, it is convenient to denote the probability of $s_i \in S$ by p_i. We can construct a vector in R^N consisting of the probabilities:

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix}.$$

⋆对于样本空间 S，它具有 N 个元素，在此空间熵的分布 P，可以很方便地进行符号标记。我们构造一个向量，其元素包含每一个样本取值的概率：

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix}$$

Because the probabilities have to add up to unity, the set of all probability distributions forms a *simplex* in R^N, namely

⋆因为概率必须加起来为 1，所有概率分布的集合形成一个R^N中的单纯形，即

$$K = \left\{ p \in R^N : \sum_{i=1}^{N} p_i = 1 \right\}.$$

We can consider the entropy to be a function defined on this simplex. Since it is a continuous function, extreme values will occur at the vertices of this simplex, at points where all except one of the probabilities are zero. If $\mathbf{p}_v$ is a vertex, then the entropy there will be

$$H(\mathbf{p}_v) = (N-1).0.\log(0) + 1.\log(1).$$

⋆我们可以把熵看成是这个单纯形的函数。由于它是一个连续的函数，所以在这个单纯形的顶点上会出现极值，而在其他点为 0。如果$\mathbf{p}_v$是一个顶点，那么熵就是

$$H(\mathbf{p}_v) = (N-1).0.\log(0) + 1.\log(1).$$

The logarithm of zero is not defined, but the limit of $x\log(x)$ as x tends to 0 exists and is equal to zero. If we take the limiting values, we see that at any vertex, $H(\mathbf{p}_v) = 0$, as $\log(1) = 0$. This is the minimum value of the entropy function.

⋆零的对数没有被定义，但是趋于 0 的极限存在并且等于零。如果我们采取极限值，我们可以看到在任何顶点都有$H(\mathbf{p}_v) = 0$，如$\log(1) = 0$。这是熵函数的最小值

The entropy function has a maximum value at an interior point of the simplex. To find it we can use *Lagrange multipliers*.

⋆熵函数在单纯形的内点具有最大值。要找到它，我们可以使用拉格朗日乘子。

THEOREM 1.1

If we have a sample space with N elements, the maximum value of the entropy function is $\log(N)$.

⋆定理 1.1 如果我们有一个含有 N 个元素的样本空间，那么熵的最大值是$\log(N)$。

（译者按：这个证明，可以采取很多方法，书本的 PROOF 采用了拉格朗日乘子进行求偏导数，证明了这个定理。）

1.7 A Useful Inequality

＊1.7 节　一个有用的不等式

LEMMA 1.1

If $p_1, p_2, \cdots, p_N$ *and* $q_1, q_2, \cdots, q_N$ *are all non-negative numbers that satisfy the conditions* $\sum_{i=1}^{N} p_n = 1$ *and* $\sum_{i=1}^{N} q_n = 1$*, then*

$$-\sum_{i=1}^{N} p_i \log(p_i) \leqslant -\sum_{i=1}^{N} p_i \log(q_i)$$

with equality if and only if $p_i = q_i$ *for all* i.

＊**引理 1.1**　如果p_1，p_2，⋯，p_N和q_1，q_2，⋯，q_N都是非负数，且各自的和都为 1，那么有如下不等式：

$$-\sum_{i=1}^{N} p_i \log(p_i) \leqslant -\sum_{i=1}^{N} p_i \log(q_i)$$

当且仅当$p_i = q_i$（$i=1$，2，⋯，N）时等式成立。

（译者按：这个不等式的证明比较简单，主要是不等式$\ln x \leqslant x-1$的转化。）

The inequality can also be written in the form

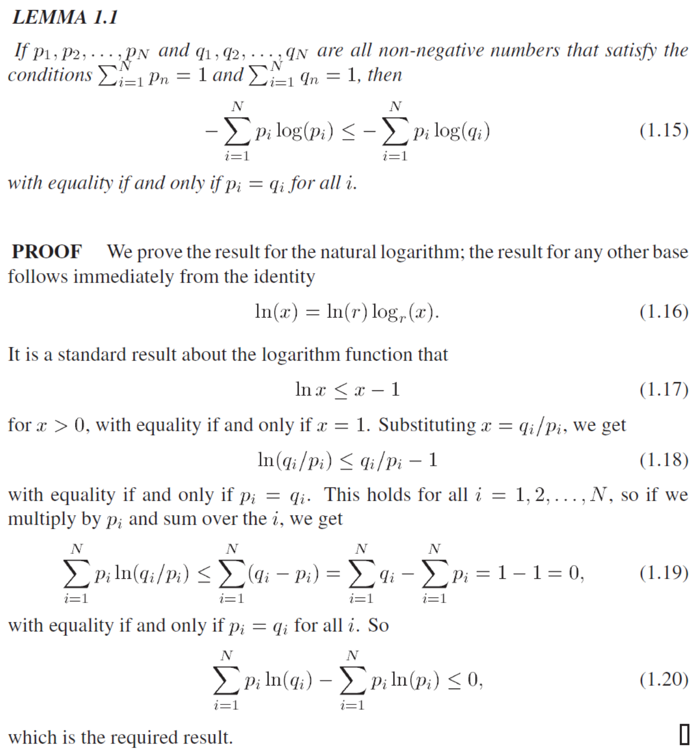
$$\sum_{i=1}^{N} p_i \log(q_i/p_i) \leqslant 0,$$

with equality if and only if $p_i = q_i$ for all i.

＊这个不等式还可以写成$\sum_{i=1}^{N} p_i \log(q_i/p_i) \leqslant 0$这种形式，当且仅当$p_i = q_i$（$i=1$，2，⋯，$N$）时等式成立。

Note that putting $q_i = 1/N$ for all i in this inequality gives us an alternative proof that the maximum value of the entropy function is $\log(N)$.

＊如果令$q_i = 1/N$，那么从这个不等式就可以看出，熵的最大值就是$\log(N)$。



LEMMA 1.1

If $p_1, p_2, \ldots, p_N$ *and* $q_1, q_2, \ldots, q_N$ *are all non-negative numbers that satisfy the conditions* $\sum_{i=1}^{N} p_n = 1$ *and* $\sum_{i=1}^{N} q_n = 1$*, then*

$$-\sum_{i=1}^{N} p_i \log(p_i) \leq -\sum_{i=1}^{N} p_i \log(q_i) \tag{1.15}$$

with equality if and only if $p_i = q_i$ *for all* i.

PROOF　We prove the result for the natural logarithm; the result for any other base follows immediately from the identity

$$\ln(x) = \ln(r)\log_r(x). \tag{1.16}$$

It is a standard result about the logarithm function that

$$\ln x \leq x - 1 \tag{1.17}$$

for $x > 0$, with equality if and only if $x = 1$. Substituting $x = q_i/p_i$, we get

$$\ln(q_i/p_i) \leq q_i/p_i - 1 \tag{1.18}$$

with equality if and only if $p_i = q_i$. This holds for all $i = 1, 2, \ldots, N$, so if we multiply by p_i and sum over the i, we get

$$\sum_{i=1}^{N} p_i \ln(q_i/p_i) \leq \sum_{i=1}^{N} (q_i - p_i) = \sum_{i=1}^{N} q_i - \sum_{i=1}^{N} p_i = 1 - 1 = 0, \tag{1.19}$$

with equality if and only if $p_i = q_i$ for all i. So

$$\sum_{i=1}^{N} p_i \ln(q_i) - \sum_{i=1}^{N} p_i \ln(p_i) \leq 0, \tag{1.20}$$

which is the required result. □

1.8 Joint Probability Distribution Functions

⋆1.8 节 联合概率分布函数

There are many situations in which it is useful to consider sample spaces that are the Cartesian product of two or more sets.

⋆在很多情况下，考虑两个或多个样本空间集合的笛卡尔积是有用的。

DEFINITION 1.6 Cartesian Product

Let $S=\{s_1, s_2, \cdots, s_M\}$ *and* $T=\{t_1, t_2, \cdots, t_N\}$ *be two sets. The Cartesian product of* S *and* T *is the set* $S\times T=\{(s_i, t_j) : 1\leqslant i\leqslant M, 1\leqslant j\leqslant N\}$.

⋆定义 1.6 笛卡尔积 集合$S=\{s_1, s_2, \cdots, s_M\}$和集合$T=\{t_1, t_2, \cdots, t_N\}$做笛卡尔乘积，得到有序二元对集合$S\times T=\{(s_i, t_j) : 1\leqslant i\leqslant M, 1\leqslant j\leqslant N\}$。

The extension to the Cartesian product of more than two sets is immediate.

⋆笛卡儿积的多集合推广很显然。

DEFINITION 1.7 Joint Probability Distribution

A joint probability distribution *is a probability distribution on the Cartesian product of a number of sets.*

⋆定义 1.7 联合分布律 联合概率分布是一个定义在集合的笛卡儿积上面的概率分布。

If we have S and T as above, then a joint probability distribution function assigns a probability to each pair (s_i, t_j). We can denote this probability by p_{ij}. Since these values form a probability distribution, we have

$$0\leqslant p_{ij}\leqslant 1$$

for $1\leqslant i\leqslant M$, $1\leqslant j\leqslant N$, and

$$\sum_{i=1}^{M}\sum_{j=1}^{N} p_{ij}=1$$

⋆对于上述的S和T，联合概率分布对每一个有序对(s_i, t_j)都赋予了一个概率值。可以把这个概率记为p_{ij}。因为这些数值构成了一个概率分布，所以$0\leqslant p_{ij}\leqslant 1$，而且$\sum_{i=1}^{M}\sum_{j=1}^{N} p_{ij}=1$。

If P is the joint probability distribution function on $S\times T$, the definition of entropy becomes

$$H(P)=-\sum_{i=1}^{M}\sum_{j=1}^{N}P(s_i, t_j)\log(P(s_i, t_j))=-\sum_{i=1}^{M}\sum_{j=1}^{N}p_{ij}\log(p_{ij})$$

⋆如果 P 是一个定义在$S\times T$上的概率分布函数，那么定义在这个分布上的熵就是：

$$H(P)=-\sum_{i=1}^{M}\sum_{j=1}^{N}P(s_i, t_j)\log(P(s_i, t_j))=-\sum_{i=1}^{M}\sum_{j=1}^{N}p_{ij}\log(p_{ij})$$

If we want to emphasize the spaces S and T, we will denote the entropy of the joint probability distribution on $S\times T$ by $H(P_{S\times T})$ or simply by $H(S, T)$. This is known as the joint entropy of S and T.

⋆如果要特意强调一下样本空间，可以把熵的联合分布形式写成$H(P_{S\times T})$，或者简记为$H(S, T)$。这被称为S与T的联合熵。

If there are probability distributions P_S and P_T on S and T, respectively, and these are independent, the joint probability distribution on $S\times T$ is given by

$$p_{ij}=P_S(s_i)P_T(t_j)$$

for $1\leqslant i\leqslant M$, $1\leqslant j\leqslant N$. If there are correlations between the s_i and t_j, then this formula does not apply.

⋆对于分别定义在S和T上的概率分布P_S和P_T，如果它们独立，那么联合分布的概率取值与各自的概率分布有如下关系：

$$p_{ij}=P_S(s_i)P_T(t_j)$$

其中$1\leqslant i\leqslant M$，$1\leqslant j\leqslant N$。如果两者之间存在相关关系，那么这个公式就不适用。

DEFINITION 1.8 Marginal Distinction

If P is a joint probability distribution function on $S\times T$*, the marginal distribution on* S *is* $P_S : S\rightarrow[0,1]$ *given by*

$$P_S(s_i)=\sum_{j=1}^{N}P(s_i, t_j)$$

for $1\leqslant i\leqslant N$ *and the marginal distribution on* T *is* $P_T : T\rightarrow[0,1]$ *given by*

$$P_T(s_i)=\sum_{i=1}^{N}P(s_i, t_j)$$

for $1\leqslant j\leqslant N$.

⋆**定义 1.8** 边际分布 如果 P 是一个定义在$S\times T$上面的联合概率分布函数，那么S上面的边际概率定义为映射$P_S : S\rightarrow[0,1]$，表示为：

$$P_S(s_i)=\sum_{j=1}^{N}P(s_i, t_j)$$

同样，T上面的边际分布为映射$P_T : T\rightarrow[0,1]$，定义为

$$P_T(t_j)=\sum_{i=1}^{M}P(s_i, t_j)$$

There is a simple relation between the entropy of the joint probability distribution function and that of the marginal distribution functions.

⋆联合分布熵与边际分布熵之间有一个简单的关系。

THEOREM 1.2

If P is a joint probability distribution function on $S \times T$, P_S and P_T are the marginal distributions on S and T, respectively, then

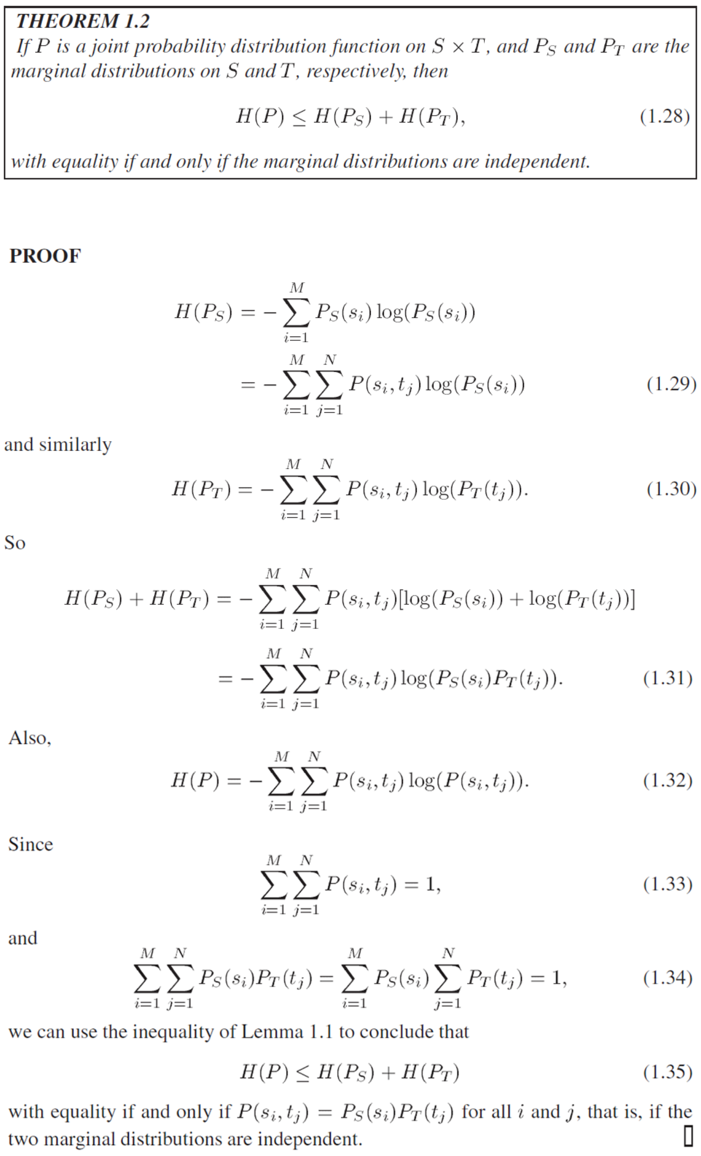
$$H(P) \leqslant H(P_S) + H(P_T)$$

with the equality if and only if the marginal distributions are independent.

＊定理 1.2：P 是$S \times T$上面的联合分布函数，P_S与P_T是分别定义在 S 和 T 上面的边际分布，那么有如下不等式成立

$$H(P) \leqslant H(P_S) + H(P_T)$$

当且仅当两个边际分布相互独立的时候等号成立。



THEOREM 1.2

If P is a joint probability distribution function on $S \times T$, and P_S and P_T are the marginal distributions on S and T, respectively, then

$$H(P) \leq H(P_S) + H(P_T), \tag{1.28}$$

with equality if and only if the marginal distributions are independent.

PROOF

$$\begin{aligned} H(P_S) &= -\sum_{i=1}^{M} P_S(s_i)\log(P_S(s_i)) \\ &= -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log(P_S(s_i)) \end{aligned} \tag{1.29}$$

and similarly

$$H(P_T) = -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log(P_T(t_j)). \tag{1.30}$$

So

$$\begin{aligned} H(P_S) + H(P_T) &= -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)[\log(P_S(s_i)) + \log(P_T(t_j))] \\ &= -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log(P_S(s_i)P_T(t_j)). \end{aligned} \tag{1.31}$$

Also,

$$H(P) = -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log(P(s_i,t_j)). \tag{1.32}$$

Since

$$\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j) = 1, \tag{1.33}$$

and

$$\sum_{i=1}^{M}\sum_{j=1}^{N} P_S(s_i)P_T(t_j) = \sum_{i=1}^{M} P_S(s_i)\sum_{j=1}^{N} P_T(t_j) = 1, \tag{1.34}$$

we can use the inequality of Lemma 1.1 to conclude that

$$H(P) \leq H(P_S) + H(P_T) \tag{1.35}$$

with equality if and only if $P(s_i,t_j) = P_S(s_i)P_T(t_j)$ for all i and j, that is, if the two marginal distributions are independent. ▯

1.9 Conditional Probability and Bayes' Theorem

⋆1.9 节 条件概率与贝叶斯理论

DEFINITION 1.9 Conditional Probability
If S is a sample space with a probability distribution function P, and E and F are events in S, the conditional probability of E given F *is*

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

⋆定义 1.9 条件概率：如果 S 是一个样本空间，概率分布函数 P 定义在 S 上，E 与 F 都是空间 S 中的时间，那么在给定 F 的条件下，E 的条件概率是

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

It is obvious that

$$P(E|F)P(F) = P(E \cap F) = P(F|E)P(E)$$

Almost as obvious is one form of *Bayes' Theorem*:

THEOREM 1.3
If S is a sample space with a probability distribution function P, and E and F are events in S, then

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

⋆定理 1.3：P 是定义在样本空间 S 上面的一个概率分布函数，E 与 F 是 S 中的事件，那么有如下等式成立

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

Bayes' Theorem is important because it enables us to derive probabilities of hypotheses from observations, as in the following example.

⋆贝叶斯定理非常重要，因为它使我们能够从观察中得出假设的概率，如下例所示。

EXAMPLE 1.7

We have two jars, A and B. Jar A contains 8 green balls and 2 red balls. Jar B contains 3 green balls and 7 red balls. One jar is selected at random and a ball is drawn from it.

We have probabilities as follows. The set of jars forms one sample space, $S = \{A, B\}$, with

$$P(\{A\}) = 0.5 = P(\{B\})$$

as one jar is as likely to be chosen as the other.

The set of colours forms another sample space, $T = \{G, R\}$. The probability of drawing a green ball is

$$P(\{G\}) = 11/20 = 0.55,$$

as 11 of the 20 balls in the jars are green. Similarly,

$$P(\{R\}) = 9/20 = 0.45.$$

We have a joint probability distribution over the colours of the balls and the jars with the probability of selecting Jar A and drawing a green ball being given by

$$P(\{(G, A)\}) = 0.4.$$

Similarly, we have the probability of selecting Jar A and drawing a red ball

$$P(\{(R, A)\}) = 0.1,$$

the probability of selecting Jar B and drawing a green ball

$$P(\{(G, B)\}) = 0.15,$$

and the probability of selecting Jar B and drawing a red ball

$$P(\{(R, B)\}) = 0.35.$$

We have the conditional probabilities: given that Jar A was selected, the probability of drawing a green ball is

$$P(\{G\}|\{A\}) = 0.8,$$

and the probability of drawing a red ball is

$$P(\{R\}|\{A\}) = 0.2.$$

Given that Jar B was selected, the corresponding probabilities are:

$$P(\{G\}|\{B\}) = 0.3,$$

and

$$P(\{R\}|\{B\}) = 0.7.$$

We can now use Bayes' Theorem to work out the probability of having drawn from either jar, given the colour of the ball that was drawn. If a green ball was drawn, the probability that it was drawn from Jar A is

$$P(\{A\}|\{G\}) = \frac{P(\{G\}|\{A\})P(\{A\})}{P(\{G\})} = 0.8 \times 0.5/0.55 = 0.73,$$

while the probability that it was drawn from Jar B is

$$P(\{B\}|\{G\}) = \frac{P(\{G\}|\{B\})P(\{B\})}{P(\{G\})} = 0.3 \times 0.5/0.55 = 0.27.$$

If a red ball was drawn, the probability that it was drawn from Jar A is

$$P(\{A\}|\{R\}) = \frac{P(\{R\}|\{A\})P(\{A\})}{P(\{R\})} = 0.2 \times 0.5/0.45 = 0.22,$$

while the probability that it was drawn from Jar B is

$$P(\{B\}|\{R\}) = \frac{P(\{R\}|\{B\})P(\{B\})}{P(\{R\})} = 0.7 \times 0.5/0.45 = 0.78.$$

(In this case, we could have derived these conditional probabilities from the joint probability distribution, but we chose not to do so to illustrate how Bayes' Theorem allows us to go from the conditional probabilities of the colours given the jar selected to the conditional probabilities of the jars selected given the colours drawn.) □

1.10 Conditional Probability Distributions and Conditional Entropy

＊1.10 节 条件概率分布与条件熵

In this section, we have a joint probability distribution P on a Cartesian product $S \times T$, where $S=\{s_1, s_2, \cdots, s_M\}$ and $T=\{t_1, t_2, \cdots, t_N\}$, with marginal distributions P_S and P_T.

＊在本节中，联合概率分布 P 是定义在笛卡儿积$S \times T$上的，其中$S=\{s_1, s_2, \cdots, s_M\}$，$T=\{t_1, t_2, \cdots, t_N\}$，而且 P 在 S 和 T 上面有边缘分布P_S与P_T。

DEFINITION 1.10 Conditional Probability of s_i ***given*** t_j

For $s_i \in S$ *and* $t_j \in T$, the conditional probability of s_i given t_j *is*

$$P(s_i|t_j)=\frac{P(s_i, t_j)}{P_T(t_j)}=\frac{P(s_i, t_j)}{\sum_{i=1}^{M} P(s_i, t_j)}$$

＊定义 1.10 给定t_j下s_i的条件概率：$s_i \in S$，$t_j \in T$，给定t_j下s_i的条件概率是

$$P(s_i|t_j)=\frac{P(s_i, t_j)}{P_T(t_j)}=\frac{P(s_i, t_j)}{\sum_{i=1}^{M} P(s_i, t_j)}$$

（这就是一般意义的条件概率）

DEFINITION 1.11 Conditional Probability Distribution given t_j

For a fixed t_j, *the conditional probabilities* $P(s_i|t_j)$ *sum to 1 over i, so they form a probability distribution on S, the conditional probability distribution given* t_j. *We will denote this by* $P_{S|t_j}$.

＊定义 1.11 给定t_j下的条件概率分布：对于给定的t_j，存在一组条件概率$P(s_i|t_j)$，其和为 1，所以这一组概率构成了 S 上的一个概率分布，称为给定t_j的条件概率分布。记为$P_{S|t_j}$。

（联合分布表中的一行，或者一列）

（这里要明白**条件概率**与**条件概率分布**的两个内涵。）

DEFINITION 1.12 Conditional Entropy given t_j

The conditional entropy given t_j *is the entropy of the conditional probability distribution on S given* t_j. *It will be denoted* $H(P_{S|t_j})$.

$$H(P_{S|t_j})=-\sum_{i=1}^{M} P(s_i|t_j)\log(P(s_i|t_j))$$

＊定义 1.12 给定t_j下的条件熵：给定t_j下的条件熵指的是给定t_j下的条件概率分布的熵，记作$H(P_{S|t_j})$，

$$H(P_{S|t_j})=-\sum_{i=1}^{M} P(s_i|t_j)\log(P(s_i|t_j))$$

一列“1 化”之后的熵。

DEFINITION 1.13 Conditional Probability Distribution on S ***given*** T

The conditional probability distribution on S given T *is the weighted average of the conditional probability distributions given* t_j *for all* j. *It will be denoted* $P_{S|T}$.

$$P_{S|T}(s_i)=\sum_{j=1}^{N} P_T(t_j) P_{S|t_j}(s_i)$$

＊定义 1.13 给定 T 下的定义在 S 上的条件概率分布：给定 T 下 S 条件概率分布时给出所有的t_j时，得到的一个加权平均，记作$P_{S|T}$。

$$P_{S|T}(s_i)=\sum_{j=1}^{N} P_T(t_j) P_{S|t_j}(s_i)$$

（如果我没有算错，这个量恒等于 1）

DEFINITION 1.13 Conditional Entropy given T

The conditional entropy given T *is the weighted average of the conditional entropies on S given* t_j *for all* $t_j \in T$. *It will be denoted* $H(P_{S|T})$.

$$H(P_{S|T})=-\sum_{j=1}^{N} P_T(t_j) \sum_{i=1}^{M} P(s_i|t_j)\log(P(s_i|t_j))$$

Since $P_T(t_j)P(s_i|t_j)=P(s_i, t_j)$, we can re-write this as

$$H(P_{S|T})=-\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j)\log(P(s_i|t_j))$$

＊定义 1.13 给定T下的条件熵：给定T下的条件熵时给定t_j时，S 上条件熵的加权平均，记作$H(P_{S|T})$。

$$H(P_{S|T})=-\sum_{j=1}^{N} P_T(t_j) \sum_{i=1}^{M} P(s_i|t_j)\log(P(s_i|t_j))$$

因为$P_T(t_j)P(s_i|t_j)=P(s_i, t_j)$，所以上式可以改写为

$$H(P_{S|T})=-\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j)\log(P(s_i|t_j))$$

六、实验体会

作为一门专业课，必然要有一个固定的教材，用一批教材上课，也需要有一本主教材。在这里我选择内容较为精炼的 *Fundamentals of Information Theory and Coding Design* 进行学习。

首先要明确什么是信息。信息的表现，不限于一个一个的 message，如果抽象一下，就会发现信息是定义在一个有限取值空间上面的随机变量。如此一来，我们就可以对信息进行纯数学的考察。有的时候我们说某一篇文章的信息量非常大，或者说某一句话的信息量十分大，但是这里的信息量并不是信息论中所研究的信息，前者太过于复杂了，很难入手。可以把一个信息源看作一个机器，这台机器每隔一秒钟就会在屏幕上随机（取值空间有限）打印一个字符，然后下一秒湮灭这个字符，紧接着产生一个新的字符。在这种情况下，信息就是这个机器的打印行为。如果有两台机器都可以打印，只是打印的概率不同，那么我们可以通过一种数学模型来衡量两台机器的信息量。

假设两台机器的打印空间是一样的，比如都是英文字母表。A 机器的打印比较平均，一段时间内，26 个字母的打印数量基本相当（依概率），但是 B 机器就不一样了，B 打印前 13 个字母的概率比后 13 个字母的概率大很多。很显然，在这种情况下，260 秒过后，A 机器每个字母平均打印了 10 次（依概率），B 机器则不然，1-13 号明显多于 10 次，14-26 号明显少于 10 次。

考虑一种极端情况，B 机器只打印前 13 个字母，即 B 的字母表只剩下了一半。这样一来，260 秒过后，A 与 B 的结果迥然不同。对于一套文字系统，字母表越复杂，单个字所表示的含义越丰富，但是对这个字母系统进行编码时，一个字符占的 bit 会更多；当一套文字系统比较简单时，单个字所代表的含义越少，但是对这个字母系统进行编码时，一个字符占的 bit 会更少。就如 ASCII 编码与汉字编码，两套体系的体系完全不一样，前者远远小于后者。但是汉字言简意赅。

当然，以上内容完全是依据经验而谈。

七、参考文献