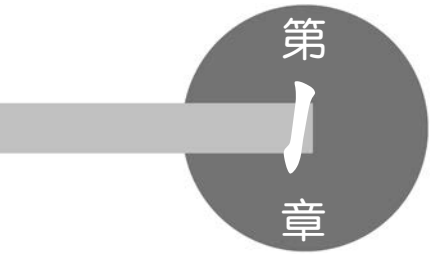

贝叶斯统计基本概念

俗话说，万事开头难。为了增强读者的学习兴趣，本章从一个贝叶斯统计的真实应用案例开始，介绍贝叶斯统计的基本概念和贝叶斯公式，概述贝叶斯统计学的历史和发展趋势以及与经典统计学的比较，最后，详细讨论了贝叶斯方法在人工智能领域的一个应用。

1.1 引　　言

1.1.1 一个美国书呆子的故事

在2012年美国总统大选期间，一个一直都被人称作“书呆子”的美国人纳特·西尔弗(Nate Silver，生于1978年1月13日)用以统计为主要工具的模型准确预测了美国50个州的选举结果。在大选日当天早晨，他的模型最新预测到时任总统巴拉克·奥巴马(Barack Obama)有90.9%的可能获得多数选举人票从而连任，而选举结果确确实实就是奥巴马总统赢得了这次美国总统大选。于是，他凭借自己的模型及其准确的预测打败了所有时事政治记者、政党媒体顾问和政治评论员。“你们知道谁是今晚(大选日当夜)的赢家吗?”美国全国广播公司新闻节目主播自问自答，“是纳特·西尔弗。”其实，早在2008年的美国总统大选期间，西尔弗就准确预测了美国50个州中49个州的选举结果。两次极为准确的预测，让这个“书呆子”扬眉吐气、名声大震，各种荣誉接踵而来，甚至于被至少4所大学授予了荣誉博士学位，当然也让我们统计和数据科学工作者大感骄傲。西尔弗的预测模型有什么神秘之处呢？那就是利用了大数据和我们将要学习的贝叶斯统计理论与方法。

1.1.2 贝叶斯统计简史

贝叶斯统计学是以英国人托马斯·贝叶斯(Thomas Bayes，1702—1761)的名字命名的。贝叶斯是一位英国牧师，但他却热衷于概率统计等科学研究，还是英国皇家学会会员。遗憾的是，现在人们对他的生平却知之甚少，甚至没有人知道贝叶斯的相貌如何，现存所有他的画像都是传说，并不能证实是他的真容。贝叶斯统计学起源于贝叶斯逝世后公开发表的一篇论文——《论一个概率理论问题的求解》(*An Essay Towards Solving a Problem in the Doctrine of Chances*)。这篇论文在贝叶斯去世两年之后由他的朋友理查德·普莱斯(Richard Price)介绍到英国皇家学会，引起了该学会的注意和讨论，并于1763年发表在《皇家学会哲学会刊》上。在该论文中，贝叶斯首次提出了贝叶斯统计的基本思想和归纳推理方法。

51 年后,法国数学、概率与统计学、天文学和物理学家拉普拉斯(P. S. Laplace,1749—1827)出版了著作《关于概率的哲学评述》(*A Philosophical Essay on Probabilities*)。在该著作中,他将贝叶斯提出的公式进行了推广并导出了一些很有意义的新结果。然而,之后相当长的一段时间里,虽然有一些理论和应用研究,但由于其理论与经典统计学相比显得另类而且人们对它的理解还不够深刻,在应用上其又计算复杂且计算量巨大,因此贝叶斯统计理论和方法长期未被普遍接受,甚至于被一些经典学者看作一种旁门左道。直到 20 世纪中叶,一批统计学家,如杰弗里斯(Jeffreys,1939,1961)、萨维奇(Savage,1954)、雷法和施莱弗(Raiffa and Schlaifer,1961)以及伯杰(Berger, 1985,1993; 中译本,1998)等才对贝叶斯统计做了更加深入的研究,特别是罗马尼亚(匈牙利)裔美国统计学家瓦尔德(Wald,1939,1950; 中译本,1963)通过将损失函数引入统计学并利用决策概念和思想把经典统计推断纳入决策理论框架而形成了统计决策理论,这样经典统计学和贝叶斯统计学通过决策理论有机地联系到了一起,得到了很有意义的理论结果。从 20 世纪中叶开始,在一批学者的努力下,人们对贝叶斯统计在观点、方法和理论上的认识不断加深。从 20 世纪 90 年代以来,伴随着计算机科学技术的发展和有效的贝叶斯统计计算方法的发明及应用,贝叶斯统计解决了相当一批经典统计难以解决的重要实际问题,从而得到了人们极大的重视。现在,贝叶斯统计理论和方法获得了人们的普遍接受,贝叶斯统计不仅在统计学本身而且在众多学科中都得到了广泛的应用,解决了各个不同学科中大量的复杂问题。贝叶斯统计表现出了勃勃生机和欣欣向荣的景象,在统计学领域牢牢地站稳了一席之地,是现代统计学的重要组成部分。

1.1.3 经典统计方法

我们先来回顾一下经典统计学的思想方法,以便与下一小节的贝叶斯统计思想方法进行比较。回忆一下概率统计课程中概率的定义,便容易明白经典统计学思想方法也就是“频率方法”,它把概率定义为频率的极限,也就是说随着随机试验重复次数的增多,随机事件发生的频率会稳定在一个常数附近,这个常数就是该随机事件发生的概率。同时,它认为总体的数字特征(如均值、方差等)和别的参数仅仅是未知的常数,可以用样本统计量(即样本的函数)来估计。此外,它又认为样本是随机变量,从而样本统计量也是随机的,因此具有概率分布即它的抽样分布。如果样本统计量的分布可以求出,利用该分布,就可以进行区间估计和假设检验等统计推断。然而,我们知道在经典统计中寻求统计量的概率分布和进行区间估计以及假设检验等都不是容易的事,而且参数的区间估计既不容易理解也不容易解释。

1.1.4 贝叶斯统计方法

贝叶斯统计学虽然也认可经典统计学的概率定义,但它同时把概率理解为人对随机事件发生可能性的一种信念(有时被称为“可信度”),当然,这种信念不是信口开河,而是基于学识和经验的审慎度量。此外,贝叶斯统计把任意一个未知量(参数)都看作一个随机变量,可用一个概率分布去描述它。我们认为这种观点是合理的,因为即使是一个确定性的未知量,也可以把它看成随机变量的特殊情形,即服从 0—1 分布的随机变量。所以说,任一个未知量都可用一个适当的概率分布去描述。这个概率分布利用历史数据或其他历史信息或研究人员的经验和学识而确定,称为该未知量(参数)的**先验分布**。而后利用新样本信息(即抽

样信息)对先验分布进行更新,更新之后的这个新概率分布称为该未知量的**后验分布**。由此,未知参数的点估计、区间估计和假设检验等统计推断都是基于后验分布来进行,而且参数的区间估计既容易理解也容易解释,假设检验则简单明了。

经典统计学把概率定义为频率的极限,初看起来似乎客观、严谨,但是在现实世界要进行重复试验,要么需要花费大量的人力、物力,要么根本无法进行。例如,我们无法重复昨天的天气和去年的经济活动。因此,用频率的极限来定义概率在实际应用中受到了极大的限制。相反,贝叶斯统计把概率理解为人对随机事件发生可能性的信念则在实际应用中没有任何限制,因为它不需要重复,事件甚至可以一次都没有发生过。而且,在贝叶斯统计中,一旦后验分布建立,所有的统计推断都是基于后验分布来进行的。因此,至少从理论而言,贝叶斯统计推断比经典统计推断要简单明了得多。当然,现代统计学的发展趋势是,根据实际问题的条件和需要挑选经典统计方法或贝叶斯统计方法,有时甚至是综合利用这两种统计理论和方法进行统计推断。所以,不管是经典统计还是贝叶斯统计,能够解决问题的就是"好统计"!

对于经典统计学与贝叶斯统计学的比较,学完本书的内容后才能有更深刻的体会,因此希望读者在研读完本书后,再好好对它们做一个详细的比较分析。

1.2　概率空间与随机事件贝叶斯公式

1.2.1　柯氏概率论公理体系与贝叶斯公式

我们从概率论知道概率空间是三位一体的一个研究对象(Ω,F,P),其中,Ω 是样本点(基本事件)全体,也称为样本空间;F 是事件域(简单说就是所要研究的随机事件全体,包含必然事件 Ω 和不可能事件 Φ);P 是定义在事件域 F 上的概率(测度),满足以下三条公理:

(1) 非负性:对于任意事件 A,其概率 $P(A)\geqslant 0$;

(2) 规范性:必然事件 Ω 的概率等于 1,即 $P(\Omega)=1$;

(3) 可列可加性:如 $\{A_i\}_{i=1}^{\infty}$ 是一列事件,满足 $A_iA_j=\Phi(i\neq j)$(称为两两互不相容),则

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

这一公理体系称为**柯尔莫哥洛夫概率论公理体系**,是苏联著名数学家柯尔莫哥洛夫于 1933 年建立的,得到了概率统计学者们的广泛认可,从而为概率论建立了坚实的理论基础。

另外,对于任意两个事件 A、B 且 $P(A)>0$,定义在 A 发生的情形下,B 发生的条件概率为

$$P(B\mid A)=\frac{P(AB)}{P(A)}$$

从而,$P(AB)=P(A)P(B|A)$,这就是**乘法公式**。推而广之,设 $\{A_k\}_{k=1}^{n}$ 是任意 n 个随机事件,则有更一般的乘法公式

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2\mid A_1)P(A_3\mid A_1A_2)\cdots P(A_n\mid A_1A_2\cdots A_{n-1})$$

其成立的证明留作练习。

现设$\{A_i\}_{i=1}^{\infty}$是事件域 F 中的一列事件，若 $\bigcup_{i=1}^{\infty} A_i=\Omega$，且 $A_iA_j=\Phi(i\neq j)$，则称 $\{A_i\}_{i=1}^{\infty}$ 为必然事件 Ω 的一个划分(也称为 Ω 的完全事件组，这里事件的个数也可以是有限多个，比如说 n 个，这相当于 $k>n$ 时都有 $A_k=\Phi$)。显然，任一个事件 A 与其补 $\overline{A}$ 就是 Ω 的一个划分，也是最简单的一个划分。现在设 $\{A_i\}_{i=1}^{\infty}$ 为 Ω 的一个划分且 $P(A_i)>0$，则对任一个事件 $B\in F$ 有**全概率公式**

$$P(B)=\sum_{i=1}^{\infty}P(A_i)P(B\mid A_i)$$

事实上，由

$$B=B\Big(\bigcup_{i=1}^{\infty}A_i\Big)=\bigcup_{i=1}^{\infty}(A_iB)\text{ 且 }(A_iB)\cap(A_jB)=(A_iA_j)B=\Phi,\quad i\neq j$$

利用可列可加性及乘法公式就得

$$P(B)=P\Big(\bigcup_{i=1}^{\infty}A_iB\Big)=\sum_{i=1}^{\infty}P(A_iB)=\sum_{i=1}^{\infty}P(A_i)P(B\mid A_i)$$

现在将全概率公式以及乘法公式应用到条件概率 $P(A_j\mid B)$的公式就有

$$P(A_j\mid B)=\frac{P(A_jB)}{P(B)}=\frac{P(A_j)P(B\mid A_j)}{\sum_{i=1}^{\infty}P(A_i)P(B\mid A_i)},\quad j=1,2,\cdots,n,\cdots$$

这就是著名的**随机事件贝叶斯公式(定理或法则)**，也称为逆概率公式，这里$\{A_j\}$可以认为是事件 B 发生的所有可能的原因，而贝叶斯公式就是计算在已知事件 B 发生的条件下每个原因的可能性大小(即概率)，也就是说由结果去推测原因，因此叫逆概率公式。此外，在这个贝叶斯公式中，$P(A_j)$称为 A_j 的**先验概率**，因为这个概率相对于事件 B 来说是事件发生之前的，而 $P(A_j\mid B)$自然称为 A_j 的**后验概率**。

1.2.2 两例：她怀孕了吗？“非典”时期病人为何要测量体温？

贝叶斯公式与全概率公式都是概率论中的著名公式，在许多学科中都有重要应用，下面我们来看两个例子。

例 1.1 (她怀孕了吗?)根据历史资料知道，女性一次性交后怀孕的概率为 15%。假如一个女性某次性交后怀疑自己怀孕了，但又不能确定。于是，她做了个准确率为 90%的验孕测试，即 90%的怀孕案例会给出阳性反应的检验结果，同时知道该测试当未怀孕时阳性反应占 10%。她当然想知道在检验结果为阳性的条件下的怀孕概率。然而，她不懂贝叶斯统计，所以请你帮助她算出该概率。

解：已知

$P(\text{怀孕})=0.15$，　$P(\text{检测阳性}\mid\text{怀孕})=0.90$，　$P(\text{检测阳性}\mid\text{未怀孕})=0.10$

由已知得，$P(\text{未怀孕})=0.85$。由贝叶斯公式知在检验结果为阳性的条件下的怀孕概率：

$$P(\text{怀孕}\mid\text{检验阳性})=\frac{P(\text{检验阳性}\mid\text{怀孕})P(\text{怀孕})}{P(\text{检验阳性}\mid\text{怀孕})P(\text{怀孕})+P(\text{检验阳性}\mid\text{未怀孕})P(\text{未怀孕})}$$

$$=\frac{0.90\times0.15}{0.90\times0.15+0.10\times0.85}=\frac{0.135}{0.135+0.085}=0.614$$

这里 P(怀孕)$=0.15$ 就是怀孕的先验概率,P(怀孕|检验阳性)$=0.614$ 就是怀孕的后验概率,它是在观察数据(阳性测试)后怀孕概率的更新,表明如果测验呈阳性,则怀孕的可能性大大提高。

例 1.2　("非典"时期病人为何要测量体温?)"非典(SARS)"(发生在 2003 年)患者的主要病症表现为发热、干咳。根据某地区历史资料,已知人群中既发热又干咳的病人患"非典"的概率为 5%;仅发热的病人患"非典"的概率为 3%;仅干咳的病人患"非典"的概率为 1%;无上述病症而患"非典"的概率为 0.01%。现对该区 25 000 人进行检查,发现其中既发热又干咳的病人为 250 人,仅发热的病人为 500 人,仅干咳的病人为 1 000 人,试求:①该区中某人患"非典"的概率;②"非典"患者是仅发热的病人的概率。

解:引入记号

$$A=\{\text{既发热又干咳的病人}\},\quad B=\{\text{仅发热的病人}\},$$
$$C=\{\text{仅干咳的病人}\},\quad D=\{\text{无明显症状的人}\},$$
$$E=\{\text{"非典" 患者}\}$$

易知 A、B、C、D 构成了一个划分。根据对该区 25 000 人进行检查的结果,有

$$P(A)=\frac{250}{25\,000},\quad P(B)=\frac{500}{25\,000},\quad P(C)=\frac{1\,000}{25\,000},$$
$$P(D)=\frac{25\,000-(250+500+1\,000)}{25\,000}=\frac{23\,250}{25\,000}$$

由全概率公式得患"非典"的概率:

$$\begin{aligned}P(E)&=P(A)P(E\mid A)+P(B)P(E\mid B)+P(C)P(E\mid C)+P(D)P(E\mid D)\\&=\frac{250}{25\,000}\times 5\%+\frac{500}{25\,000}\times 3\%+\frac{1\,000}{25\,000}\times 1\%+\frac{23\,250}{25\,000}\times 0.01\%=0.001\,593\end{aligned}$$

由贝叶斯公式知,"非典"患者是仅发热的病人的概率:

$$P(B\mid E)=\frac{P(B)P(E\mid B)}{P(E)}=\frac{\frac{500}{25\,000}\times 3\%}{0.001\,593}=0.376\,647\,8$$

同理,可以算出"非典"患者是既发热又干咳、仅干咳、无明显症状的病人的概率分别为

$$P(A\mid E)=\frac{P(A)P(E\mid A)}{P(E)}=\frac{\frac{250}{25\,000}\times 5\%}{0.001\,593}=0.313\,873\,2$$
$$P(C\mid E)=\frac{P(C)P(E\mid C)}{P(E)}=\frac{\frac{1\,000}{25\,000}\times 1\%}{0.001\,593}=0.251\,098\,6$$
$$P(D\mid E)=\frac{P(D)P(E\mid D)}{P(E)}=\frac{\frac{23\,250}{25\,000}\times 0.01\%}{0.001\,593}=0.058\,380\,41$$

不难看出

$$P(A\mid E)+P(B\mid E)+P(C\mid E)+P(D\mid E)=1$$

而一个人患"非典"时最可能的症状是发热。这就是在"非典"时期动不动就要测量病人体温的原因。

1.2.3 案例：贝叶斯方法在人工智能领域的应用之一

案例 1.1 (自动语音识别——神奇的语音输入法)你的手机里安装了讯飞语音输入法或其他语音输入法了吗？是不是觉得它很神奇呢？想不想知道它为什么能够把你说的话转换为文字呢？这个转换过程其实就是自动语音识别。简单地说，自动语音识别是指由机器自动将语音信号转换为文字的方法和过程。人类的语言可以说是各种信息里最复杂和最动态的一种，著名语言学家乔姆斯基(A. N. Chomsky)和信息论的祖师爷香农 (C. Shannon)等学者都关注过自动语音识别问题，然而那时自动语音识别并没有获得很大进展。在这个领域率先取得突破的是捷克裔美国语音和语言处理大师贾里尼克(F. Jelinek)。从 20 世纪 60 年代开始，贾里尼克开创性地将语音识别问题看成一个通信问题，认为语音识别就是根据接收到的信号序列推测说话人实际发出的信号序列(即说的话)和要表达的意思，并且用贝叶斯公式和两个隐马尔可夫模型建立起统计语音识别系统，把对应的一套模型称为声学模型和语言模型，从而极大地改变了这一领域的研究方向。此外，他还与其他合作者提出了数字通信领域最重要的算法之一——BCJR(L. R. Bahl, J. Cocke, F. Jelinek, J. Raviv, 1974)算法。难能可贵的是，这种统计语音识别系统不但能够识别静态的词库里的语音，而且对动态变化的词库语音具有很好的适应性，即对新出现的词汇，只要这个词已经被高频使用，可用于训练的数据量足够多，系统就能通过训练而正确地识别之。这实际上表明贝叶斯公式对新词汇语音信息有非常好的适应能力。由于本书的性质，这里我们不可能对问题展开详细的讨论，有兴趣者可以去研读有关人工智能的文献资料。但我们从已经开发出来的语音输入法产品知道这种统计语音识别系统是非常成功的！这是贝叶斯方法在人工智能领域的一个重要应用。

1.3 三种信息与先验分布

在 1.1 节中，我们初步了解到统计学中有两个主要学派：经典统计学派与贝叶斯统计学派。在本节，我们从这两个学派使用的信息种类来讨论它们的异同。首先我们来了解统计推断问题中存在的三种信息。

1.3.1 总体与总体信息

我们从已学课程知道统计学中总体就是根据一定的目的和要求所确定的研究对象的全体。例如，如果要统计调查全国大学男生的身高，那么，我们就可以把全国大学男生的集合作为总体，而大学男生身高这个指标就是关于该总体的一个数量，可以用一个符号 X 来标记它。由于在对随机抽出的一个大学男生具体测量之前，并不知道该大学男生的确切身高，而且人的身高是受遗传、营养等随机因素影响而确定的，所以 X 是一个随机变量从而服从某种概率分布。再如，我们要考察一个经济指标 X[可以把它设想为某一只股票的收益率或一个国家的 GDP(国内生产总值)]，由于受各种各样的随机因素的影响，X 是一个随机变量，它的所有可能取值就构成了一个总体，并且也服从某一种概率分布。由于一个随机变量的概率分布完全刻画了该随机变量的统计规律性，因此，我们实际上甚至可以抽象地把这个随机变量的概率分布看作总体。**总体信息**就是我们对总体概率分布的了解或知识，一般而

言，对总体信息最大的了解是知道总体概率分布所属的分布族。例如，若我们知道总体服从正态分布 $N(\mu,\sigma^2)$，虽然这时两个参数还是未知的，我们也知道它的密度函数是一条关于总体均值对称的钟形曲线，并且它的各阶矩都存在，同时也知道第一个参数 μ 是该分布的均值，而第二个参数 σ^2 是该分布的方差。当然，总体到底服从怎样的概率分布族对一个全新研究问题而言通常不得而知，这正是统计学的一个分支——非参数统计所要研究的。显而易见，要获得总体信息往往必须投入大量的人力、物力。例如，美国军队为了获得某种新的电子元件的寿命分布，购买了上万个此种电子元件，做了大量的寿命实验，获得大量数据后才确认其寿命概率分布是什么。简而言之，总体信息非常重要，获得它虽然不容易但又是必须做的，因为它是统计推断的基础。

1.3.2　样本信息

为了对所研究的总体有更多的了解，我们必须从总体抽取（观察或收集）一定的样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$，这些样本给我们提供的信息就是**样本信息**，也称为抽样信息。样本信息两种最重要的表现形式是样本的联合分布与样本统计量的抽样分布，其次是样本对总体特征的各种估计，如样本均值、样本方差（标准差）等。样本是统计学（无论是频率学派还是贝叶斯学派）的粮食，没有样本就如同巧妇难为无米之炊一样，做不成统计学上的任何事情，也就没有统计学了。

仅仅基于总体信息和样本信息进行统计推断的统计学理论和方法称为**经典统计学**。它的历史悠久，但大发展却是 19 世纪末到 20 世纪上半叶。由于统计学家皮尔逊（K. Pearson 1857—1936）、费雪（R. A. Fisher 1890—1962）和奈曼（J. Neyman 1894—1981）等人的杰出工作，经典统计学理论得到空前的发展，成为当时统计学的主流。20 世纪下半叶，经典统计学在工业、农业、医学、经济、金融、管理、军事等领域获得了广泛的应用，并取得了巨大的成功；同时，在这些领域又不断提出新的统计问题，于是又反过来促进经典统计学的进一步发展。但是，伴随着经典统计学的持续发展与广泛应用，它本身的缺陷与某些方面的矛盾之处也逐渐暴露出来了。另外，也存在一些经典统计学难以解决的重要问题。

1.3.3　先验信息与先验分布

先验信息是指在抽样之前对所研究的统计问题的了解或知识，一般来说，先验信息主要来源于研究者的知识和经验以及历史资料（数据），而且常常是零散的，需要提炼加工才可以应用。

先验信息是人们对所研究的统计问题长期观察或研究积累起来的重要历史信息，理应善加利用到统计推断中来，以提高统计推断的质量。从后面的章节我们可以看到，经典统计学由于忽视了先验信息的使用，有时会导致不合理的结论。关于先验信息在帮助人们进行推断的作用，请看下面有趣的例子。

例 1.3　统计学家萨维奇（L. J. Savage，1962）曾考察两个统计实验：

（1）一位常饮奶茶的妇女声称，对于一杯奶茶，她能辨别先倒进杯子里的是茶还是奶。对此做了 10 次试验，她都正确地说出了。

（2）一位音乐家声称，他能从一页乐谱辨别出是海顿（Haydn）还是莫扎特（Mozart）的作品。在 10 次这样的试验中，他都正确辨别了。

现在的问题是被实验者完全是在猜测吗？假如被实验者完全是在猜测，则每次成功的概率为 0.5，那么 10 次都猜中的概率为 $2^{-10}=0.000\,976\,6$，这是一个很小的概率，是几乎不可能发生的，所以假设“被实验者完全是在猜测”是不对的，被实验者每次成功的概率要比 0.5 大得多。换句话说，这不是纯粹的猜测，而是这两位被实验者都有丰富的专业经验，是经验帮助他们作出了正确判断。由此可见，经验(也是一种先验信息)在推断中不可忽视，应善加利用才是正确之举。

例 1.4 (产品质量管理问题)有一句话说得好，“产品质量是企业的生命线”。企业能否生存下去，其产品质量是关键因素之一。我们可以用一个指标来衡量产品质量的高低，那就是不合格品率。为了了解产品的质量，某厂每天都要抽检 5 件产品，以获得不合格品率 θ 的估计。经过 100 个工作日后就积累了大量的数据，通过整理得表 1.1。

表 1.1 产品抽查数据

不合格品	出现次数	频率
0	94	0.94
1	3	0.03
2	2	0.02
3	1	0.01
4	0	0.00
5	0	0.00

根据这些历史资料(就是一种先验信息)，对过去产品的不合格率就可以构造一个概率分布，如表 1.2 所示。

表 1.2 不合格品率先验概率分布

不合格品率 θ	0.0	0.2	0.4	0.6	0.8	1.0
先验概率	0.94	0.03	0.02	0.01	0.00	0.00

这里就是用频率来近似得到先验概率。从表 1.2 可以看出，不合格品率 θ 大于等于 0.2 的先验概率：

$$P(\theta \geqslant 0.2)=0.03+0.02+0.01=0.06$$

是一个相当小的数。

对先验信息进行提炼加工获得的分布就是先验分布。在这个例子中，先验分布(表 1.2)综合了该厂过去产品的质量情况。我们看到这个分布的概率绝大部分集中在 $\theta=0$ 附近。因此，该产品可认为是“信得过产品”。如果以后的多次抽检结果与历史资料提供的先验分布是一致或更好的，质检单位就可以按照要求授予它是“免检产品”，或者每月抽检一两次就足够了，这样，就省去了大量的人力、物力。可见先验信息在统计推断及统计应用中是大有用武之地的。当然，如果以后的多次抽检结果与先验分布有较大的区别，那么我们就应该考虑利用新样本对先验分布进行更新，以期获得更符合实际的新分布——后验分布，这正是贝叶斯统计所要做的重要工作。

基于总体信息、样本信息和先验信息进行统计推断的理论与方法被称为**贝叶斯统计学**。从使用信息的角度看，它与经典统计学的差别在于是否利用先验信息。贝叶斯学派重视先

验信息的收集、挖掘和提炼，并综合先验信息形成先验分布，将其应用到统计推断中来，以提高统计推断的质量。

1.4　一般形式的贝叶斯公式与后验分布

1.4.1　知识准备

首先回忆一下在概率论中有关随机向量和条件分布的几个概念。我们以二维情形为例，设(X,Y)是二维随机向量且分布密度为$f(x,y)$，则X和Y的边际密度分别是

$$f_X(x)=\int_{\mathbf{R}} f(x,y)\mathrm{d}y,\quad f_Y(y)=\int_{\mathbf{R}} f(x,y)\mathrm{d}x$$

其中，$\mathbf{R}$表示实数集，而Y在X已知的条件密度是

$$f(y\mid x)=\frac{f(x,y)}{f_X(x)}=\frac{f(x,y)}{\int_{\mathbf{R}} f(x,y)\mathrm{d}y}$$

从而又有

$$f(x,y)=f(y\mid x)f_X(x)=f(x\mid y)f_Y(y)$$

其次引入高等数学中的两个重要函数：贝塔函数和伽玛函数。它们在贝叶斯统计中经常用到，值得记住。它们分别定义如下：

$$\beta(z,w)=\int_0^1 t^{z-1}(1-t)^{w-1}\mathrm{d}t,\quad \Gamma(z)=\int_0^\infty \mathrm{e}^{-t}t^{z-1}\mathrm{d}t$$

它们有两个重要性质：

$$\Gamma(z+1)=z\Gamma(z),\quad \beta(z,w)=\frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)}$$

第一个性质表明伽玛函数是阶乘$n!=n\cdot(n-1)!$的推广，第二个性质说明贝塔函数和伽玛函数密切关联。

最后，引入一个在贝叶斯统计中常用的分布族，即贝塔分布族$\mathrm{Beta}(a,b)$，其中$a>0$，$b>0$是两个参数。贝塔分布的密度函数如下：

$$\mathrm{Beta}(x\mid a,b)=\frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\quad x\in(0,1)$$

（这意味着x取其他值情形下密度为零）并且具有性质

$$\mathrm{Mode}(X)=\frac{a-1}{a+b-2},\quad E(X)=\frac{a}{a+b},\quad \mathrm{Var}(X)=\frac{ab}{(a+b)^2(a+b+1)}$$

当$a=b=1$时，贝塔分布的密度函数变成$\mathrm{Beta}(x\mid a=1,b=1)=1$，$x\in(0,1)$，这正是均匀分布$\mathrm{U}(0,1)$的密度，所以均匀分布$\mathrm{U}(0,1)$是一个特殊的贝塔分布。图 1.1 为贝塔分布族在四组参数值下的密度函数曲线，我们看到在不同的参数值下密度函数曲线变化很大。

1.4.2　编程语言 R 与其软件包

本书从下一小节开始就要求读者用 R 软件进行统计计算和作图，并把这一要求贯穿全书，目的是通过动手使用软件让读者培养起自己的数据感和体验研读贝叶斯统计的乐趣，从而激发起对贝叶斯统计学的兴趣。

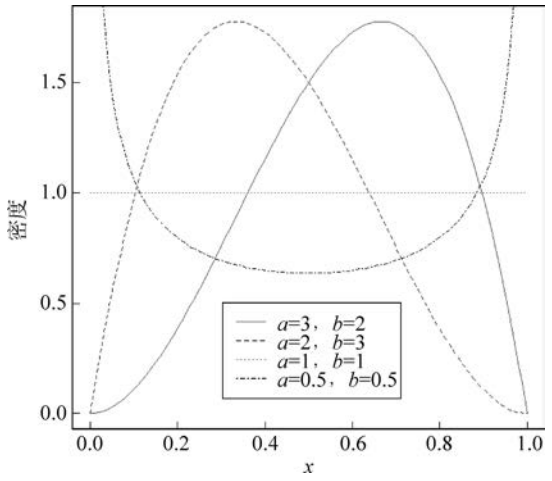

图 1.1 贝塔分布族在四组参数值下的密度函数曲线

"R you ready for R?"这是国外高校校园里一句时髦的问句,它表明了 R 语言在国外高校盛行的程度。那么 R 到底是何方神圣而在校园如此盛行呢? R 是著名的贝尔实验室(Bell Laboratories)的编程语言 S 的实现版,最初的两位设计者是当时任教于新西兰奥克兰大学的 Ross Ihaka 教授和 Robert Gentleman 教授。由他们的名字拼写,大家可以看出这套软件系统叫 R 的原因了。现在 R 由其核心团队负责维护和发展,每半年左右会更新一次。R 是用于统计计算和绘图的编程语言与软件环境; R 是一个自由、免费、源代码开放的软件包; R 是一套完整地用于数据处理、统计计算和制图的软件系统。R 的功能还包括:数据的输入、输出以及存储;数组运算(其数组种类丰富,向量、矩阵运算功能尤其强大)。由于全球各个领域学者的贡献,R 有成千上万用于不同领域的软件包,但它的基本包为 base,我们可从其官网或官网镜像,如 https://cloud.r-project.org/,下载并安装,本书安装的版本是 R-4.2.1-win。由于基本包 base 实际上还包括 stats 和 graphics 等诸多包,所以安装好 base 后,我们不但可以进行各种算术计算,也可以进行通常的统计计算(建模)和绘图了。

为了方便初学者的学习和实践,本书制作了一个专用 R 包——BayesianStat(但进行了压缩),并把书中所有案例数据和主要程序都放入此包,可通过扫描本书后面的二维码进行下载。读者免费下载此包到自己的电脑后,解压就得到文件夹 BayesianStat,然后将此文件夹复制到安装好的基本包 base 所带文件夹 library 中即可应用,library 文件夹的一个路径示例是

C:\Program Files\R\R-4.2.1\library

如你用的是苹果 Mac OS 操作系统,如苹果笔记本,同样可以安装与使用这个 R 包。如你已装旧版的,则要先删除,再安装这个新版。从现在开始,我们就要充分利用 R 软件来进行贝叶斯统计的学习了。

在编程和使用计算机软件的时候,要特别注意计算机中绝对值最大的数和绝对值最小的非零数都是有界的。例如,从数学来说,

$$\log(0.01^{180}) \equiv 180 \times \log(0.01)$$

但是,在用 R 软件计算上式左右两边时,我们得到

```
log(0.01^180)
[1] - Inf
180 * log(0.01)
[1] - 828.9306
0.01^180
[1] 0
```

这就是说，上式左边的计算结果是负无穷大，而右边计算结果是－828.930 6，同时不难知道后者才是正确的。左边的计算之所以会出现错误的结果，是因为 0.01^180 的计算结果已经小到计算机把它等同于零了。所以，对这个数学上的恒等式，在计算机上只能用右边的表达式来计算。

另外，第三方 R 包的常用安装法有两种：菜单法和命令法。例如，我们要用菜单法安装 R 包 mvnormtest，那么，在打开 R 的控制台 R console 后，我们可以在左上角看到“程序包”菜单，单击它，进入菜单并看到好几个子菜单，首先单击子菜单“设定 CRAN 镜像”，则会弹出一个小视窗，在里面挑选一个离你最近的镜像网站并单击它，则确定了镜像网站。然后，再次打开“程序包”菜单，单击子菜单“安装程序包”，则同样会弹出一个小视窗，里面有上万的第三方 R 包，找到我们要安装的 R 包 mvnormtest，单击它，那么就会下载安装了。如果用命令法安装，则是在控制台 R console 内、提示符“＞”后，敲入命令并执行（在 R 编程时，应使用英文输入法，否则容易出错）：

```
install.packages("mvnormtest", repos = "https://cloud.r-project.org/")
```

那么就会下载并安装好 mvnormtest。最后，为了检查是否真安装好了，可以敲入命令并执行：

```
library(mvnormtest)          # 将 R 包 mvnormtest 引入控制台 R console
```

如果命令顺利执行不出错，则说明安装成功。这是使用任何一个第三方 R 包都必须首先执行的命令。

1.4.3　一般形式的贝叶斯公式

现在我们要对一个总体 X 进行统计推断，假设其分布密度为 $p(x|\theta)$，其中，θ 是未知参数，之所以写成条件密度的形式是因为在贝叶斯统计中未知参数 θ 被看成随机变量了。进一步，假设参数 θ 已经有了先验分布 $\pi(\theta)$，而且从总体 X 那里得到了新样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$。现在的问题是怎样利用样本对先验分布 $\pi(\theta)$ 进行更新，以期得到更适当的分布。我们知道样本信息综合体现在其联合分布密度 $p(\boldsymbol{x}|\theta)$ 中，而且如果样本是简单随机样本，则

$$p(\boldsymbol{x}\mid\theta)=\prod_{i=1}^{n}p(x_i\mid\theta)$$

现在假设更新后的分布是 $\pi(\theta|\boldsymbol{x})$，即 θ 的以样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 为已知条件的分布。根据条件密度的公式，$\pi(\theta|\boldsymbol{x})$ 可以写成

$$\pi(\theta\mid\boldsymbol{x})=\frac{h(\boldsymbol{x},\theta)}{m(\boldsymbol{x})}$$

其中,$h(\boldsymbol{x},\theta)$是 $\boldsymbol{x}$ 和参数 θ 的联合密度,$m(\boldsymbol{x})$是 $\boldsymbol{x}$ 的边际密度,而且

$$m(\boldsymbol{x})=\int_{\Theta}h(\boldsymbol{x},\theta)\mathrm{d}\theta \quad (\Theta \text{ 是参数空间})$$

另外,利用先验分布 $\pi(\theta)$和样本的分布密度 $p(\boldsymbol{x}|\theta)$,我们可得样本 $\boldsymbol{x}$ 和参数 θ 的联合密度

$$h(\boldsymbol{x},\theta)=p(\boldsymbol{x}\mid\theta)\pi(\theta)$$

于是

$$\pi(\theta\mid\boldsymbol{x})=\frac{h(\boldsymbol{x},\theta)}{m(\boldsymbol{x})}=\frac{p(\boldsymbol{x}\mid\theta)\pi(\theta)}{\int_{\Theta}p(\boldsymbol{x}\mid\theta)\pi(\theta)\mathrm{d}\theta}$$

显而易见,这个公式把总体信息、样本信息和先验信息都综合进去了。这就是**密度函数形式的贝叶斯公式**(定理或法则),其中,$\pi(\theta|\boldsymbol{x})$被称为 θ 的后验分布,它是集中了总体、样本和先验三种信息后对于先验分布 $\pi(\theta)$的更新,以期得到参数 θ 更符合实际的分布。贝叶斯公式是整个贝叶斯统计学的奠基石,初看简单,其实复杂,值得在理解的基础上记住!

如果 θ 是离散参数,其先验分布可用先验分布列$\{\pi(\theta_j)\}$来表示,则后验分布也是离散形式,而且容易得到

$$\pi(\theta_j\mid\boldsymbol{x})=\frac{p(\boldsymbol{x}\mid\theta_j)\pi(\theta_j)}{\sum_i p(\boldsymbol{x}\mid\theta_i)\pi(\theta_i)},\quad j=1,2,\cdots$$

这个公式与事件形式的贝叶斯公式是何其相似!

注:

(1) 从贝叶斯公式显而易见,无论是样本分布 $p(\boldsymbol{x}|\theta)$还是先验分布 $\pi(\theta)$,乘以一个非零常数都不会改变后验分布 $\pi(\theta|\boldsymbol{x})$。

(2) 当得到样本观察值 $\boldsymbol{x}$ 后,样本分布密度 $p(\boldsymbol{x}|\theta)$也就是熟知的似然函数,并常常记为 $l(\theta)=l(\theta|\boldsymbol{x})=p(\boldsymbol{x}|\theta)$以表明这是 θ 的函数。因此,$p(\boldsymbol{x}|\theta)\pi(\theta)=l(\theta|\boldsymbol{x})\pi(\theta)$。

(3) 先验分布 $\pi(\theta)$当然也有参数(如 λ),但是在这里假定它已知了,所以没有写出来。如果它未知或为了强调而写出来,那就是 $\pi(\theta)=\pi(\theta|\lambda)$,并且我们称先验分布中的参数为**超参数**。

(4) 这里对贝叶斯公式的解释是从经典统计的视角出发的,但是,也可以从其他视角来解释该公式。此处的要点是,一个事件发生了,那么是什么原因促使它发生的。贝叶斯公式给出了计算原因概率大小的一种方法。

1.4.4 计算后验分布的例

在本小节,我们通过例子来加深对贝叶斯公式的理解。

例 1.5 (例 1.4 续)该工厂为了进一步改善产品质量,采用了更先进可行的技术,不合格品率 θ 因此有可能发生变化。为了对 θ 的先验分布进行更新,我们来计算 θ 的后验分布。为此,我们对 n 件产品进行独立检测,不合格品出现的个数记为 X,显然,X 服从二项分布 $\mathrm{Bin}(n,\theta)$,即

$$P(X=x\mid\theta)=p(x\mid\theta)=C_n^x\theta^x(1-\theta)^{n-x},\quad x=0,1,\cdots,n$$

再根据贝叶斯公式和 θ 的先验分布(表 1.2),我们就可以把 θ 的后验分布算出来,其一般表达式是

$$\pi(\theta_j \mid x)=\frac{p(x \mid \theta_j)\pi(\theta_j)}{\sum_i p(x \mid \theta_i)\pi(\theta_i)},\quad x=0,1,2,\cdots,n;\ j=1,2,\cdots,6$$

在 R 平台中利用如下命令就可以把以二项分布 Bin(n,θ)为总体、参数 θ 为离散情形的后验概率分布具体算出来，例如，若 $n=10$，$x=0$，则可以算得相应的后验概率分布表 1.3。从该表可以看出，通过采用新技术，产品质量有了很大的提高。为了理解整个计算过程，请读者手工计算出 $\theta_1=0.0$ 的后验概率。以下就是所用的 R 命令：

```
library(BayesianStat)    # 计算后验概率的命令 Bindiscrete 在此包中
theta<-c(0,0.2,0.4,0.6,0.8,1)
prior<-c(0.94,0.03,0.02,0.01,0.00,0.00)
Bindiscrete(x=0, n=10, pi=theta, pi.prior=prior, n.pi=6)
```

表 1.3　不合格品率后验概率分布表

不合格品率 θ	0.0	0.2	0.4	0.6	0.8	1.0
后验概率	0.996 5	0.003 4	0.000 1	0.000 0	0.000 0	0.000 0

这里，参变量 x 是样本值；n 是样本量；pi 是不合格品率 θ 的取值向量；pi.prior 是 θ 的先验概率向量；n.pi 是 θ 的取值个数。另外，最后这个命令可以同时得到先验概率与后验概率的比较图(图 1.2)。该图形象地把后验概率相对于先验概率的变化表示出来，从该图可以看出不合格品率 $\theta=0$ 的后验概率比先验概率大，而其他情形的后验概率都不大于先验概率，这就更生动形象地说明了产品质量有了很大的提高。

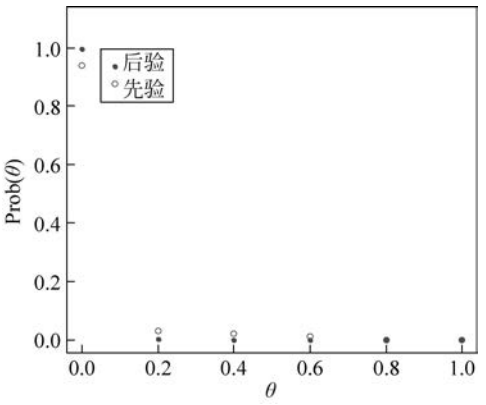

图 1.2　不合格品率先验与后验概率比较

现在假设在检测该产品之前我们对不合格品率 θ 没有任何先验信息(比如，这是新产品)。在这种情况下，贝叶斯建议用区间(0,1)上的均匀分布 U(0,1)作为 θ 的先验分布，因为该分布在区间(0,1)上机会均等地取到每一点。贝叶斯的这个建议被后人称为**贝叶斯假设**。这时 θ 的先验分布密度为

$$\pi(\theta)=\begin{cases}1, & 0<\theta<1\\ 0, & \text{其他场合}\end{cases}$$

于是，样本 X 与参数 θ 的联合分布

$$h(x,\theta)=p(x \mid \theta)\pi(\theta)=C_n^x\theta^x(1-\theta)^{n-x},\quad x=0,1,\cdots,n;\ 0<\theta<1$$

而 X 的边缘分布

$$m(x)=C_n^x\int_0^1\theta^x(1-\theta)^{n-x}\,\mathrm{d}\theta=C_n^x\,\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}=\frac{1}{n+1},\quad x=0,1,\cdots,n$$

利用贝叶斯公式，最后可得 θ 的后验分布

$$\pi(\theta \mid x)=\frac{h(x,\theta)}{m(x)}=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1},\quad 0<\theta<1$$

这正是参数为 $x+1$ 和 $n-x+1$ 的贝塔分布 Beta($x+1$,$n-x+1$)。

在 R 平台中利用如下命令就可以把以二项分布 Bin(n,θ)为总体,参数 θ 服从贝塔分布的先验密度和后验分布密度图形画出来(图 1.3)。

```
library(BayesianStat)
Binbeta(x, n = 10, a = 1, b = 1, pi = seq(0.01, 0.999, by = 0.001), plot = TRUE)
```

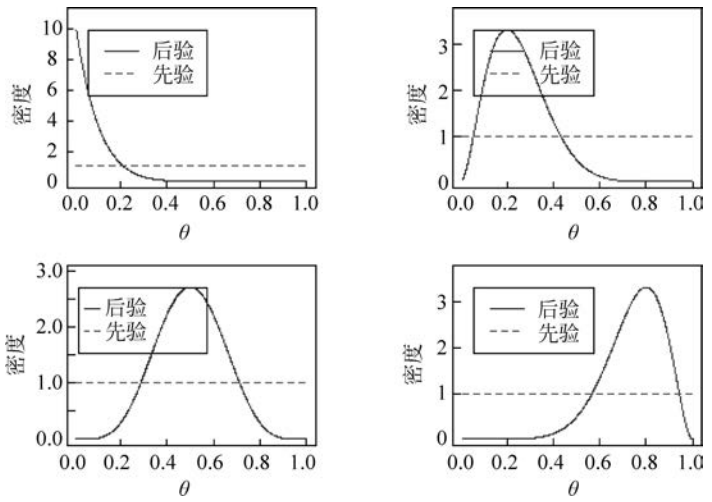

图 1.3 均匀分布先验与二项分布形成的后验分布密度图

在函数 Binbeta 中,参变量 x 是样本值;n 是样本量;a 和 b 是贝塔分布的两个参数(在本例中,因为先验是(0,1)区间上的均匀分布,所以 $a=b=1$);pi 是不合格品率 θ 的取值向量;plot 是逻辑变量(取"TRUE"表示要作图;取"FALSE"表示不要作图)。注意:这里样本量(产品抽取个数)$n=10$,按照从左到右、从上到下的顺序各图对应的样本值分别是 $x=0$,$x=2$,$x=5$,$x=8$。从图 1.3 可以看出,随着样本值的变化,后验密度曲线也发生了重大变化。换句话说,样本对先验分布产生了重大影响,先验被实质性更新了。

1.4.5 案例:贝叶斯方法在人工智能领域的应用之二

案例 1.2 (朴素贝叶斯分类器,本案例只要求了解之)无论是在科技领域还是在日常生活中,对事物进行分类都是常做的事。例如,每个使用电子邮件的人都知道在电邮中常常会收到垃圾邮件。如果电邮系统较好,那么收到垃圾邮件的数量就较少,因为垃圾邮件绝大部分都被电邮系统通过分类识别或者说过滤出来而排除了。在分类这个领域有一种分类方法就是**朴素贝叶斯分类器**(Naive Bayes Classifier)。它是基于贝叶斯定理而建立起来的一套统计机器学习分类算法,是一种有监督分类方法(supervised classification),也是一种虽然简单但很有效的贝叶斯分类方法。

具体而言,设 Y 是一个**类变量**(class variable,对应于经典统计的响应或因变量),共有 K 个类别为 y_1,y_2,…,y_K(在机器学习中,它们被称为标签,Y 在一次观察中取一个标签为值)。例如,对于电邮过滤问题,Y 可以有两个标签(正常邮件,垃圾邮件)。而 $\boldsymbol{X}=(X_1$,

$X_2,\cdots,X_n$)是个**特征**或**属性向量**(feature or attribute vector),其各个分量从各自的侧面描述或者说解释类变量 Y,被称为**特征变量**(对应于经典统计的独立或自变量)。如果我们得到一个 $\boldsymbol{X}$ 的样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$,那么,由贝叶斯公式,$(Y=y)$的后验概率

$$p(y\mid\boldsymbol{x})=\frac{p(x_1,x_2,\cdots,x_n\mid y)p(y)}{p(x_1,x_2,\cdots,x_n)}\propto p(x_1,x_2,\cdots,x_n\mid y)p(y)$$

其中,符号∝表示正比于,因为在 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$给定后边际 $p(x_1,x_2,\cdots,x_n)$是常数(详情可见本书第 2 章)。为了能够较容易地计算出这个后验概率,我们"天真地"(naive)假设在已知类别的条件下,特征向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$的分量是相互独立的,并称这一性质为**类条件独立性**,于是后验概率

$$p(y\mid\boldsymbol{x})\propto p(x_1,x_2,\cdots,x_n\mid y)p(y)=p(y)\prod_{i=1}^{n}p(x_i\mid y)$$

如果 y_{k^*} 满足

$$y_{k^*}=\arg\max_y\left[p(y)\prod_{i=1}^{n}p(x_i\mid y)\right]$$

即$(Y=y_{k^*})$的后验概率最大,则认为相应的对象(比如说,一封电邮)属于第 k^* 类,或者说,预测结果是第 k^* 类。显然,为了计算后验概率,还要知道 $p(x_i\mid y)$的概率分布(称为**类条件分布**)。常用的有正态分布 $\mathrm{N}(\mu,\sigma^2)$、伯努利分布 $\mathrm{Bernoulli}(\theta)$、多项分布 $\mathrm{Multinom}(n;\theta)$(自然包括二项分布)等。另外,为了简化和更精确地计算,往往取对数后再来计算,于是

$$y_{k^*}=\arg\max_y\left[\log p(y)+\sum_{i=1}^{n}\log p(x_i\mid y)\right]$$

下面我们来看一个具体的案例。在 R 包 BayesianStat 中有两个文本数据集[①],一个为 trainset.csv,另一个为 testset.csv。它们都是一些英文帖子(推文),其中一部分帖子是关于 Data Science 的,其余的与 Data Science 无关。此外,在样本集 trainset.csv 中,一个帖子是不是关于 Data Science 的已经知道,并且当它是关于 Data Science 的时被标记为"1",否则,被标记为"0"。这样,trainset.csv 由三列构成:第一列 Data_Science 是帖子的标签;第二列 Date 是帖子发表的日期;第三列 Tweet 是帖子文本。类似这样的数据集在机器学习中被称为**训练集**(train set),它是用来训练即估计模型的(训练的过程就是学习)。它的前六行可从下面的 R 程序中见到。另一个文本数据集 testset.csv 的第二、三列也分别是帖子发表的日期、帖子文本,但第一列 id 是识别码,也就是序号。类似这样的文本数据集在机器学习中被称为**测试集**(test set),它是用来测试并评估模型的。其最后三行同样可从下面的 R 程序中见到。

本案例分析的过程是,首先训练模型,然后测试评估模型,具体步骤如下:

第一步,安装要用到的 R 程序包;

第二步,引入训练集并清理之,即把不影响判断帖子属性的一些符号、词语片段删除;

第三步,将文本数据转化为数值特征(这一过程被称为**文本数据的向量化**,其实得到的结果是一个 324×1 837 阶稀疏数值矩阵);

第四步,将标签列因子化;

① 原始数据来自 https://dx.doi.org/10.6084/m9.figshare.2062551.v1。

第五步,训练(估计)模型,这里选用的是多项分布朴素贝叶斯分类模型;

第六步,引入测试集并清理之;

第七步,将测试集中的文本数据转化为数值特征矩阵;

第八步,测试并评估训练所得模型。

从下面的 R 程序我们看到,模型预测得到测试集的最后六条帖子对应的标签分别是 0,0,1,1,0,1,即倒数第四、三、一条帖子是关于 Data Science 的,而另三条帖子则不是。这里我们感兴趣的是,那事实上到底是不是如此呢?我们将这六条帖子显示出来(在下面 R 程序的最后),然后看看它们的内容并判断一下,我们会毫无争议地得到一样的结论。这就是说,所得模型对帖子的分类能力相当优秀。

```
install.packages('naivebayes', repos = "https://cloud.r-project.org/")
install.packages('mgsub')
install.packages('text2vec')
library(BayesianStat)
pat <- R.home("library/BayesianStat/data/trainset.csv")
train <- read.csv(file = pat, header = TRUE)     #引入训练集
head(train)
 Data_Science      Date                                                    Tweet
1            0  11/04/15                  Oh... It is even worse... They are playing …
2            1  11/12/13        … Mavericks Issues Resolved http://t.co/i7qAPuR8EN
3            0  02/03/14             … this stellar artwork http://t.co/IYnxU8FSVS
4            0  25/02/14                        … But I was happy not carrying a tube
5            1  16/12/13             … year in review, 2013 http://t.co/NO7MVgMSpK
6            1  13/11/13        … guide to memory usage in R http://t.co/kxU5kS2sHw
library(mgsub)
# 删除所有类似 http://t.co/i7qAPuR8EN 的片段
train$Tweet <- mgsub(train$Tweet, "http\\S+", "")
head(train)
 Data_Science      Date                                                    Tweet
1            0  11/04/15                 Oh... It is even worse... They are playing …
2            1  11/12/13                      RStudio OS X Mavericks Issues Resolved
3            0  02/03/14           A Hubble glitch has produced this stellar artwork
4            0  25/02/14                        … But I was happy not carrying a tube
5            1  16/12/13                Data and visualization year in review, 2013
6            1  13/11/13                         A detailed guide to memory usage in R
train$Tweet[319]
[1] "There you go! #AusvArg is on! As much as I would love Argentina to be in the final, I am
supporting #Australia #OfficeSweepstake"
# 删除所有帖子中的#符号
train$Tweet <- mgsub(train$Tweet, "#", "")
train$Tweet[319]
[1] "There you go! AusvArg is on! As much as I would love Argentina to be in the final, I am
supporting Australia OfficeSweepstake"
library(text2vec)
xtrain <- train$Tweet
it <- itoken(xtrain, preprocess_function = tolower, tokenizer = word_tokenizer)
v <- create_vocabulary(it)
vectorizer <- vocab_vectorizer(v)
```

```
xtrain_dtm <- create_dtm(it, vectorizer)
dim(xtrain_dtm)
[1]  324 1837
ytrain <- train$Data_Science
yf <- factor(ytrain)
library(naivebayes)
mnbMod <- multinomial_naive_bayes(x = xtrain_dtm, y = yf)
pat <- R.home("library/BayesianStat/data/testset.csv")
test <- read.csv(file = pat, header = TRUE)      #引入测试集
tail(test,3)
     id       Date                                                          Tweet
161 161   2002/9/14          knitr in a knutshell tutorial http://t.co/ixSQOifbBK
162 162  2011/12/13    … to get data, a music video parody http://t.co/sILKbdfB2H
163 163   2021/2/14         … with R, from Graham Williams http://t.co/x3683TyF9q
test$Tweet <- mgsub(test$Tweet, "http\\S+", "")
test$Tweet <- mgsub(test$Tweet, "#", "")
xtest <- test$Tweet
itest <- itoken(xtest, preprocess_function = tolower, tokenizer = word_tokenizer)
xtest_dtm <- create_dtm(itest, vectorizer)
mnbPredClas <- predict(mnbMod, newdata = xtest_dtm, type = "class")
mnbPredProb <- predict(mnbMod, newdata = xtest_dtm, type = "prob")
tail(mnbPredClas)
[1] 0 0 1 1 0 1
Levels: 0 1
test$Tweet[158:163]                              #测试集的最后六条帖子
[1] "The Martian Is Hands Down The Best Thriller Of The Year, the book is great "
[2] "Halfpenny ruled out of World Cup rugby  "
[3] "CRAN now has 5000 R packages "
[4] "knitr in a knutshell tutorial "
[5] "Up all night to get data, a music video parody "
[6] "A survival guide to Data Science with R, from Graham Williams "
```

类条件独立性这一假设确实是天真的，因为在实际应用中的大多数场合，这个假设其实并不成立。然而，令人惊喜的是，在类条件独立性假设下，不但计算大大化简，而且朴素贝叶斯分类器得到的分类结果在许多场合是相当好。此外，朴素贝叶斯分类器还有一大优点，就是与其他更复杂的方法相比，朴素贝叶斯分类器计算速度非常快，这对于分类大规模的文本等数据是必需的。正因为有这些优势，朴素贝叶斯分类器已被广泛应用于许多领域。

本章要点小结

本章简要介绍了贝叶斯统计学的历史、现状以及思想方法，并将它与经典统计学进行了初步的比较，同时也分析了贝叶斯方法在人工智能领域应用的两个案例。重点是如下三种形式的贝叶斯公式（定理）及其含义：

$$P(A_j \mid B)=\frac{P(A_jB)}{P(B)}=\frac{P(A_j)P(B \mid A_j)}{\sum_{i=1}^{\infty}P(A_i)P(B \mid A_i)} \quad j=1,2,\cdots,n,\cdots$$

$$\pi(\theta_j \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \theta_j)\pi(\theta_j)}{\sum_i p(\boldsymbol{x} \mid \theta_i)\pi(\theta_i)}, \quad j = 1, 2, \cdots$$

$$\pi(\theta \mid \boldsymbol{x}) = \frac{h(\boldsymbol{x}, \theta)}{m(\boldsymbol{x})} = \frac{p(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int_\Theta p(\boldsymbol{x} \mid \theta)\pi(\theta)\mathrm{d}\theta}$$

它们是整个贝叶斯统计学的基础和重中之重。

思考与练习

1.1 统计推断可能用到哪三种信息？如何界定经典统计和贝叶斯统计？

1.2 简要陈述贝叶斯统计的思想和历史。

1.3 一些著作把贝叶斯的开山之作“*An Essay Towards Solving a Problem in the Doctrine of Chances*”中的术语 chances 翻译为“机遇”而不是“概率”。你认为是否正确并说出你的依据(提示：从互联网上下载此论文并浏览一下)。

1.4 设$\{A_k\}_{k=1}^n$是任意 n 个随机事件，证明更一般的乘法公式

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2)\cdots P(A_n \mid A_1A_2\cdots A_{n-1})$$

1.5 用自己的语言总结各种形式的贝叶斯公式。

1.6 试分别定义先验分布和后验分布。

1.7 有一种前列腺癌标记(prostate cancer marker，PSA)检测法的特性如下：如果成年男性犯了前列腺癌，则检测结果呈阳性的概率高达 90%，同时，如果成年男性未犯前列腺癌，则检测结果呈阳性的概率为 5%。现在某大学男生进行前列腺癌标记检测，结果呈阳性。虽然他根据自己的了解知道大学男生犯前列腺癌的概率只有 0.001%，但还是非常害怕，很想知道他确实犯前列腺癌的概率，遗憾的是他没有学好贝叶斯统计学，所以请你赶快帮助他算出这个概率。另外，根据计算出的概率，你对该大学男生有什么建议？

1.8 为研究产品质量，我们从一批产品中抽取 8 个产品进行检验，结果发现 3 个不合格品。现设 θ 是这批产品的不合格率，并且先验分布有两种情形为

$$\theta \sim \mathrm{U}(0,1)$$

$$\theta \sim \pi(\theta) = \begin{cases} 2(1-\theta), & \theta \in (0,1) \\ 0, & \theta \notin (0,1) \end{cases}$$

试分别求 θ 的后验分布。

1.9 手工计算例 1.5 中当 $n=10, x=0$ 时，$\theta_1=0.0$ 的后验概率。

1.10 用 R 命令计算例 1.5 中当 $n=10, x=1$ 且先验概率为表 1.2 时的后验概率并说明结果的意义。

1.11 用 R 命令做出例 1.5 中当 $n=10$ 且先验为均匀分布 U(0,1)时的先验密度和后验密度图并加以解释(x 分别取 1,3,4,6,9,10)。

1.12 为了提高某产品的质量，公司经理考虑改进生产设备，预计需投资 90 万元，但从投资效果看，下属部门有两种意见：

(1) θ_1：改进生产设备后，高质量产品可占 90%。

(2) θ_2：改进生产设备后，高质量产品可占 70%。

但经理根据过去的经验认为，θ_1 的可信程度只有 40%，θ_2 的可信程度是 60%，即

$$\pi(\theta_1)=0.4,\quad \pi(\theta_2)=0.6$$

经理不想仅仅用过去的经验来做决策，因此通过小规模试验观其结果再定夺，为此做了一项试验，试验结果如下：

$$A=\{\text{试制 5 个产品，全是高质量的产品}\}$$

经理对这次试验结果很高兴，希望用此试验结果来修改他原先对 θ_1 和 θ_2 的看法，即要去求后验概率 $\pi(\theta_1|A)$与 $\pi(\theta_2|A)$。如今已有先验概率 $\pi(\theta_1)$和 $\pi(\theta_2)$，还需要两个条件概率 $P(A|\theta_1)$与 $P(A|\theta_2)$，这可用二项分布算得

$$P(A\mid\theta_1)=0.9^5=0.590,\quad P(A\mid\theta_2)=0.7^5=0.168$$

由于经理没有学过贝叶斯统计，请你帮助将后验概率 $\pi(\theta_1|A)$与 $\pi(\theta_2|A)$算出来（虽然其下属已经帮助计算过）。

经过实验 A 后，更新的概率使经理对增加投资以改进质量的兴趣增大，但是为了慎重起见，他还想再做一次小规模的试验，观其结果再做决策。此次试验结果如下：

$$B=\{\text{试制 10 个产品，9 个是高质量产品}\}$$

经理希望用此试验结果对 θ_1 与 θ_2 再做一次更新，为此把上次试验的后验概率看作这次的先验概率，即

$$\pi(\theta_1)=0.7,\quad \pi(\theta_2)=0.3$$

用与上次试验同样的方法，请你再帮助经理把新的后验概率算出来。观察新后验概率后，你会向经理提什么建议？最后你对贝叶斯方法有什么新的认识？