数据贵过黄金?

Joseph Malkevitch ／文　　曾铁勇／译

在不同的时期，不同的地方，黄金一直备受人类社会的推崇。2011 年，某些地方黄金的价格每盎司突破了 1900 美元。多年来，黑金一直是对石油（目前每桶超过 100 美元）的重要性和价值的形象比喻。但是数据有可能成为未来真正有价值的商品吗？如果是这样，这与数学有什么关系呢？

2012 年的四月是数学宣传月，这一年的宣传主题是：数学、统计和海量数据。

本文的目的是要探讨短语“数据挖掘”的含义，以及为了促进这个领域的发展采用了哪些数学工具和思想。数据挖掘的起源之一是减小数据的占用空间，从而产生了对收集到的数据进行分析的需求。本文将给出一个非数学问题引起数学界关注后，相关的数学又如何发展的经典例子。人们不止一次看到，过去所发展的数学，只因为其“美”及智力上的吸引力，常常是洞悉新的应用环境的工具。在描述问题的背景后，我将给出一些引起广泛关注的和数据相关的例子，并探讨涉及数据挖掘的“人工智能”的方法。

无处不在的数据

什么是“数据”？作为多重领域使用的术语，数据最常代表的是“事实”或数值形式的统计。然而，有时这个术语又被用来表示将要被分析或用来做决断的信息，最近出现的对于数据的“定义”又涉及到用计算机处理或分析的数字、符号、字符串和表格。

虽然数据长伴我们，但利用数据进行更深入的探索并（或）影响决策的想法则相对较新。几何和代数之根源可上溯千年，而“数据”数学则是近代产物，在过去的 150 年间发展迅猛。产生这种现象的部分原因是人们需要计算或分析大规模数据以获得重要信息，但这往往非常耗时，且受制于人为错误。因此，人们很自然地求助计算机来大规模采集和分析数据，以达到快速和准确的目的。

在看到不因噪音或偶然测量误差而出现的

模式这一意义下，要了解数据，需要懂得概率论。概率论和统计学是在许多方式下共存的两个科目。然而，概率是一个困难与微妙的学科。尽管概率论的数学基础相当坚实，但在数学之外的学科中用概率来数学建模，却异常复杂。当我们给出一个陈述：一枚硬币出现正面的机会稍大于出现背面的的机会，比如假设出现正面的概率为 0.501，而出现背面的概率为 0.499，如何诠释这个陈述呢？如果在投掷方式不影响硬币出现正面或者背面的情况下（注意：某些技高一筹者可以多次投掷一枚硬币使得每次都出现正面）多次投掷硬币，将得到一个关于出现 H（代表正面）和 T（代表背面）的模式。例如，投掷一枚硬币 10 次而得到下面的模式：

TTHHHTTTTH

现在，基于这个有限的、很小的硬币投掷集合，出现正面的相对频率为 4/10（10 次投掷出现 4 次正面），而出现背面的相对频率为 6/10（10 次投掷出现 6 次背面）。由此你也许看到了这里出现的一个困难：对于任何一次固定数量的投掷，甚至是一次非常大数量的投掷，得到出现正面的相对频率为 0.501 和出现背面的相对频率为 0.499 的情况十分罕见！概率的“稳定的相对频率”的解释为：从长远来看，随着投掷次数变得越来越多，得到正面和背面的相对频率将分别为 0.501 和 0.499。不幸的是，似乎没有方法使得产生这种论点背后的直觉很“严谨”。除此之外，有些说法比如，这个电厂在未来 10 年将会发生核事故的概率为 0.00000001，或者明天在波士顿某些地区将会下雨的概率为 4/10 都没有很清晰的含义。

现代概率的观点是概率是受制于某些规则（公理）的系统，概率“直观的”性质产生于这个系统。这些规则意味着当有人使用“相对频率”的观点时，在一定意义上不会被误导。然而，这些年出现了其它的途径来“解释”概率的含义。显然，对于滚动骰子、投掷硬币及孩子的出生性别模式，可以较合理地理解概率意义下稳定的相对频率。然而，对于一个核电厂的燃料棒在未来 10 年内发生熔毁的可能性，概率的相对频率意味着什么呢？出现这类事件的历史非常短，所以对其稳定的相对频率的理解变得毫无意义。即使对于天气报告中经常出现的预报，比如明天下雨的可能性（概率）是 80% 的预报，我们应该如何理解呢？

因数学家长期纠结于对概率这个概念所赋予的意义，导致了许多不同的“矛盾的”观点的出现。有鉴于此，概率和统计学的杰出贡献者伦纳德 • 萨维奇（Leonard Savage, 1917-1971）指出：“众所公认，统计学在某种程度上依赖于概率。但是，对于概率是什么以及它如何与统计学相联系，很少像今天如此激烈地争论并产生完全不同的观点。”

伦纳德 • 萨维奇（1917-1971）

问题之一是如何使用共同的语言（无论是英语、法语等）来表达不同环境下涉及到的“噪声”、“随机性”、“可能性”，或者是“意外”。放射性衰变机制显然不同于龙卷风会将袭击哪个州，或将会在哪天袭击这样的问题。

目前对于概率意义有多种诠释。作为一个例子，概率论的一种解释涉及到可根据经验获得的知识来帮助主观判断。这种方法也称为贝叶斯概率（即使是这个术语也存在几种不同的版本），它试图量化当和目标事件相关的某事件发生的概率正知后，如何来修正目标事件发生的概率。由这种角度产生的“概率”与数学家的正规方法产生的“概率”遵循相同的基本规则。然而，基于概率的不同表示所得到的推理方法也会有所不同。因此，会存在一些情况，因为采用不同的概率表示方法，做决断时就要从不同的可能中选择一种。估计这个复杂的议题将长期被数学家、统计学家和哲学家大范围地讨论。

概率和统计的先驱

在数学中，重要的思想凭空出世是非常罕见的。许多国家的科学家都为概率和统计的发展做出了贡献。本节将简要介绍一小部分重要贡献者。毫无疑问，概率论中相对频率的早期先驱之一是法国哲学家和数学家布莱士·帕斯卡（Blaise Pascal, 1623-1662）。帕斯卡的出发点是赌徒在实际赌局中的机会问题，在此前提下他提出了一些较深刻的见解。

布莱士·帕斯卡（1623-1662）

英国牧师托马斯·贝叶斯（Thomas Bayes, 1702-1761）对概率论有了更深刻的理解，并且作出了极其重要的贡献。

托马斯·贝叶斯（1702-1761）

贝叶斯的著名成果涉及到条件概率这个概念，这些条件概率可以针对于人们在实验中看到的或者在某些假设下产生的。这样实验或假设出现的结果称为事件。当无偏地投掷硬币（这时出现正面或背面的相对频率假定为 1/2）10 次，那么 10 次中正好出现 7 次背面的概率是多少，或者 10 次中正好出现 7 次正面的概率是多少？假设生男生女的概率为 1/2，以及不同的出生事件是彼此独立的（即一个小孩的出生不影响其他小孩的出生），那么，一对夫妇生的前两个小孩都是女孩的概率是多少？如果有序的字符串 GGB 表示第一个小孩是女孩，第二个小孩是女孩，第三个小孩是男孩，那么所询问的概率可表示为 $P(GG)$ 是多少？我们也可以利用角标来表示出生的顺序，如 B_1, G_2, G_1B_2 分别代表三个事件：第一个小孩是男孩，第二个是女孩，第一个是女孩且第二个是男孩。现在假设我们知道第一个小孩是女孩。问两个小孩都是女孩的概率是多少？这里我们问的是在给定第一个小孩是女孩的条件下，两个小孩都是女孩的“条件概率”。更一般地，令 $P(X|Y)$ ＝事件 Y 发生的条件下事件 X 发生的概率，也可以记 $P(Y|X)$ ＝事件 X 发生的条件下事件 Y 发生的概率。不难看出，直观上假设 $P(Y)$ 不为零，则

$$P(X\,|\,Y)=\frac{P(X\text{且}Y)}{P(Y)};$$

利用两个集合 X 和 Y 的交集的符号，上式可以转变为

$$P(X\,|\,Y)=\frac{P(X\cap Y)}{P(Y)}.$$

经过片刻的思考，我们就知道当 $P(X)$ 和 $P(Y)$ 都不为零时，$P(X|Y)$ 和 $P(Y|X)$ 并不需相等。

例如，对于上面的小孩出生问题，我们可以知道 $P(G_1G_2|G_1) = 1/2$, $P(G_1|G_1G_2) = 1$; 这里我们用到的下面的事实：$P(G_1G_2) = 1/4$, $P(G_1) = 1/2$.

在计算条件概率时，贝叶斯给出了著名的贝叶斯定理，这个定理在贝叶斯死后才得以

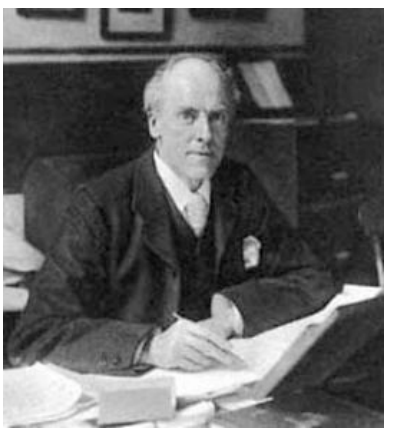
发表。直观上理解，贝叶斯定理提供了一个框架，可以从中排序出对导致事件发生的“因素”（事件）有“相对”影响力的事件。更具体地说，假设我们有一系列事件 $X_1, X_2, \cdots, X_k$，事件间相互排斥，即只可以有一个事件发生。因此，在同样一个空间中，一个实验的结果属于由这些事件组成的集合。假定 E 是一个概率不为零的事件。假设我们知道 E 一定发生，那么如何计算 $P(X_i|E)$？贝叶斯定理是一个“公式”，由这个公式可以计算出 $P(X_i|E)$ 的值，即已知 E 发生，一个特定因素 X_i 发生的概率。

数学史的较早期，即 18 世纪或者更早时，几乎所有对数学做出重大贡献的人同时也是伟大的物理学家。如牛顿（Isaac Newton, 1643-1727），欧拉（Leonhard Euler, 1707-1783），拉普拉斯（Pierre-Simon Laplace, 1749-1827），勒让德（Adrien-Marie Legendre, 1752-1833），高斯（Karl Friedrich Gauss, 1777-1855），他们不仅是杰出的数学家同时也是杰出的物理学家。作为物理学家，他们对数据中的“噪声”，即物理量的测量误差，均有所关注。统计和概率论的某些理论正是源于这种考虑。上面提到的数学家高斯以不同方式思考这个问题，由此产生了“最小二乘法”。正态分布的相关思想亦由此产生。另外，统计学不仅关注自然科学，而且关注社会科学。在这方面的先驱者之一是比利时的数学家和科学家阿道夫 • 凯特勒（Adolphe Quetelet, 1796-1874），他利用正态曲线来研究人类的特点，并从分析的角度来研究犯罪规律。其工作属于将统计思想用于社会学的早期实践。

阿道夫 • 凯特勒（1796-1874）

在现代，越来越多的人对统计及统计与概率论的结合做出贡献。在这里，我将要展示的不仅是来自不同背景不同国家的人们对统计和概率的贡献，更要说明有多少成果根源于近代。当然了，把这个话题说透彻大概需要一本书的内容。

卡尔 • 皮尔逊（Karl Pearson, 1857-1936）在英格兰出生及去世。皮尔逊曾就读于剑桥大学，而他的大部分职业生涯在伦敦大学的大学学院度过。皮尔逊对利用数学的思想来研究进化感兴趣，这也促使他提出了新的统计思想和方法。他 1894 年提出的标准偏差这个术语为众多文科修读统计的学生所知，标准偏差也成为度量数据发散程度的通用术语。皮尔逊还促进了使用大样本的统计检验思想的发展。



卡尔 • 皮尔逊 (1857-1936)

像所有的数学分支一样，在精细的审视下，统计学的发展有着丰富而复杂的历史，其推动者的身份不仅仅是数学家，他们同时也从事其他学术领域的智力活动。凯恩斯（John Maynard Keynes, 1883-1946）在经济学领域大名鼎鼎，然而他在剑桥大学学习了数学，并于 1921 年出版了关于概率论的一本重要专著。凯恩斯的工作被数学家、哲学家罗素勋爵关注，同时也引起了弗兰克 • 拉姆齐（Frank Ramsey）的注意。除了组合数学方面的著名成就（今天我们称为拉姆齐定理），拉姆齐同样在概率论和统计学方面做出了重要的贡献。最初凯恩斯的概率观点趋向于稳定的相对频率方法，但随着时间的推移，他的观点逐渐趋向于主观性更强的信念体系，可能在经济学中这个更有吸引力吧。