数学之美番外篇：平凡而又神奇的贝叶斯方法

BY 刘未鹏 – SEPTEMBER 21, 2008

POSTED IN: 数学, 机器学习与人工智能, 计算机科学

概率论只不过是把常识用数学公式表达了出来。

——拉普拉斯

记得读本科的时候，最喜欢到城里的计算机书店里面去闲逛，一逛就是好几个小时；有一次，在书店看到一本书，名叫贝叶斯方法。当时数学系的课程还没有学到概率统计。我心想，一个方法能够专门写出一本书来，肯定很牛逼。后来，我发现当初的那个朴素归纳推理成立了——这果然是个牛逼的方法。

——题记

目录

0. 前言

这是一篇关于贝叶斯方法的科普文，我会尽量少用公式，多用平白的语言叙述，多举实际例子。更严格的公式和计算我会在相应的地方注明参考资料。贝叶斯方法被证明是非常 general 且强大的推理框架，文中你会看到很多有趣的应用。

1. 历史

托马斯·贝叶斯（Thomas Bayes）同学的详细生平在这里。以下摘一段 wikipedia 上的简介：

> *所谓的贝叶斯方法源于他生前为解决一个"逆概"问题写的一篇文章，而这篇文章是在他死后才由他的一位朋友发表出来的。在贝叶斯写这篇文章之前，人们已经能够计算"正向概率"，如"假设袋子里面有N个白球，M个黑球，你伸手进去摸一把，摸出黑球的概率是多大"。而一个自然而然的问题是反过来："如果我们事先并不知道袋子里面黑白球的比例，而是闭着眼睛摸出一个（或好几个）球，观察这些取出来的球的颜色之后，那么我们可以就此对袋子里面的黑白球的比例作出什么样的推测"。这个问题，就是所谓的逆概问题。*

实际上，贝叶斯当时的论文只是对这个问题的一个直接的求解尝试，并不清楚他当时是不是已经意识到这里面包含着的深刻的思想。然而后来，贝叶斯方法席卷了概率论，并将应用延伸到各个问题领域，所有需要作出概率预测的地方都可以见到贝叶斯方法的影子，特别地，贝叶斯是机器学习的核心方法之一。这背后的深刻原因在于，现实世界本身就是不确定的，人类的观察能力是有局限性的（否则有很大一部分科学就没有必要做了——设想我们能够直接观察到电子的运行，还需要对原子模型争吵不休吗？），我们日常所观察到的只是事物表面上的结果，沿用刚才那个袋子里面取球的比方，我们往往只能知道从里面取出来的球是什么颜色，而并不能直接看到袋子里面实际的情况。这个时候，我们就需要提供一个猜测（hypothesis，更为严格的说法是"假设"，这里用"猜测"更通俗易懂一点），所谓猜测，当然就是不确定的（很可能有好多种乃至无数种猜测都能满足目前的观测），**但也绝对不是两眼一抹黑瞎蒙——具体地说，我们需要做两件事情：1. 算出各种不同猜测的可能性大小。2. 算出最靠谱的猜测是什么。第一个就是计算特定猜测的后验概率，对于连续的猜测空间则是计算猜测的概率密度函数。第二个则是所谓的模型比较，模型比较如果不考虑先验概率的话就是最大似然方法。**

1.1 一个例子：自然语言的二义性

下面举一个自然语言的不确定性的例子。当你看到这句话：

> *The girl saw the boy with a telescope.*

你对这句话的含义有什么猜测？平常人肯定会说：那个女孩拿望远镜看见了那个男孩（即你对这个句子背后的实际语法结构的猜测是：The girl saw-with-a-telescope the boy ）。然而，仔细一想，你会发现这个句子完全可以解释成：那个女孩看见了那个拿着望远镜的男孩（即：The girl saw the-boy-with-a-telescope ）。那为什么平常生活中我们每个人都能够迅速地对这种二义性进行消解呢？这背后到底隐藏着什么样的思维法则？我们留到后面解释。

1.2 贝叶斯公式

贝叶斯公式是怎么来的？

我们还是使用 wikipedia 上的一个例子：

> *一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。有了这些信息之后我们可以容易地计算"随机选取一个学生，他（她）穿长裤的概率和穿裙子的概率是多大"，这个就是前面说的"正向概率"的计算。然而，假设你走在校园中，迎面走来一个穿长裤的学生（很不幸的是你高度近似，你只看得见他（她）穿的是否长裤，而无法确定他（她）的性别），你能够推断出他（她）是男生的概率是多大吗？*

一些认知科学的研究表明（《决策与判断》以及《Rationality for Mortals》第12章：小孩也可以解决贝叶斯问题），我们对形式化的贝叶斯问题不擅长，但对于以频率形式呈现的等价问题却很擅长。在这里，我们不妨把问题重新叙述成：你在校园里面随机游走，遇到了 N 个穿长裤的人（仍然假设你无法直接观察到他们的性别），问这 N 个人里面有多少个女生多少个男生。

你说，这还不简单：算出学校里面有多少穿长裤的，然后在这些人里面再算出有多少女生，不就行了？

我们来算一算：假设学校里面人的总数是 U 个。60% 的男生都穿长裤，于是我们得到了 U * P(Boy) * P(Pants|Boy) 个穿长裤的（男生）（其中 P(Boy) 是男生的概率 = 60%，这里可以简单的理解为男生的比例；P(Pants|Boy) 是条件概率，即在 Boy 这个条件下穿长裤的概率是多大，这里是 100%，因为所有男生都穿长裤）。40% 的女生里面又有一半（50%）是穿长裤的，于是我们又得到了 U * P(Girl) * P(Pants|Girl) 个穿长裤的（女生）。加起来一共是 U * P(Boy) * P(Pants|Boy) + U * P(Girl) * P(Pants|Girl) 个穿长裤的，其中有 U * P(Girl) * P(Pants|Girl) 个女生。两者一比就是你要求的答案。

下面我们把这个答案形式化一下：我们要求的是 P(Girl|Pants)（穿长裤的人里面有多少女生），我们计算的结果是 U * P(Girl) * P(Pants|Girl) / [U * P(Boy) * P(Pants|Boy) + U * P(Girl) * P(Pants|Girl)] 。容易发现这里校园内人的总数是无关的，可以消去。于是得到

P(Girl|Pants) = P(Girl) * P(Pants|Girl) / [P(Boy) * P(Pants|Boy) + P(Girl) * P(Pants|Girl)]

注意，如果把上式收缩起来，分母其实就是 P(Pants)，分子其实就是 P(Pants, Girl) 。而这个比例很自然地就读作：在穿长裤的人（ P(Pants) ）里面有多少（穿长裤）的女孩（ P(Pants, Girl) ）。

上式中的 Pants 和 Boy/Girl 可以指代一切东西，所以其一般形式就是：

P(B|A) = P(A|B) * P(B) / [P(A|B) * P(B) + P(A|~B) * P(~B)]

收缩起来就是：

P(B|A) = P(AB) / P(A)

其实这个就等于：

P(B|A) * P(A) = P(AB)

难怪拉普拉斯说**概率论只是把常识用数学公式表达了出来。**

然而，后面我们会逐渐发现，看似这么平凡的贝叶斯公式，背后却隐含着非常深刻的原理。

2. 拼写纠正

经典著作《人工智能：现代方法》的作者之一 Peter Norvig 曾经写过一篇介绍如何写一个拼写检查/纠正器的文章（原文在这里，徐宥的翻译版在这里，这篇文章很深入浅出，强烈建议读一读），里面用到的就是贝叶斯方法，这里我们不打算复述他写的文章，而是简要地将其核心思想介绍一下。

首先，我们需要询问的是："问题是什么？"

问题是我们看到用户输入了一个不在字典中的单词，我们需要去猜测："这个家伙到底真正想输入的单词是什么呢？"用刚才我们形式化的语言来叙述就是，我们需要求：

P(我们猜测他想输入的单词 | 他实际输入的单词)

这个概率。并找出那个使得这个概率最大的猜测单词。显然，我们的猜测未必是唯一的，就像前面举的那个自然语言的歧义性的例子一样；这里，比如用户输入：thew，那么他到底是想输入 the，还是想输入 thaw？到底哪个猜测可能性更大呢？幸运的是我们可以用贝叶斯公式来直接出它们各自的概率，我们不妨将我们的多个猜测记为 h1 h2 .. （h 代表 hypothesis），它们都属于一个有限且离散的猜测空间 H（单词总共就那么多而已），将用户实际输入的单词记为 D（D 代表 Data，即观测数据），于是

P(我们的猜测1 | 他实际输入的单词)

可以抽象地记为：

P(h1 | D)

类似地，对于我们的猜测2，则是 P(h2 | D)。不妨统一记为：

P(h | D)

运用一次贝叶斯公式，我们得到：

P(h | D) = P(h) * P(D | h) / P(D)

对于不同的具体猜测 h1 h2 h3 .. ，P(D) 都是一样的，所以在比较 P(h1 | D) 和 P(h2 | D) 的时候我们可以忽略这个常数。即我们只需要知道：

P(h | D) ∝ P(h) * P(D | h)（注：那个符号的意思是"正比例于"，不是无穷大，注意符号右端是有一个小缺口的。）

这个式子的抽象含义是：对于给定观测数据，一个猜测是好是坏，取决于"这个猜测本身独立的可能性大小（先验概率，Prior）"和"这个猜测生成我们观测到的数据的可能性大小"（似然，Likelihood）的乘积。具体到我们的那个 thew 例子上，含义就是，用户实际是想输入 the 的可能性大小取决于 the 本身在词汇表中被使用的可能性（频繁程度）大小（先验概率）和 想打 the 却打成 thew 的可能性大小（似然）的乘积。

下面的事情就很简单了，对于我们猜测为可能的每个单词计算一下 P(h) * P(D | h) 这个值，然后取最大的，得到的就是最靠谱的猜测。

一点注记：Norvig 的拼写纠正器里面只提取了编辑距离为 2 以内的所有已知单词。这是为了避免去遍历字典中每个单词计算它们的 P(h) * P(D | h)，但这种做法为了节省时间带来了一些误差。但话说回来难道我们人类真的回去遍历每个可能的单词来计算他们的后验概率吗？不可能。实际上，根据认知神经科学的观点，我们首先根据错误的单词做一个 bottom-up 的关联提取，提取出有可能是实际单词的那些候选单词，这个提取过程就是所谓的基于内容的提取，可以根据错误单词的一些模式片段提取出有限的一组候选，非常快地缩小的搜索空间（比如我输入 explaination，单词里面就有充分的信息使得我们的大脑在常数时间内把可能性 narrow down 到 explanation 这个单词上，至于具体是根据哪些线索——如音节——来提取，又是如何在生物神经网络中实现这个提取机制的，目前还是一个没有弄清的领域）。然后，我们对这有限的几个猜测做一个 top-down 的预测，看看到底哪个对于观测数据（即错误单词）的预测效力最好，而如何衡量预测效率则就是用贝叶斯公式里面的那个 P(h) * P(D | h) 了——虽然我们很可能使用了一些启发法来简化计算。后面我们还会提到这样的 bottom-up 的关联提取。

3. 模型比较与奥卡姆剃刀

3.1 再访拼写纠正

介绍了贝叶斯拼写纠正之后，接下来的一个自然而然的问题就来了："为什么？"为什么要用贝叶斯公式？为什么贝叶斯公式在这里可以用？我们可以很容易地领会为什么贝叶斯公式用在前面介绍的那个男生女生长裤裙子的问题里是正确的。但为什么这里？

为了回答这个问题，一个常见的思路就是想想：非得这样吗？因为如果你想到了另一种做法并且证明了它也是靠谱的，那么将它与现在这个一比较，也许就能得出很有价值的信息。那么对于拼写纠错问题你能想到其他方案吗？

不管怎样，一个最常见的替代方案就是，选择离 thew 的编辑距离最近的。然而 the 和 thaw 离 thew 的编辑距离都是 1 。这可咋办捏？你说，不慌，那还是好办。我们就看到底哪个更可能被错打为 thew 就是了。我们注意到字母 e 和字母 w 在键盘上离得很紧，无名指一抽筋就不小心多打出一个 w 来，the 就变成 thew 了。而另一方面 thaw 被错打成 thew 的可能性就相对小一点，因为 e 和 a 离得较远而且使用的指头相差一个指头（一个是中指一个是小指，不像 e 和 w 使用的指头靠在一块——神经科学的证据表明紧邻的身体设施之间容易串位）。OK，很好，因为你现在已经是在用最大似然方法了，或者直白一点，你就是在计算那个使得 P(D | h) 最大的 h 。

而贝叶斯方法计算的是什么？是 P(h) * P(D | h) 。多出来了一个 P(h) 。我们刚才说了，这个多出来的 P(h) 是特定猜测的先验概率。为什么要掺和进一个先验概率？刚才说的那个最大似然不是挺好么？很雄辩地指出了 the 是更靠谱的猜测。有什么问题呢？既然这样，我们就从给最大似然找茬开始吧——我们假设两者的似然程度是一样或非常相近，这样不就难以区分哪个猜测更靠谱了吗？比如用户输入tlp，那到底是 top 还是 tip ？（这个例子不怎么好，因为 top 和 tip 的词频可能仍然是接近的，但一时想不到好的英文单词的例子，我们不妨就假设 top 比 tip 常见许多吧，这个假设并不影响问题的本质。）这个时候，当最大似然不能作出决定性的判断时，先验概率就可以插手进来给出指示——"既然你无法决定，那么我告诉你，一般来说 top 出现的程度要高许多，所以更可能他想打的是 top "）。

以上只是最大似然的一个问题，即并不能提供决策的全部信息。

最大似然还有另一个问题：即便一个猜测与数据非常符合，也并不代表这个猜测就是更好的猜测，因为这个猜测本身的可能性也许就非常低。比如 MacKay 在《Information Theory：Inference and Learning Algorithms》里面就举了一个很好的例子：-1 3 7 11 你说是等差数列更有可能呢？还是 -X^3 / 11 + 9/11*X^2 + 23/11 每项把前项作为 X 带入后计算得到的数列？此外曲线拟合也是，平面上 N 个点总是可以用 N-1 阶多项式来完全拟合，当 N 个点近似但不精确共线的时候，用 N-1 阶多项式来拟合能够精确通过每一个点，然而用直线来做拟合/线性回归的时候却会使得某些点不能位于直线上。你说到底哪个好呢？多项式？还是直线？一般地说肯定是越低阶的多项式越靠谱（当然前提是也不能忽视"似然"P(D | h)，明摆着一个多项式分布您愣是去拿直线拟合也是不靠谱的，这就是为什么要把它们两者乘起来考虑。），原因之一就是低阶多项式更常见，先验概率（ P(h) ）较大（原因之二则隐藏在 P(D | h) 里面），这就是为什么我们要用样条来插值，而不是直接搞一个 N-1 阶多项式来通过任意 N 个点的原因。

以上分析当中隐含的哲学是，观测数据总是会有各种各样的误差，比如观测误差（比如你观测的时候一个 MM 经过你一不留神，手一抖就是一个误差出现了），所以如果过分去寻求能够完美解释观测数据的模型，就会落入所谓的数据过配（overfitting）的境地，一个过配的模型试图连误差（噪音）都去解释（而实际上噪音又是不需要解释的），显然就过犹不及了。所以 P(D | h) 大不代表你的 h （猜测）就是更好的 h。还要看 P(h) 是怎样的。所谓奥卡姆剃刀精神就是说：如果两个理论具有相似的解释力度，那么优先选择那个更简单的（往往也正是更平凡的，更少繁复的，更常见的）。

过分匹配的另一个原因在于当观测的结果并不是因为误差而显得"不精确"而是因为真实世界中对数据的结果产生贡献的因素太多太多，跟噪音不同，这些偏差是一些另外的因素集体贡献的结果，不是你的模型所能解释的——噪音那是不需要解释——一个现实的模型往往只提取出几个与结果相关度很高，很重要的因素（cause）。这个时候观察数据会倾向于围绕你的有限模型的预测结果呈正态分布，于是你实际观察到的结果就是这个正态分布的随机取样，这个取样很可能受到其余因素的影响偏离你的模型所预测的中心，这个时候便不能贪心不足地试图通过改变模型来"完美"匹配数据，因为那些使结果偏离你的预测的贡献因素不是你这个有限模型里面含有的因素所能概括的，硬要打肿脸充胖子只能导致不实际的模型，举个教科书例子：身高和体重的实际关系近似于一个二阶多项式的关系，但大家都知道并不是只有身高才会对体重产生影响，物理世界影响体重的因素太多太多了，有人身材高大却瘦得跟稻草，有人却是横长竖不长。但不可否认的是总体上来说，那些特殊情况越是特殊就越是稀少，呈围绕最普遍情况（胖瘦适中）的正态分布，这个分布就保证了我们的身高——体重相关模型能够在大多数情况下做出靠谱的预测。但是——刚才说了，特例是存在的，就算不是特例，人有胖瘦，密度也有大小，所以完美符合身高——体重的某个假想的二阶多项式关系的人是不存在的，我们又不是欧几里德几何世界当中的理想多面体，所以，当我们对人群随机抽取了 N 个样本（数据点）试图对这 N 个数据点拟合出一个多项式的话就得注意，它肯定得是二阶多项式，我们要做的只是去根据数据点计算出多项式各项的参数（一个典型的方法就是最小二乘）；它肯定不是直线（我们又不是稻草），也不是三阶多项式四阶多

项式.. 如果硬要完美拟合 N 个点，你可能会整出一个 N-1 阶多项式来——设想身高和体重的关系是 5 阶多项式看看？

3.2 模型比较理论（Model Comparasion）与贝叶斯奥卡姆剃刀（Bayesian Occam's Razor）

实际上，模型比较就是去比较哪个模型（猜测）更可能隐藏在观察数据的背后。其基本思想前面已经用拼写纠正的例子来说明了。我们对用户实际想输入的单词的猜测就是模型，用户输错的单词就是观测数据。我们通过：

$$P(h \mid D) \propto P(h) * P(D \mid h)$$

来比较哪个模型最为靠谱。前面提到，光靠 P(D | h)（即"似然"）是不够的，有时候还需要引入 P(h) 这个先验概率。奥卡姆剃刀就是说 P(h) 较大的模型有较大的优势，而最大似然则是说最符合观测数据的（即 P(D | h) 最大的）最有优势。整个模型比较就是这两方力量的拉锯。我们不妨再举一个简单的例子来说明这一精神：你随便找枚硬币，掷一下，观察一下结果。好，你观察到的结果要么是"正"，要么是"反"（不，不是少林足球那枚硬币:P），不妨假设你观察到的是"正"。现在你要去根据这个观测数据推断这枚硬币掷出"正"的概率是多大。根据最大似然估计的精神，我们应该猜测这枚硬币掷出"正"的概率是 1，因为这个才是能最大化 P(D | h) 的那个猜测。然而每个人都会大摇其头——很显然，你随机摸出一枚硬币这枚硬币居然没有反面的概率是"不存在的"，我们对一枚随机硬币是否一枚有偏硬币，偏了多少，是有着一个先验的认识的，这个认识就是绝大多数硬币都是基本公平的，偏得越多的硬币越少见（可以用一个 beta 分布来表达这一先验概率）。将这个先验正态分布 p(θ)（其中 θ 表示硬币掷出正面的比例，小写的 p 代表这是概率密度函数）结合到我们的问题中，我们便不是去最大化 P(D | h)，而是去最大化 P(D | θ) * p(θ)，显然 θ = 1 是不行的，因为 P(θ=1) 为 0，导致整个乘积也为 0。实际上，只要对这个式子求一个导数就可以得到最值点。

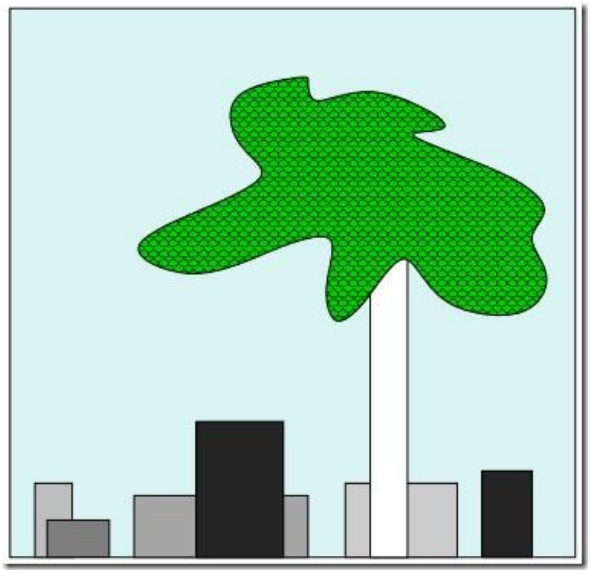
以上说的是当我们知道先验概率 P(h) 的时候，光用最大似然是不靠谱的，因为最大似然的猜测可能先验概率非常小。然而，有些时候，我们对于先验概率一无所知，只能假设每种猜测的先验概率是均等的，这个时候就只有用最大似然了。实际上，统计学家和贝叶斯学家有一个有趣的争论，统计学家说：我们让数据自己说话。言下之意就是要摒弃先验概率。而贝叶斯支持者则说：数据会有各种各样的偏差，而一个靠谱的先验概率则可以对这些随机噪音做到健壮。事实证明贝叶斯派胜利了，胜利的关键在于所谓先验概率其实也是经验统计的结果，譬如为什么我们会认为绝大多数硬币是基本公平的？为什么我们认为大多数人的肥胖适中？为什么我们认为肤色是种族相关的，而体重则与种族无关？先验概率里面的"先验"并不是指先于一切经验，而是仅指先于我们"当前"给出的观测数据而已，在硬币的例子中先验指的只是先于我们知道投掷的结果这个经验，而并非"先天"。

然而，话说回来，有时候我们必须得承认，就算是基于以往的经验，我们手头的"先验"概率还是均匀分布，这个时候就必须依赖用最大似然，我们用前面留下的一个自然语言二义性问题来说明这一点：

> *The girl saw the boy with a telescope.*

到底是 The girl saw-with-a-telescope the boy 这一语法结构，还是 The girl saw the-boy-with-a-telescope 呢？两种语法结构的常见程度都差不多（你可能会觉得后一种语法结构的常见程度较低，这是事后偏见，你只需想想 The girl saw the boy with a book 就知道了。当然，实际上从大规模语料统计结果来看后一种语法结构的确稍稍不常见一丁点，但是绝对不足以解释我们对第一种结构的强烈倾向）。那么到底为什么呢？

我们不妨先来看看 MacKay 在书中举的一个漂亮的例子：



图中有多少个箱子？特别地，那棵书后面是一个箱子？还是两个箱子？还是三个箱子？还是.. 你可能会觉得树后面肯定是一个箱子，但为什么不是两个呢？如下图：

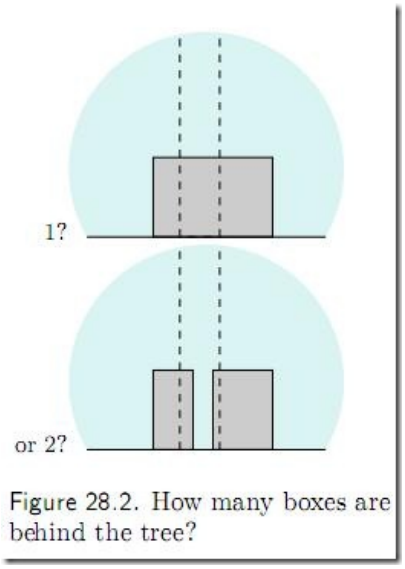


Figure 28.2. How many boxes are behind the tree?

很简单，你会说：要是真的有两个箱子那才怪了，怎么就那么巧这两个箱子刚刚好颜色相同，高度相同呢？

用概率论的语言来说，你刚才的话就翻译为：猜测 h 不成立，因为 P(D | h) 太小（太巧合）了。我们的直觉是：巧合（小概率）事件不会发生。所以当一个猜测（假设）使得我们的观测结果成为小概率事件的时候，我们就说"才怪呢，哪能那么巧捏？！"

现在我们可以回到那个自然语言二义性的例子，并给出一个完美的解释了：如果语法结构是 The girl saw the-boy-with-a-telecope 的话，怎么那个男孩偏偏手里拿的就是望远镜——一个可以被用来 saw-with 的东东捏？这也忒小概率了吧。他咋就不会拿本书呢？拿什么都好。怎么偏偏就拿了望远镜？所以唯一的解释是，这个"巧合"背后肯定有它的必然性，这个必然性就是，如果我们将语法结构解释为 The girl saw-with-a-telescope the boy 的话，就跟数据完美吻合了——既然那个女孩是用某个东西去看这个男孩的，那么这个东西是一个望远镜就完全可以解释了（不再是小概率事件了）。

自然语言二义性很常见，譬如上文中的一句话：

> *参见《决策与判断》以及《Rationality for Mortals》第12章：小孩也可以解决贝叶斯问题*

就有二义性：到底是参见这两本书的第 12 章，还是仅仅是第二本书的第 12 章呢？如果是这两本书的第 12 章那就是咄咄怪事了，怎么恰好两本书都有第 12 章，都是讲同一个问题，更诡异的是，标题还相同呢？

注意，以上做的是似然估计（即只看 P(D | h) 的大小），不含先验概率。通过这两个例子，尤其是那个树后面的箱子的例子我们可以看到，似然估计里面也蕴含着奥卡姆剃刀：树后面的箱子数目越多，这个模型就越复杂。单个箱子的模型是最简单的。似然估计选择了更简单的模型。

这个就是所谓的贝叶斯奥卡姆剃刀（**Bayesian Occam's Razor**），因为这个剃刀工作在贝叶斯公式的似然（P(D | h)）上，而不是模型本身（P(h)）的先验概率上，后者是传统的奥卡姆剃刀。关于贝叶斯奥卡姆剃刀我们再来看一个前面说到的曲线拟合的例子：如果平面上有 N 个点，近似构成一条直线，但绝不精确地位于一条直线上。这时我们既可以用直线来拟合（模型1），也可以用二阶多项式（模型2）拟合，也可以用三阶多项式（模型3），.. ，特别地，用 N-1 阶多项式便能够保证肯定能完美通过 N 个数据点。那么，这些可能的模型之中到底哪个是最靠谱的呢？前面提到，一个衡量的依据是奥卡姆剃刀：越是高阶的多项式越是繁复和不常见。然而，我们其实并不需要依赖于这个先验的奥卡姆剃刀，因为有人可能会争辩说：你怎么就能说越高阶的多项式越不常见呢？我偏偏觉得所有阶多项式都是等可能的。好吧，既然如此那我们不妨就扔掉 P(h) 项，看看 P(D | h) 能告诉我们什么。我们注意到越是高阶的多项式，它的轨迹弯曲程度越是大，到了八九阶简直就是直上直下，于是我们不仅要问：一个比如说八阶多项式在平面上随机生成的一堆 N 个点偏偏恰好近似构成一条直线的概率（即 P(D | h)）有多大？太小太小了。反之，如果背后的模型是一条直线，那么根据该模型生成一堆近似构成直线的点的概率就大得多了。这就是贝叶斯奥卡姆剃刀。

这里只是提供一个关于贝叶斯奥卡姆剃刀的科普，强调直观解释，更多理论公式请参考 MacKay 的著作《Information Theory：Inference and Learning Algorithms》第 28 章。

3.3 最小描述长度原则

贝叶斯模型比较理论与信息论有一个有趣的关联：

P(h | D) ∝ P(h) * P(D | h)

两边求对数，将右式的乘积变成相加：

ln P(h | D) ∝ ln P(h) + ln P(D | h)

显然，最大化 P(h | D) 也就是最大化 ln P(h | D)。而 ln P(h) + ln P(D | h) 则可以解释为模型（或者称"假设"、"猜测"）h 的编码长度加上在该模型下数据 D 的编码长度。使这个和最小的模型就是最佳模型。

而究竟如何定义一个模型的编码长度，以及数据在模型下的编码长度则是一个问题。更多可参考 Mitchell 的《Machine Learning》的 6.6 节，或 Mackay 的 28.3 节）

3.4 最优贝叶斯推理

所谓的推理，分为两个过程，第一步是对观测数据建立一个模型。第二步则是使用这个模型来推测未知现象发生的概率。我们前面都是讲的对于观测数据给出最靠谱的那个模型。然而很多时候，虽然某个模型是所有模型里面最靠谱的，但是别的模型也并不是一点机会都没有。譬如第一个模型在观测数据下的概率是 0.5 。第二个模型是 0.4 ，第三个是 0.1 。如果我们只想知道对于观测数据哪个模型最可能，那么只要取第一个就行了，故事到此结束。然而很多时候我们建立模型是为了推测未知的事情的发生概率，这个时候，三个模型对未知的事情发生的概率都会有自己的预测，仅仅因为某一个模型概率稍大一点就只听他一个人的就太不民主了。所谓的最优贝叶斯推理就是将三个模型对于未知数据的预测结论加权平均起来（权值就是模型相应的概率）。显然，这个推理是理论上的制高点，无法再优了，因为它已经把所有可能性都考虑进去了。

只不过实际上我们是基本不会使用这个框架的，因为计算模型可能非常费时间，二来模型空间可能是连续的，即有无穷多个模型（这个时候需要计算模型的概率分布）。结果还是非常费时间。所以这个被看作是一个理论基准。

4. 无处不在的贝叶斯

以下我们再举一些实际例子来说明贝叶斯方法被运用的普遍性，这里主要集中在机器学习方面，因为我不是学经济的，否则还可以找到一堆经济学的例子。

4.1 中文分词

贝叶斯是机器学习的核心方法之一。比如中文分词领域就用到了贝叶斯。Google 研究员吴军在《数学之美》系列中就有一篇是介绍中文分词的，这里只介绍一下核心的思想，不做赘述，详细请参考吴军的文章（这里）。

分词问题的描述为：给定一个句子（字串），如：

南京市长江大桥

如何对这个句子进行分词（词串）才是最靠谱的。例如：

1. 南京市/长江大桥

2. 南京/市长/江大桥

这两个分词，到底哪个更靠谱呢？

我们用贝叶斯公式来形式化地描述这个问题，令 X 为字串（句子），Y 为词串（一种特定的分词假设）。我们就是需要寻找使得 P(Y|X) 最大的 Y ，使用一次贝叶斯可得：

P(Y|X) ∝ P(Y)*P(X|Y)

用自然语言来说就是 这种分词方式（词串）的可能性 乘以 这个词串生成我们的句子的可能性。我们进一步容易看到：可以近似地将 P(X|Y) 看作是恒等于 1 的，因为任意假想的一种分词方式之下生成我们的句子总是精准地生成的（只需把分词之间的分界符号扔掉即可）。于是，我们就变成了去最大化 P(Y) ，也就是寻找一种分词使得这个词串（句子）的概率最大化。而如何计算一个词串：

W1, W2, W3, W4 ..

的可能性呢？我们知道，根据联合概率的公式展开：P(W1, W2, W3, W4 ..) = P(W1) * P(W2|W1) * P(W3|W2, W1) * P(W4|W1,W2,W3) * .. 于是我们可以通过一系列的条件概率（右式）的乘积来求整个联合概率。然而不幸的是随着条件数目的增加（P(Wn|Wn-1,Wn-2,..,W1) 的条件有 n-1 个），数据稀疏问题也会越来越严重，即便语料库再大也无法统计出一个靠谱的 P(Wn|Wn-1,Wn-2,..,W1) 来。为了

缓解这个问题，计算机科学家们一如既往地使用了"天真"假设：我们假设句子中一个词的出现概率只依赖于它前面的有限的 k 个词（k 一般不超过 3，如果只依赖于前面的一个词，就是2元语言模型（2-gram），同理有 3-gram 、4-gram 等），这个就是所谓的"有限地平线"假设。虽然这个假设很傻很天真，但结果却表明它的结果往往是很好很强大的，后面要提到的朴素贝叶斯方法使用的假设跟这个精神上是完全一致的，我们会解释为什么像这样一个天真的假设能够得到强大的结果。目前我们只要知道，有了这个假设，刚才那个乘积就可以改写成：P(W1) * P(W2|W1) * P(W3|W2) * P(W4|W3) ..（假设每个词只依赖于它前面的一个词）。而统计 P(W2|W1) 就不再受到数据稀疏问题的困扰了。对于我们上面提到的例子"南京市长江大桥"，如果按照自左到右的贪婪方法分词的话，结果就成了"南京市长/江大桥"。但如果按照贝叶斯分词的话（假设使用 3-gram），由于"南京市长"和"江大桥"在语料库中一起出现的频率为 0，这个整句的概率便会被判定为 0 。从而使得"南京市/长江大桥"这一分词方式胜出。

一点注记：有人可能会疑惑，难道我们人类也是基于这些天真的假设来进行推理的？不是的。事实上，统计机器学习方法所统计的东西往往处于相当表层（shallow）的层面，在这个层面机器学习只能看到一些非常表面的现象，有一点科学研究的理念的人都知道：越是往表层去，世界就越是繁复多变。从机器学习的角度来说，特征（feature）就越多，成百上千维度都是可能的。特征一多，好了，高维诅咒就产生了，数据就稀疏得要命，不够用了。而我们人类的观察水平显然比机器学习的观察水平要更深入一些，为了避免数据稀疏我们不断地发明各种装置（最典型就是显微镜），来帮助我们直接深入到更深层的事物层面去观察更本质的联系，而不是在浅层对表面现象作统计归纳。举一个简单的例子，通过对大规模语料库的统计，机器学习可能会发现这样一个规律：所有的"他"都是不会穿 bra 的，所有的"她"则都是穿的。然而，作为一个男人，却完全无需进行任何统计学习，因为深层的规律就决定了我们根本不会去穿 bra 。至于机器学习能不能完成后者（像人类那样的）这个推理，则是人工智能领域的经典问题。至少在那之前，声称统计学习方法能够终结科学研究（原文）的说法是纯粹外行人说的话。

4.2 统计机器翻译

统计机器翻译因为其简单，自动（无需手动添加规则），迅速成为了机器翻译的事实标准。而统计机器翻译的核心算法也是使用的贝叶斯方法。

问题是什么？统计机器翻译的问题可以描述为：给定一个句子 e，它的可能的外文翻译 f 中哪个是最靠谱的。即我们需要计算：P(f|e) 。一旦出现条件概率贝叶斯总是挺身而出：

P(f|e) ∝ P(f) * P(e|f)

这个式子的右端很容易解释：那些先验概率较高，并且更可能生成句子 e 的外文句子 f 将会胜出。我们只需简单统计（结合上面提到的 N-Gram 语言模型）就可以统计任意一个外文句子 f 的出现概率。然而 P(e|f) 却不是那么好求的，给定一个候选的外文局子 f，它生成（或对应）句子 e 的概率是多大呢？我们需要定义什么叫 "对应"，这里需要用到一个分词对齐的平行语料库，有兴趣的可以参考《Foundations of Statistical Natural Language Processing》第 13 章，这里摘选其中的一个例子：假设 e 为：John loves Mary 。我们需要考察的首选 f 是：Jean aime Marie（法文）。我们需要求出 P(e|f) 是多大，为此我们考虑 e 和 f 有多少种对齐的可能性，如：

John (Jean) loves (aime) Marie (Mary)

就是其中的一种（最靠谱的）对齐，为什么要对齐，是因为一旦对齐了之后，就可以容易地计算在这个对齐之下的 P(e|f) 是多大，只需计算：

P(John|Jean) * P(loves|aime) * P(Marie|Mary)

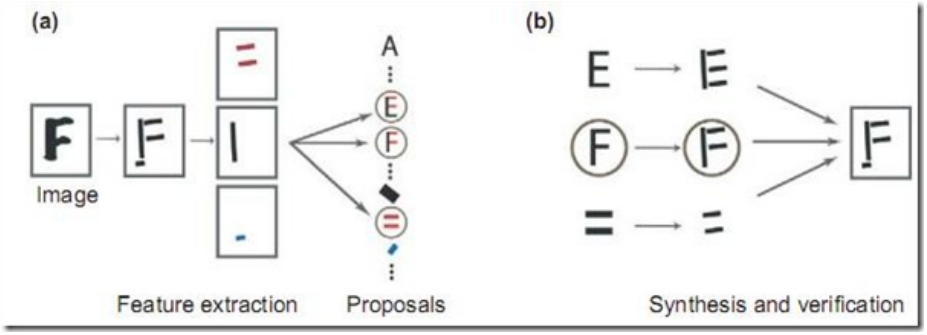
即可。

然后我们遍历所有的对齐方式，并将每种对齐方式之下的翻译概率 ∑ 求和。便可以获得整个的 P(e|f) 是多大。

一点注记：还是那个问题：难道我们人类真的是用这种方式进行翻译的？highly unlikely 。这种计算复杂性非常高的东西连三位数乘法都搞不定的我们才不会笨到去使用呢。根据认知神经科学的认识，很可能我们是先从句子到语义（一个逐层往上（bottom-up）抽象的 folding 过程），然后从语义根据另一门语言的语法展开为另一门语言（一个逐层往下（top-down）的具体化 unfolding 过程）。如何可计算地实现这个过程，目前仍然是个难题。（我们看到很多地方都有 bottom-up/top-down 这样一个对称的过程，实际上有人猜测这正是生物神经网络原则上的运作方式，对视觉神经系统的研究尤其证明了这一点，Hawkins 在《On Intelligence》里面提出了一种 HTM（Hierarchical Temporal Memory）模型正是使用了这个原则。）

4.3 贝叶斯图像识别，Analysis by Synthesis

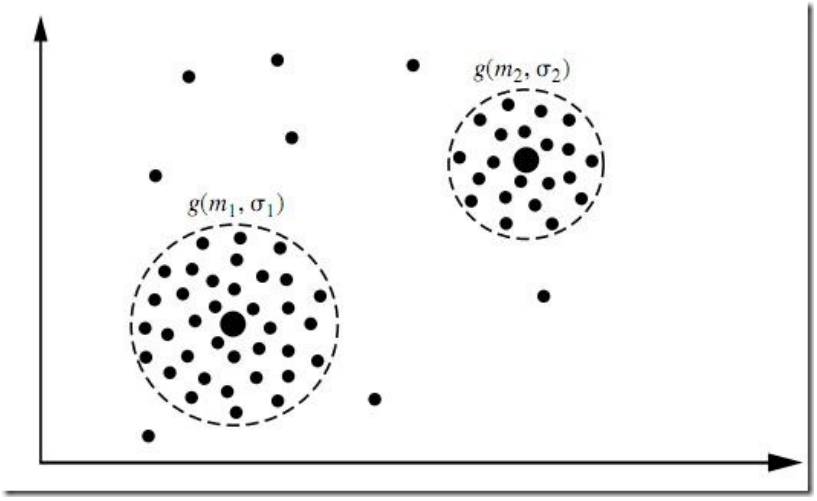
贝叶斯方法是一个非常 general 的推理框架。其核心理念可以描述成：Analysis by Synthesis（通过合成来分析）。06 年的认知科学新进展上有一篇 paper 就是讲用贝叶斯推理来解释视觉识别的，一图胜千言，下图就是摘自这篇 paper：



首先是视觉系统提取图形的边角特征，然后使用这些特征自底向上地激活高层的抽象概念（比如是 E 还是 F 还是等号），然后使用一个自顶向下的验证来比较到底哪个概念最佳地解释了观察到的图像。

4.4 EM 算法与基于模型的聚类

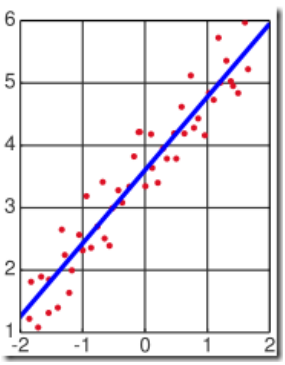
聚类是一种无指导的机器学习问题，问题描述：给你一堆数据点，让你将它们最靠谱地分成一堆一堆的。聚类算法很多，不同的算法适应于不同的问题，这里仅介绍一个基于模型的聚类，该聚类算法对数据点的假设是，这些数据点分别是围绕 K 个核心的 K 个正态分布源所随机生成的，使用 Han JiaWei 的《Data Ming：Concepts and Techniques》中的图：



图中有两个正态分布核心，生成了大致两堆点。我们的聚类算法就是需要根据给出来的那些点，算出这两个正态分布的核心在什么位置，以及分布的参数是多少。这很明显又是一个贝叶斯问题，但这次不同的是，答案是连续的且有无穷多种可能性，更糟的是，只有当我们知道了哪些点属于同一个正态分布圈的时候才能够对这个分布的参数作出靠谱的预测，现在两堆点混在一块我们又不知道哪些点属于第一个正态分布，哪些属于第二个。反过来，只有当我们对分布的参数作出了靠谱的预测时候，才能知道到底哪些点属于第一个分布，那些点属于第二个分布。这就成了一个先有鸡还是先有蛋的问题了。为了解决这个循环依赖，总有一方要先打破僵局，说，不管了，我先随便整一个值出来，看你怎么变，然后我再根据你的变化调整我的变化，然后如此迭代着不断互相推导，最终收敛到一个解。这就是 EM 算法。

EM 的意思是"Expectation-Maximazation"，在这个聚类问题里面，我们是先随便猜一下这两个正态分布的参数：如核心在什么地方，方差是多少。然后计算出每个数据点更可能属于第一个还是第二个正态分布圈，这个是属于 Expectation 一步。有了每个数据点的归属，我们就可以根据属于第一个分布的数据点来重新评估第一个分布的参数（从蛋再回到鸡），这个是 Maximazation 。如此往复，直到参数基本不再发生变化为止。这个迭代收敛过程中的贝叶斯方法在第二步，根据数据点求分布的参数上面。

4.5 最大似然与最小二乘



学过线性代数的大概都知道经典的最小二乘方法来做线性回归。问题描述是：给定平面上 N 个点，（这里不妨假设我们想用一条直线来拟合这些点——回归可以看作是拟合的特例，即允许误差的拟合），找出一条最佳描述了这些点的直线。

一个接踵而来的问题就是，我们如何定义最佳？我们设每个点的坐标为 (Xi, Yi) 。如果直线为 y = f(x) 。那么 (Xi, Yi) 跟直线对这个点的"预测"：(Xi, f(Xi)) 就相差了一个 ΔYi = |Yi – f(Xi)| 。最小二乘就是说寻找直线使得 (ΔY1)^2 + (ΔY2)^2 + .. （即误差的平方和）最小，至于为什么是误差的平方和而不是误差的绝对值和，统计学上也没有什么好的解释。然而贝叶斯方法却能对此提供一个完美的解释。

我们假设直线对于坐标 Xi 给出的预测 f(Xi) 是最靠谱的预测，所有纵坐标偏离 f(Xi) 的那些数据点都含有噪音，是噪音使得它们偏离了完美的一条直线，一个合理的假设就是偏离路线越远的概率越小，具体小多少，可以用一个正态分布曲线来模拟，这个分布曲线以直线对 Xi 给出的预测 f(Xi) 为中心，实际纵坐标为 Yi 的点 (Xi, Yi) 发生的概率就正比于 EXP[-(ΔYi)^2]。(EXP(..) 代表以常数 e 为底的多少次方）。

现在我们回到问题的贝叶斯方面，我们要想最大化的后验概率是：

P(h|D) ∝ P(h) * P(D|h)

又见贝叶斯！这里 h 就是指一条特定的直线，D 就是指这 N 个数据点。我们需要寻找一条直线 h 使得 P(h) * P(D|h) 最大。很显然，P(h) 这个先验概率是均匀的，因为哪条直线也不比另一条更优越。所以我们只需要看 P(D|h) 这一项，这一项是指这条直线生成这些数据点的概率，刚才说过了，生成数据点 (Xi, Yi) 的概率为 EXP[-(ΔYi)^2] 乘以一个常数。而 P(D|h) = P(d1|h) * P(d2|h) * .. 即假设各个数据点是独立生成的，所以可以把每个概率乘起来。于是生成 N 个数据点的概率为 EXP[-(ΔY1)^2] * EXP[-(ΔY2)^2] * EXP[-(ΔY3)^2] * .. = EXP**{-[(ΔY1)^2 + (ΔY2)^2 + (ΔY3)^2 + ..**]} 最大化这个概率就是要最小化 (ΔY1)^2 + (ΔY2)^2 + (ΔY3)^2 + .. 。熟悉这个式子吗？

5. 朴素贝叶斯方法

朴素贝叶斯方法是一个很特别的方法，所以值得介绍一下。我们用朴素贝叶斯在垃圾邮件过滤中的应用来举例说明。

5.1 贝叶斯垃圾邮件过滤器

问题是什么？问题是，给定一封邮件，判定它是否属于垃圾邮件。按照先例，我们还是用 D 来表示这封邮件，注意 D 由 N 个单词组成。我们用 h+ 来表示垃圾邮件，h- 表示正常邮件。问题可以形式化地描述为求：

P(h+|D) = P(h+) * P(D|h+) / P(D)

P(h-|D) = P(h-) * P(D|h-) / P(D)

其中 P(h+) 和 P(h-) 这两个先验概率都是很容易求出来的，只需要计算一个邮件库里面垃圾邮件和正常邮件的比例就行了。然而 P(D|h+) 却不容易求，因为 D 里面含有 N 个单词 d1, d2, d3, .. ，所以P(D|h+) = P(d1,d2,..,dn|h+) 。我们又一次遇到了数据稀疏性，为什么这么说呢？P(d1,d2,..,dn|h+) 就是说在垃圾邮件当中出现跟我们目前这封邮件一模一样的一封邮件的概率是多大！开玩笑，每封邮件都是不同的，世界上有无穷多封邮件。瞧，这就是数据稀疏性，因为可以肯定地说，你收集的训练数据库不管里面含了多少封邮件，也不可能找出一封跟目前这封一模一样的。结果呢？我们又该如何来计算 P(d1,d2,..,dn|h+) 呢？

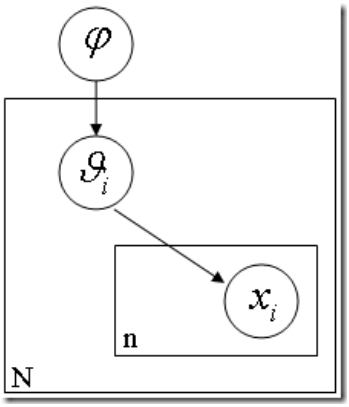
我们将 P(d1,d2,..,dn|h+) 扩展为：P(d1|h+) * P(d2|d1, h+) * P(d3|d2,d1, h+) * .. 。熟悉这个式子吗？这里我们会使用一个更激进的假设，我们假设 di 与 di-1 是完全条件无关的，于是式子就简化为 P(d1|h+) * P(d2|h+) * P(d3|h+) * .. 。这个就是所谓的条件独立假设，也正是朴素贝叶斯方法的朴素之处。而计算 P(d1|h+) * P(d2|h+) * P(d3|h+) * .. 就太简单了，只要统计 di 这个单词在垃圾邮件中出现的频率即可。关于贝叶斯垃圾邮件过滤更多的内容可以参考这个条目，注意其中提到的其他资料。

一点注记：这里，为什么有这个数据稀疏问题，还是因为统计学习方法工作在浅层面，世界上的单词就算不再变多也是非常之多的，单词之间组成的句子也是变化多端，更不用说一篇文章了，文章数目则是无穷的，所以在这个层面作统计，肯定要被数据稀疏性困扰。我们要注意，虽然句子和文章的数目是无限的，然而就拿邮件来说，如果我们只关心邮件中句子的语义（进而更高抽象层面的"意图"（语义，意图如何可计算地定义出来是一个人工智能问题），在这个层面上可能性便大大缩减了，我们关心的抽象层面越高，可能性越小。单词集合和句子的对应是多对一的，句子和语义的对应又是多对一的，语义和意图的对应还是多对一的，这是个层级体系。神经科学的发现也表明大脑的皮层大致有一种层级结构，对应着越来越抽象的各个层面，至于如何具体实现一个可放在计算机内的大脑皮层，仍然是一个未解决问题，以上只是一个原则（principle）上的认识，只有当 computational 的 cortex 模型被建立起来了之后才可能将其放入电脑。

5.2 为什么朴素贝叶斯方法令人诧异地好——一个理论解释

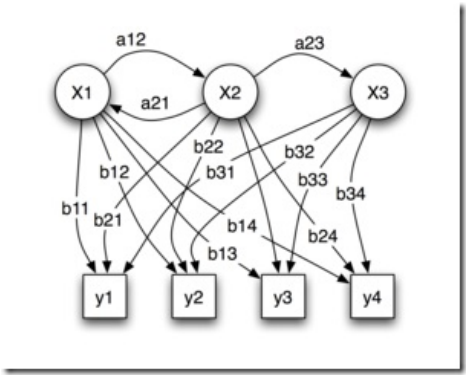
朴素贝叶斯方法的条件独立假设看上去很傻很天真，为什么结果却很好很强大呢？就拿一个句子来说，我们怎么能鲁莽地声称其中任意一个单词出现的概率只受到它前面的 3 个或 4 个单词的影响呢？别说 3 个，有时候一个单词的概率受到上一句话的影响都是绝对可能的。那么为什么这个假设在实际中的表现却不比决策树差呢？有人对此提出了一个理论解释，并且建立了什么时候朴素贝叶斯的效果能够等价于非朴素贝叶斯的充要条件，这个解释的核心就是：有些独立假设在各个分类之间的分布都是均匀的所以对于似然的相对大小不产生影响；即便不是如此，也有很大的可能性各个独立假设所产生的消极影响或积极影响互相抵消，最终导致结果受到的影响不大。具体的数学公式请参考这篇 paper 。

6. 层级贝叶斯模型



层级贝叶斯模型是现代贝叶斯方法的标志性建筑之一。前面讲的贝叶斯，都是在同一个事物层次上的各个因素之间进行统计推理，然而层次贝叶斯模型在哲学上更深入了一层，将这些因素背后的因素（原因的原因，原因的原因，以此类推）囊括进来。一个教科书例子是：如果你手头有 N 枚硬币，它们是同一个工厂铸出来的，你把每一枚硬币掷出一个结果，然后基于这 N 个结果对这 N 个硬币的 θ（出现正面的比例）进行推理。如果根据最大似然，每个硬币的 θ 不是 1 就是 0（这个前面提到过的），然而我们又知道每个硬币的 p(θ) 是有一个先验概率的，也许是一个 beta 分布。也就是说，每个硬币的实际投掷结果 Xi 服从以 θ 为中心的正态分布，而 θ 又服从另一个以 Ψ 为中心的 beta 分布。层层因果关系就体现出来了。进而 Ψ 还可能依赖于因果链上更上层的因素，以此类推。

6.1 隐马可夫模型（HMM）



吴军在数学之美系列里面介绍的隐马可夫模型（HMM）就是一个简单的层级贝叶斯模型：

> *那么怎么根据接收到的信息来推测说话者想表达的意思呢？我们可以利用叫做"隐含马尔可夫模型"（Hidden Markov Model）来解决这些问题。以语音识别为例，当我们观测到语音信号 o1,o2,o3 时，我们要根据这组信号推测出发送的句子 s1,s2,s3。显然，我们应该在所有可能的句子中找最有可能性的一个。用数学语言来描述，就是在已知 o1,o2,o3,…的情况下，求使得条件概率 P (s1,s2,s3,…|o1,o2,o3….) 达到最大值的那个句子 s1,s2,s3,…*

吴军的文章中这里省掉没说的是，s1, s2, s3, .. 这个句子的生成概率同时又取决于一组参数，这组参数决定了 s1, s2, s3, .. 这个马可夫链的先验生成概率。如果我们将这组参数记为 λ，我们实际上要求的是：P(S|O, λ)（其中 O 表示 o1,o2,o3,.. ，S表示 s1,s2,s3,..）

> *当然，上面的概率不容易直接求出，于是我们可以间接地计算它。利用贝叶斯公式并且省掉一个常数项，可以把上述公式等价变换成*
>
> *P(o1,o2,o3,…|s1,s2,s3….) * P(s1,s2,s3,…)*
>
> *其中*

> *P(o1,o2,o3,...|s1,s2,s3....) 表示某句话 s1,s2,s3...被读成 o1,o2,o3,...的可能性，而 P(s1,s2,s3,...) 表示字串 s1,s2,s3,...本身能够成为一个合乎情理的句子的可能性，所以这个公式的意义是用发送信号为 s1,s2,s3...这个数列的可能性乘以 s1,s2,s3.. 本身可以一个句子的可能性，得出概率。*

这里，s1,s2,s3...本身可以一个句子的可能性其实就取决于参数 λ，也就是语言模型。所以简而言之就是发出的语音信号取决于背后实际想发出的句子，而背后实际想发出的句子本身的独立先验概率又取决于语言模型。

7. 贝叶斯网络

吴军已经对贝叶斯网络作了科普，请直接跳转到这里。更详细的理论参考所有机器学习的书上都有。

参考资料

一堆机器学习，一堆概率统计，一堆 Google，和一堆 Wikipedia 条目，一堆 paper 。

部分书籍参考《机器学习与人工智能资源导引》。

TAGS: 数学, 机器学习与人工智能, 计算机科学

52 条评论 18 条转载 239 条新浪微博 7 条腾讯微博 从旧到新排序

spidertiger
Cool! I like this article!
2009年2月8日 回复

Jeffye
Very good.
Bayesian is very very useful.
2009年2月13日 回复

ollydbg
好文章，我已经把你的文章放入中文wiki里面了

http://zh.wikipedia.org/wiki/贝叶斯定理
2009年2月19日 回复


刘未鹏
谢谢 ollydbg
2009年2月19日 回复

willin
這樣解說太棒了！我要仔細再研究...
2009年3月3日 回复

FuYi
文章写得太棒了！把我一直头痛的许多问题解释的非常清楚
2009年3月15日 回复

thejinchao
赞！
这篇文章一定要好好读一下
2009年3月23日 回复

pc
good explication!
2009年3月24日 回复

Samuel.D

拜读牛文！

2009年3月26日 回复

chxy

好，很好，很好很强大

2009年3月27日 回复

flexin`

牛阿，学习了！

2009年4月3日 回复

jinzhi

你好，想问一下"06 年的认知科学新进展上有一篇 paper 就是讲用贝叶斯推理来解释视觉识别的"，这篇文章的题目是什么，谢谢啦

2009年4月7日 回复

路人

.....贝叶斯已经萎了 而且过去也不是一篇 是他妈一堆

8月18日 回复

路人

对不起啊 说的不确切 朴素贝叶斯已经萎了 现在大多就是在模型上做下推广 做context 或者做segment配合的识别之类的 .... 现在随机场还是很火的 这个吧 刘大哥也不是这个领域的 不过是听说或者偶尔看见了这paper 就顺口一说 底下的小娃娃就觉得我操牛逼死了这paper 就算你是小娃娃也要动脑筋啊。。。唉 有点误人子弟啊 怎么说哦 你这种心态去看看经典、基础的paper比较好 别人一说我操这paper牛逼就赶紧下下来 别人说我操这跟踪做的牛逼 就赶紧搜源码存好 这种心态你最后就只能当个码农 看看算法的书 然后觉得自己牛逼死了

8月18日 回复

ccgk

真的很强悍！

说得太好了！

2009年4月30日 回复

soo

作为机器学习方面的专家，我不得不说，文章还是不错的

2009年5月20日 回复

hbsztsyx

一个字，牛；

两个字，佩服；

三个字，太好了；

受教了，谢了！

2009年6月11日 回复

jiasha

我忽然觉得heckman模型里是不是也用了贝叶斯方法~

2009年6月18日 回复

kai

你好！我觉得你的文章写得很好。不过你提到，机器学习需要对数据进行统计才能得到"所有男人都不穿bra"这个结论，而人可以通过对深层规律的推理得知。你认为两者获得知识的方式不同，然后认为"声称统计学习方法能够终结科学研究(原文)的说法是纯粹外行人说的话"。我觉得不管结论是否正确，首先这个推理是不全面的。因为如果两种方法能在一定时间内得到同样的结论，那可以认为两者是等价的。但你没有证明两者不是等价的。

2009年6月30日 回复

挑灯看剑

好文章啊，受益匪浅！

牛人！！！

2009年7月20日 回复

davansy

帮将来的我顶，虽然现在还看不懂.....！顶，未鹏牛！

2009年10月11日 回复

liguow

看了开头就忍不住要来拜一下。 您绝对是牛人！！！

2009年10月13日 回复

splendid sun

受教了，把复杂理论知识用通俗生动的语言表述出来，这种文章我最喜欢了！

2009年12月5日 回复

Philome

谢谢哦！看看增长下自己的见识&

2010年1月23日 回复

socrat

有一个地方不是很清楚，还请不吝赐教：P
关于MacKay 树后面是一个箱子还是两个箱子的那个例子，
h表示的是猜测，猜测可能是一个箱子或两个箱子
D表示的是图中的观察到的情况，就是树后面露出两截箱子。
那么P(D|h)不就应该是：我当h这个猜测是真的，然后再算这个猜测为真的情况下观察到图例子中情况的概率。
这样的话，无论我的猜测是一个箱子还是两个箱子，我观察到的都是树后面两截箱子，所以P(D|h)应该都等于1才对啊。
至于你说的两个箱子的假设，存在的可能性很低，其实应该是说P(h)很低吧？
这样的话，这个例子是否本应该用来说明奥卡姆剃刀，而非贝叶斯奥卡姆剃刀？

2010年5月26日 回复

asuwill

"这样的话，无论我的猜测是一个箱子还是两个箱子，我观察到的都是树后面两截箱子，所以P(D|h)应该都等于1才对啊"
你说的这句话就不对。P(D|h)是指已知h发生的情况下，D发生的概率。就你问的问题来说，如果h是：树后有两个箱子，那么P(D|h)是：树后有两个箱子，看起来如图所示的概率。这个概率不大，正如作者在文章中所说，两个箱子高度、颜色都一致，同时被树挡住，我们的经验告诉我们，不太可能

2011年3月10日 回复

kklots

"这样的话，无论我的猜测是一个箱子还是两个箱子，我观察到的都是树后面两截箱子，所以P(D|h)应该都等于1才对啊"
你说的这句话就不对。P(D|h)是指已知h发生的情况下，D发生的概率。就你问的问题来说，如果h是：树后有两个箱子，那么P(D|h)是：树后有两个箱子，看起来如图所示的概率。

我觉得你们讨论的问题不是一个问题。
前者讨论的应该是：P(h|D)，其中h是：树后面有两个一模一样的箱子，且箱子缝隙刚好被树挡住。那么P(D|h)应该是：树后有两个一模一样的箱子，且箱子缝隙刚好被树挡住，在这种情况下，出现如图情况的概率。很明显，这时的P(D|h)应该为1，但树后有两个一模一样的箱子的概率P(h)很低，导致结果不可能。
后者讨论的是：P(h|D)，其中h是：树后面有两个箱子。那么P(D|h)应该是：树后有两个箱子，出现如图情况的概率。这时的P(D|h)应该很小，因为有两个箱子的情况有很多种，但偏偏出现两个箱子的高低、颜色相同，且同时被树挡住的情况很少见。

2011年12月15日 回复

Naich

Very Useful, thanks

2010年6月24日 回复

Brianlan

好文~！
深刻且易懂~
订阅了~

2010年11月30日 回复

非法人

有一点科学研究的理念的人都知道：越是往表层去，世界就越是繁复多变.....
有体会，我能想到的最形象的比喻是 google地球
现在的研究领域过细是否能与之链接？

2011年2月8日 回复

hercy

有理有据，图文并茂，通俗易懂，非常赞！

2011年2月26日 回复

liuyizhe

理解到这个程度的人，不在少数，但是能这么清晰，条例分明的写出来的人，实在是少之又少啊赞！

2011年3月28日 回复

non-bayes

贝叶斯很神奇？
http://fur.ly/5g3g

2011年5月6日 回复

river

很有用 很好懂

2011年7月13日 回复

breakinen

真的谢谢你的文章!
现在是晚上1点17我还在看这篇~
第一次感觉到数学之美
第一次改变了对上学期刚考完试的概率论的仇视态度:P

谢谢!

2011年8月6日 回复

bluesjay

好文
按发音的正确翻译应该是: 贝斯。
就像Greenwich

2011年8月23日 回复

Paul

文章非常好，深入浅出，给理论一个非常直观的解释。文笔排版也很好，非常值得阅读。

2011年8月30日 回复

Edwin

从头到尾读完，受益匪浅！ 博主耐心地花费心思用浅显的语言写出这么长一篇文章，这种知识分享的精神值得尊敬。

2011年10月26日 回复

Directory

wowooo. 讲解的很不错，惭愧是学数学的，收藏你的博客。

2011年12月7日 回复

深蓝

发现一个错误，P(h | D) ∝ P(h) * P(D | h)，不能推出ln P(h | D) ∝ ln P(h) + ln P(D | h)，取对数以后就不再是正比关系了

2月12日 回复

zhuwenxiang

楼上 ln是凸函数可保持递增性 正比和正比例不是一个概念

3月10日 回复

libin198783

晕 你这样的运算估计就没有正比干系了 任何一个类似的式子 经过变换都不是正比

3月12日 回复

tcipc

∝是正相关的意思，前面正巧是正比，后面这样写也没错

4月11日 回复

tcipc

讲的太好了，以前我只知道有这么一个公式，现在才知道他有这么深的内涵。

4月11日 回复

pumps on sale

经典啊

4月11日 回复

黄成

发现一处错误：

“

一点注记：Norvig 的拼写纠正器里面只提取了编辑距离为 2 以内的所有已知单词。这是为了避免去遍历字典中每个单词计算它们的 P(h) * P(D | h)，但这种做法为了节省时间带来了一些误差。但话说回来难道我们人类真的回去遍历每个可能的单词来计算他们的后验概率吗？

”

粗体处好像应该是“会”

8月1日 回复

cschen

写得很赞，通俗易懂，学习了

8月30日 回复

廖廖斋

通俗易懂啊……

9月12日 回复

亚宁

牛文一篇

9月12日 回复

masikkk

通俗易懂，非常感谢

9月21日 回复

罗健明

很不错~

9月24日 回复

kun

初次相见，秉烛夜读！不错……

10月5日 回复

使用社交帐号登录: 微博 QQ 人人 豆瓣 开心 更多»

说点什么吧 ...

不想登录？直接点击发布即可作为游客留言。 发 布

刘未鹏 | Mind Hacks正在使用多说