什么是贝叶斯方法?

作者：王江源 来源：爱科学

本文原地址：https://www.iikx.com/news/statistics/6064.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

什么是**贝叶斯方法**?提到统计，就绕不开概率。

概率是什么?什么决定了概率?这是一个哲学问题。先按下不表，因为我也不会!

一般认为，概率是一个事件发生的可能性大小。

在经典力学的框架下，只要知道了一个系统中各个粒子的初始状态，无论这个系统多么复杂，都可以根据牛顿经典动力学方程模拟出系统在任何时刻的状态的状态。换句话说，世界是确定的，毫无随机性可言。

但实际上，我们不可能获得系统所有的信息，只能根据有限的观测数据去猜。怎么猜，就是通过各种统计建模。

贝叶斯观点认为，概率是一个人对于一个事件发生的信心，即认为一个事件有多大的可能性会发生。如果认为一定会发生，概率就是1;认为一定不会发生，概率就是0;有可能发生，概率就介于0-1之间。

这种概率会随着观测数据的变化而变化。一开始我们没有数据，只能根据原有经验猜测一个大概的概率，称为**先验概率**
。随着观测数据的
积累，我们会逐渐修正我们的先验概
率，修正后的概率称为**后验概率**。数据越多，我们的后验概率就越接近于真实的情况。

举个例子，来了一个新同学小明，我们不知道他多过去。如果让我们猜测他期末考试及格的可能性，因为我们不知道他既往的考试成绩，所以合理的猜测是50%，作为先验概率。后来发现小明几乎不去上课，去了也是碎觉，我们可以认为他是个学渣，期末考试“十有八九”不及格。结果小明期末考试分数很高，那么我们对他的考试能力可能会有改观。如果小明第二次、第三次都考得很好，我们可能会认为他是学神，妈妈再也不用担心他的考试成绩了。

说了这么多，我们来简单推导贝叶斯的公式。

假设A、B是两个事件，两者同时发生的概率记为P(AB)。

分步来看，AB同时发生的概率=A发生的概率×A发生的情况下B发生的概率。

写成公式就是：

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

这里的AB是对称的，所以也可以写成：

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

也就是说：

P(A)P(B|A) =P(B)P(A|B)

两边除以P(A):

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

这就是贝叶斯公式的基本形式。

在实际应用中，我们把B称为假设，用H表示，把A称为证据或数据，用D表示，贝叶斯公式变成：

P(H|D) = (P(H)P(D|H)) / P(D)

我们从另外的角度来看这个公式：

(1)左边的P(H|D)是指在证据的情况下假设发生的概率，称为**后验概率**。

(2)右边的P(H)是在没有任何证据的情况下H发生的概率，称为**先验概率**。

(3)右边的P(D|H)是在该假设发生的情况下，出现该证据的概率，称为**似然度**。

(4)右边的P(D)是一个归一化因子，保证所有概率之和等于1。数值上来说，它等于所有假设下先验概率与似然度乘积的和，即所有假设下分子的和。

其中后验概率是我们想知道的，先验概率是已知的，似然度也可以根据已知的数据去算得。

这里解释一下什么是假设。假设是一个随机变量的所有可能发生的情况，比如考试结果有及格和不及格两种情况，及格或不及格都是考试结果的一种假设。这里的假设要满足两个条件：

(1)互斥性：及所以假设中最多只有一个假设会发生。

(2)完备性：假设必须包含所有情况，且至少有一个假设会发生。

贝叶斯推断的步骤：

贝叶斯的典型应用是求某一个特定假设的后验概率。

(1)第一步：找出所有可能的假设。

(2)第二步：求出所有假设的**先验概率**。

(3)第三步：求出所有假设的**似然度**。

(4)第四步：对于各种假设，用先验概率乘以似然度，然后归一化，根据贝叶斯公式求出后验概率。

下面举一个经典的烂大街的例子来说明贝叶斯公式的应用过程。

已知某种疾病在人群中的患病率为1/10000，现在发明了一种新的检测方法用于诊断该疾病。这种方法测定1000名患者，结果显示其中阳性者999例，阴性者1例;又测定了1000名正常人，其中阴性者995例，阳性者5例。请问：对于一个人，该检测结果为阳性，请问患病的概率是多少。

在回答这个问题之前，我们把检测方法整理为一个四格表：

	患病	未患病
结果为阳性	999	5
结果为阴性	1	995

容易求出：
该检查方法的：
敏感性 = 999/1000 = 0.999
特异性 = 995/1000 = 0.995
阳性预测值 = 999/（999+5） = 0.995
阴性预测值 = 995/（995+1） = 0.999
从检查效果的角度来讲，这个检测方法已经非常好了，敏感性、特异性都非常好。

回到问题中，这个结果为阳性的患者患病的概率是多少呢？
一个简单粗暴的回答是：999/（999+5） = 0.995，即阳性预测值。

真的是这样吗？
如果我们考虑先验概率，结果就大不一样，因为这个病在人群中发生的概率很低！

我们用贝叶斯来计算：
这里的假设有两个：患病与不患病，先验概率都知道。似然度可以根据检验方法的数据来计算。

假设H	证据D	先验概率P(H)	似然度P(D|H)	归一化因子P(D)	后验概率P(H|D)
患病	阳性	0.0001	0.999(999/100	0.0051	0.02

			0)		
未患病	阳性	0.9999	0.005(5/1000)	0.0051	0.98

可以看出，患病的概率为2%，不患病的概率为98%！
那是不是说明这个检测方法没有用呢？当然不是！检测前，患病概率是0.01%，检测阳性后，患病概率增加为2%，提高了200倍。
这个时候该怎么办？复查一次！

如果还是阳性，结果将如何。这时的先验概率就变成了0.02。
重复上面的计算过程：

假设H	证据D	先验概率P(H)	似然度P(D|H)	归一化因子P(D)	后验概率P(H|D)
患病	阳性	0.02	0.999(999/1000)	0.025	0.80
未患病	阳性	0.98	0.005(5/1000)	0.025	0.20

这时患病概率上升至80%！
再复查一次，仍为阳性！

假设H	证据D	先验概率P(H)	似然度P(D|H)	归一化因子P(D)	后验概率P(H|D)
患病	阳性	0.8	0.999(999/1000)	0.8	0.999
未患病	阳性	0.2	0.005(5/1000)	0.8	0.001

患病概率上升至99.9%！

对于不寻常的事情，需要不寻常的证据！

更多 统计方法 请访问 https://www.iikx.com/news/statistics/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发