02 | 贝叶斯视角下的机器学习

2018-06-07 王天一

机器学习40讲 进入课程 >

讲述：王天一
时长 17:03 大小 8.29M

在上一篇文章中，我向你介绍了频率学派对概率、统计和机器学习的理解。今天则要转换视角，看一看贝叶斯学派解决这些问题的思路。

还记得那个“九死一生”的例子吗？对其中 90% 的概率更直观、也更合理的解释是生病之后生还的可能性。之所以说频率主义的解释牵强，是因为没有哪个人能倒霉到三番五次地得这个病。当多次独立重复试验不可能实现时，就不存在从频率角度解读概率的理论基础。

虽然上面的这个例子不见得严谨，却可以用来描述频率学派的问题：对于所有的“一锤子买卖”，也就是不包含随机变量的事件来说，频率学派对概率的解读都是不成立的。

为了解决频率主义的问题，贝叶斯学派给出了一种更加通用的概率定义：概率表示的是客观上事件的可信程度（degree of belief），也可以说成是主观上主体对事件的信任程度，它

是建立在对事件的已有知识基础上的。

比方说，当一个球迷提出“明天皇家马德里战胜拉斯帕尔马斯的概率是 86%”的时候，可以理解成他对皇马获胜有 86% 的把握程度，要是买球的话自然就会在独胜上下出重注（其实贝叶斯概率正是来源于对赌博的分析）。

除了对概率的置信度解释之外，贝叶斯学派中的另一个核心内容是贝叶斯定理（Bayes' theorem），用来解决“逆向概率问题”（inverse probability problem）。

听名字就知道，逆向概率和前向概率是对应的：假定数据由一个生成模型给出，前向概率是在已知生成过程的前提下来计算数据的概率分布和数字特征，逆向概率则是在已知数据的前提下反过来计算生成过程的未知特性。贝叶斯定理的数学表达式可以写成

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

式中的 $P(H)$ 被称为**先验概率**（prior probability）；$P(D|H)$ 被称为**似然概率**（likelihood probability）；$P(H|D)$ 被称为**后验概率**（posterior probability）。

抛开乱七八糟的符号，贝叶斯定理同样可以从贝叶斯概率的角度加以解读：所谓先验概率是指根据以往经验和分析得到的概率，可以视为假设 H 初始的可信程度；与假设 H 相关的数据 D 会作为证据出现，将数据纳入考虑范围后，假设 H 的可信程度要么会增强要么会削弱。但不管增强还是削弱，得到的结果都是经过数据验证的假设的可信程度，这就是后验概率。

贝叶斯定理的意义正是在于将先验概率和后验概率关联起来，刻画了数据对于知识和信念的影响。

纳粹德国的宣传部长保罗·约瑟夫·戈培尔有句名言：“如果你说的谎言范围够大，并且不断重复，人民终会开始相信它。”从贝叶斯定理的角度看，这句话是有科学依据的空穴来风。本来谎言的先验概率 $p(lie)$，也就是初始的可信度接近于 0，而 $p(\bar{lie}) = 1 - p(lie)$ 接近于 1。可问题的关键在于似然概率——戈培尔这句话的核心是被宣传对象对将谎言说成真理的宣传的信任。宣传对象相信宣传者不说假话，意味着似然概率 $p(brainwash|lie)$ 较大，同时 $p(brainwash|\bar{lie})$ 较小。这样一来，经过宣传之后，谎言成立的后验概率就可以写成

$$p(lie|brainwash) = \frac{p(lie) \cdot p(brainwash|lie)}{p(lie) \cdot p(brainwash|lie) + p(\bar{lie}) \cdot p(brainwash|\bar{lie})}$$

稍作分析就不难发现，只要 $p(brainwash|lie) > 0.5$，谎言的后验概率就会大于先验概率。更重要的是，本次宣传得到的后验概率 $(lie|brainwash)$ 将作为下次宣传的先验概率 $p(lie)$ 出现。于是，在后验概率与先验概率不断迭代更新的过程中，$p(lie|brainwash)$ 将持续上升，谎言也就越来越接近真理了。

将贝叶斯定理应用到统计推断中，就是贝叶斯主义的统计学。频率统计理论的核心在于认定待估计的参数是固定不变的常量，而用来估计的数据是随机的变量。**贝叶斯统计则恰恰相反：它将待估计的参数视为随机变量，用来估计的数据反过来是确定的常数，讨论观测数据的概率分布才是没有意义的。**贝叶斯统计的任务就是根据这些确定的观测数据反过来推断未知参数的概率分布。

相对于频率主义的最大似然估计，贝叶斯主义在参数估计中倾向于使后验概率最大化，使用最大后验概率估计（maximum a posteriori estimation）。

频率学派认为观测数据之所以会出现是因为它出现的概率最大，因此最可能的参数就是以最大概率生成这一组训练数据的参数。最大后验估计则是将频率学派中“参数”和“数据”的角色做了个调换：参数本身是随机变量（服从先验分布），有许多可能的取值，而不同取值生成这一组观测数据（服从似然分布）也是不同的。因而最大后验概率推断的过程就是结合参数自身的分布特性，找到最可能产生观测数据的那个参数的过程。

贝叶斯定理告诉我们，**后验概率正比于先验概率和似然概率的乘积，这意味着后验概率实质上就是用先验概率对似然概率做了个加权处理。**频率主义将参数看成常量，那么似然概率就足以描述参数和数据之间的关系。贝叶斯主义则将参数看成变量，因此参数自身的特性也会影响到参数和数据之间的关系。先验概率的作用可以用下面的例子来说明（本例来自 David JC MacKay, Information Theory, Inference, and Learning Algorithms, Example 2.3）

“Jo 去进行某种疾病的检查。令随机变量 a 表示 Jo 的真实健康状况，$a = 1$ 表示 Jo 生病，$a = 0$ 表示 Jo 没病；令随机变量 b 表示 Jo 的检查结果，$b = 1$ 表示阳性，$b = 0$ 表示阴性。已知检查的准确率是 95%，也就是此病患者的检查结果 95% 会出现阳性，非此病患者的检查结果 95% 会出现阴性，同时在 Jo 的类似人群中，此病的发病率是 1%。如果 Jo 的检查结果呈阳性，那么她患病的概率是多大呢？"

直观理解，“检查的准确率是 95%”似乎说明了 Jo 患病的概率就是 95%，可事实真是这样吗？根据贝叶斯定理，患病概率可以写成

$$p(a=1|b=1)=\frac{p(b=1|a=1)\cdot p(a=1)}{p(b=1|a=1)\cdot p(a=1)+p(b=1|a=0)\cdot p(a=0)}$$

式中的 $p(b=1|a=1)=0.95$ 就是似然概率，$p(a=1)=0.01$ 则是先验概率。不难求出，Jo 患病的真正概率，也就是后验概率只有 16%!

为什么会出现这样的情况呢？对于频率学派来说，Jo 要么生病要么没病，概率的推演是在这两个确定的前提下分别进行的，所以似然概率就足以说明问题。可是阳性检查结果既有真阳性也有假阳性，两者的比例是不同的。虽然真阳性基本意味着生病，但由于先验概率较小（1%），它在所有的阳性结果中依然是少数（16%）。相比之下，假阳性结果凭借其比较大的先验概率（99%），占据了阳性结果的大部分（84%）。这个例子说明抛开先验概率谈论似然概率，是没有多少说服力的。

不难看出，先验信息在贝叶斯统计中占据着相当重要的地位。可问题在于先验信息从哪里来?

先验信息是在使用数据之前关于分析对象的已有知识，可当这种已有知识并不存在时，就不能对先验做出合理的建模。事实上，指定先验分布的必要性正是贝叶斯学派被频率学派的诟病之处，因为先验分布不可避免地会受到主观因素的影响，这与统计学立足客观的出发点背道而驰。这中间的哲学思辨在此不做探讨，你只需要知道**即使包含某些主观判断，先验信息也是贝叶斯主义中不可或缺的核心要素。**

当已有的知识实在不足以形成先验信息时，贝叶斯主义的处理方式是引入**无信息先验**（noninformative prior），认为未知参数取到所有取值的可能性都是相等的，也就是满足均匀分布。由于此时的先验概率是个常数，这个先验概率也被称为**平坦先验**（flat prior）。**在平坦先验之下，最大后验估计和最大似然估计是等效的。**

不知道你还记不记得上一篇文章末尾的例子？如果从频率主义出发，可以用最大似然估计求出 Alice 得分的概率 $\hat{p}=5/8$，而 Bob 赢得赌局的概率就是他连得三分的概率 $(1-\hat{p})^3\approx 0.0527$。

可是在贝叶斯主义看来，事情并没有这么简单，因为已有的投球结果并不能给出关于得分位置的可靠信息，5:3 的领先可能意味着 Alice 有较大的得分概率，也可能意味着 Bob 虽有有较大的得分概率却走了背字。因而在贝叶斯学派看来，处理未知参数 p 的方式不应该是武断地把它看成一个常数，而是应该从变量的角度去观察，考虑它在 [0, 1] 上所有可能的取值，再计算在所有可能的取值下 Bob 获胜概率的数学期望，从而消除 p 的不确定性对结果的影响。

在这样的思想下，Bob 获胜的概率就可以写成

$$E = \int_0^1 (1-p)^3 P(p|A=5, B=3) \mathrm{d}p$$

利用贝叶斯定理可以将上式中的条件概率写成

$$P(p|A=5, B=3) = \frac{P(A=5, B=3|p)P(p)}{\int_0^1 P(A=5, B=3|p)P(p)\mathrm{d}p}$$

在这个式子中，先验概率 $P(p)$ 是在观察到数据之前 p 的分布，因而是未知的。但由于 p 服从均匀分布，所以是个常数，也就不会对 $P(p|A=5, B=3)$ 产生影响。另一方面，$P(A=5, B=3|p)$ 可以用二项分布计算，其数值等于 $8!/(5!3!)p^5(1-p)^3$。将这一结果代入 E 的表达式，可以得到

$$E = \frac{\int_0^1 p^5(1-p)^6 \mathrm{d}p}{\int_0^1 p^5(1-p)^3 \mathrm{d}p} = 0.0909$$

显然，这与最大似然估计得到的结果是不同的。但这个结果却符合频率主义的阐释：如果用蒙特卡洛法（Monte Carlo method）进行数值仿真的话，你会发现这个 0.0909 才是符合真实情况的概率。

将贝叶斯定理应用到机器学习之中，完成模型预测和选择的任务，就是贝叶斯视角下的机器学习。**由于贝叶斯定理大量涉及各种显式变量与隐藏变量的依赖关系，通常用概率图模型来**

直观地描述。贝叶斯主义将未知参数视为随机变量，参数在学习之前的不确定性由先验概率描述，学习之后的不确定性则由后验概率描述，这中间不确定性的消除就是机器学习的作用。

与频率主义不同的是，贝叶斯学习的输出不是简单的最优估计值 $\hat{\theta}$，而是关于参数的概率分布 $p(\theta)$，从而给出了更加完整的信息。在预测问题中，贝叶斯学习给出的也不仅仅是一个可能性最大的结果，而是将所有结果及其概率以概率分布的形式完整地呈现出来。

除了在预测中提供更加完备的信息之外，贝叶斯学习在模型选择上也有它的优势。在贝叶斯主义看来，所谓不同的模型其实就是不同概率分布的参数化表示，使用的参数也有它们自己的先验分布，但所有模型的共同点是它们都能生成训练数据集，而模型选择的任务就是从这些概率分布中挑出一个最好的。

这里的“好”的标准就是数据和模型的符合程度，也叫可信度（model evidence）。可信度实际上就是归一化的似然函数 $p(D|M)$，表示的是模型 M 生成数据 D 的条件概率。当不同复杂度模型的经验风险接近的时候，就可以利用可信度来筛选模型了。

既然贝叶斯主义能够提供更加完整的信息，为什么迟迟没有取代频率主义成为主流呢？这就不得不说贝叶斯方法的缺点了：一是对未知变量的积分运算会导致极高的计算复杂度（computation complexity），这从 Alice 和 Bob 打赌的例子中就可以看出；二是对先验分布的设定（prior specification）包含一定的主观性，因而一直不招老派的统计学家待见。正是这两个原因限制了贝叶斯方法的广泛应用。

今天我和你分享了贝叶斯学派对概率、统计学和机器学习的认识方式，其要点如下：

贝叶斯学派认为概率是事件的可信程度或主体对事件的信任程度；

贝叶斯学派执行参数估计时，视参数为随机变量，视数据为确定取值；

贝叶斯学派主要使用最大后验概率法，让参数在先验信息和给定数据下的后验概率最大化；

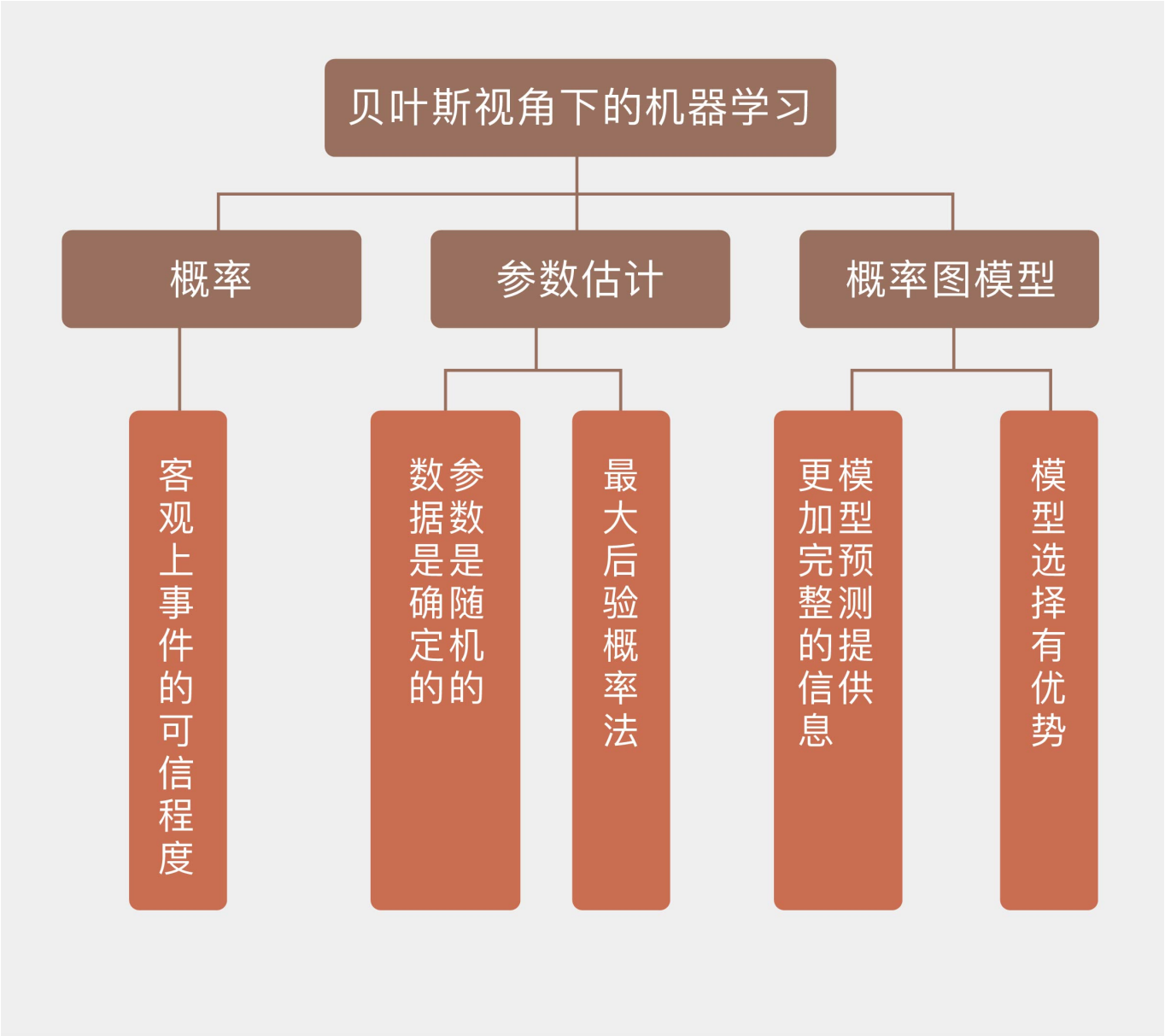
贝叶斯学派对应机器学习中的概率图模型，可以在模型预测和选择中提供更加完整的信息。

在这两篇文章中，我和你探讨了频率主义和贝叶斯主义这两个解决概率问题的基本思路，它们也是以后理解不同机器学习方法的基础。虽然两种观点各执一词，争论得不可开交，但两

者更像是一枚硬币的两面，在思想方法上没有根本性的对立，各种频率主义下的统计学习方法也可以通过贝叶斯来解释。**将两种方法论融会贯通才是理解机器学习的正确思路。**

最后再回到 Alice 和 Bob 赌局的例子，基于贝叶斯主义的方法得到了符合频率学派解释的结果，基于频率主义的最大似然估计反而做出了错误的判断，那么你是怎么看待频率学派的错误呢?

欢迎发表你的观点。



上一篇　01 | 频率视角下的机器学习

下一篇　03 | 学什么与怎么学

精选留言 (19)

写留言

风的轨迹
2018-06-12

20

王老师，综合贝叶斯主义以及频率主义这两节课，我理解总结为以下4点，不知道是否正确：
1. 在统计问题上，频率学派认为，参数是一个固定值（因为分布固定了嘛），数据是随机的，之后根据最大似然估计来求得参数值。但是这里有一个暗含的假设，就是如果参数固定那么分布也是固定的，也就是说我在讨论问题之前把模型固定好了，那么问题来了，...
展开

作者回复: 总结得非常好！但要说明的一点是一般来说，模型的形式是预先固定的（线性回归或者高斯混合或者其他）。在给定模型形式的前提下，贝叶斯可以通过后验来控制模型的参数和复杂度。

林彦
2018-06-08

4

频率学派把一种未观测到的球落在哪个位置的概率当成了唯一的概率，参数也唯一。其他位置的概率根据已观测的数据虽然小一些，但完全都用同一概率代替会造成误差。我的理解观测的次数增加会降低这种误差(这过程中球落哪的概率不变)。

作者回复: 观测次数增加，最大似然估计的结果会越来越接近真实值。

彭擦擦🐸...
2018-08-21

2

频率派和贝叶斯派在理论层面势不两立（我是坚定的站贝叶斯）

而一旦到了应用层，就是谁好用就用谁：频率派偏向于收敛，就去解决已知问题的修改（置信区间、威尔逊算法）；贝叶斯偏向于迭代，则去解决未知问题的预测（贝叶斯网络）

展开

作者回复: 贝叶斯虽然思想很棒，但运算太复杂，很多时候不接地气。要是效果差不多的话，频率方法一般是首选。

风的轨迹
2018-06-12

2

另外关于Alice和Bob的赌局，我也较一个真啊，虽说从频率学派来看，Alice赢的概率是一个确定值，但是就用8次观察的结果作为估计值也有点误差太大了吧，频率学派估计要喊冤

展开

作者回复: 你说的对，这个例子不是用来说明频率主义是错误的，而是说它在非观察变量的处理上的确存在问题。如果说投球的次数增加，最大似然的估计肯定会越来越接近真实值。但在多次重复实验难以实现时，频率主义的劣势就会凸显。

李奇科
2018-06-07

2

我认为Bayes的最大缺点在于计算量（计算时间），而不是您讲的积分复杂程度和先验问题。实际研究中会发现Bayes的公式虽然看着复杂，但不难推导。这一点恰恰是Bayes的优势。这个优势也使得Bayes模型可以很flexible。而往往frequentist的问题的数学推导会十分复杂，（ingenious）。另外先验开率提供了incorporate更多信息的device。也不好简单的说是缺点。

展开

作者回复: 贝叶斯和频率就是两种不同的思路，两者在概率、统计和机器学习里都有应用。我向大家介绍两者，并没有对哪个的偏爱，也不存在对它们优劣的评判。
频率的思路是由因及果，贝叶斯在此基础上进一步由果溯因，这是我所说“逆向概率”的含义，因为在频率学派里是没有先验后验的概念的。
贝叶斯的计算量就是来源于对积分的计算，原则上说，贝叶斯推理应该把所有的非观测变量积分掉，也就是marginalization，这是贝叶斯统计的核心。正是因为太多太复杂的积分求不出解析解，才要用复杂的计算去近似的。
贝叶斯更灵活是一点儿毛病也没有的，毕竟自带正则化特效。

code-arti...
2019-01-21

1

初出茅庐的小伙子，实践经验少，使用频率主义容易犯错。
多读书有利于在大脑中能形成准确的先验考虑。
每个人的头脑中的先验概率受父母，个人成长经历，读过的书，看过的影视剧等等因素所影响，因而对事物的判断不一样。
我们内心中的那份固执源于以往成功的经验或失败的教训

展开

never_give...
2018-06-10

1

老师，有个疑问，对于那个赌博问题，为什么只将p换成了条件概率下的p，1-p中的p不用换吗?

作者回复: (1-p)^3表示的是bob连胜3局这个事件发生的概率，在p取不同值时结果也不同，所以要积分。后面条件概率的含义是前面的p是真实值的可能性，或者说在5:3的数据给出的关于p的可信度。
你可以把p看成一个随机变量，后面的条件概率是随机变量的概率分布，(1-p)^3是随机变量的函数。

不吐槽会死...

2018-06-07

1

基础差，听得迷迷糊糊的，也不确定自己听懂没。不过作业还是要交的。我觉得频率学派这次的失败主要是因为重复的实验次数太少了吧，假如次数足够多，求出概率的极限值，那得出的结婚也差不多了。这个我觉得反而是贝叶斯学派的优势，样品比较少时计算会更加精确。

...

展开

作者回复: 虽然扔了很多次球，但从计算胜率的角度看，这场比赛其实只是一次独立的试验，根本不能依此对估计的精度做出推断。所以计算的错误并不能说明频率思想存在问题，只是对例子中的非观测变量处理不当。

你说的有道理，频率需要大量重复实验来保证精确度，但贝叶斯可以将所有不确定因素的影响都体现在结果中，这是通过数学原理保证的，与数据量无关。

正反面各0.5的概率是用来进行数学分析的理想的假设，在实际当中扔硬币其实根本不是随机事件，当所有的参数——出手角度、空气阻力等等全都已知时，硬币的正反面就是可以计算的确定结果。所以硬币这个问题要当成理想的数学模型来看，无需纠结概率差和站起来的问题。

林彦

2018-06-07

1

有具体的例子，公式推导，例子来自于难度适当的文献并给出完整的文献信息，概述理论并给出框架信息，对相关问题的与读者互动。王老师的专栏比较适合我这种入门级水平的读者更好地理解。从做老师的角度看您为学生考虑了不少。谢谢！最近我除了工作任务更多外，还在上一门有编程任务的数据分析类课程，努力挤出时间来跟上您专栏的进度。

展开

作者回复: 给足压力才能让潜力完全释放，加油！

李奇科

2018-06-07

1

对王老师的逆向概率不是很赞同。Bayes研究中不少是用generative model的

展开

阿土伯
2019-02-27

频率学派可否看成是贝叶斯学派的一种特例呢？因为频率学派假定参数是不便的，这个观点就属于先验概率吧？

小刀
2019-01-23

公式怎么都显示的看起来很难受啊？？

展开

秦龙君
2019-01-10

学习了。

展开

你不是我
2018-06-07

我认为，概率学，就是在给定的概率上，直接做计算，概率已定，所以需要大量数据作为依靠！频率学则是认为概率是随机变动的，如赌博的例子，下面第一把为3/8，如果赢了，概率就改成了4/9，所以p是流动的，谁优谁劣不好说，这可能和已有数据量有很大关系。

作者回复: 观测次数越多，最大似然的结果会越接近真实值。

李奇科
2018-06-07

您讲的Joe看病的例子难道不仅仅是一个条件概率的问题吗，我认为以此无法区分Bayes 和frequentist的优劣吧。frequentist也是承认条件概率的啊。我自己虽然也是Bayesian，但是对王老师青睐Bayes的原因无法认同

展开

韶华
2018-06-07

健康检查那个例子，我理解应该是发病的概率，而不是患病的概率，对吗？

展开

韶华

2018-06-07

健康检查那个例子，我理解应该是发病（病发生，人可以感知到）的概率，而不是患病（查出是阳性，但是人没有感觉）的概率，对吗？

作者回复: 求出来的是检查出有病实际也有病的概率，只有0.16。

Float

2018-06-07

因为频率学派将概率看作随机事件发生频率的极限值，所以我认为只有当赌博次数足够多时，最大似然估计的p值才能接近真实的p值，而仅仅靠8次赌博的结果估计得到的p值误差显然很大。

展开

作者回复: 虽然扔了很多次球，这场比赛其实只是一次独立的试验，根本不能依此对估计的精度做出推断。所以计算的错误并不能说明频率思想存在问题，只是对例子中的非观测变量处理不当。

李奇科

2018-06-07

你的那个谎言的例子中，observation又是什么呢？谈不上条件概率的不断增加吧