**生物统计学- Bayesian Statistics**

**药物所 池帅帅**

**一、什么是贝叶斯统计？**

**1、历史：**

[贝叶斯统计](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF%E7%BB%9F%E8%AE%A1" \t "https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF%E7%BB%9F%E8%AE%A1/_blank)的历史可以上溯到 16 世纪。1713 年，James Bernoulli 意识到在可用于机会游戏的演绎逻辑和每日生活中的归纳逻辑之间的区别，他提出一个著名的问题：前者

的机理如何能帮助处理后面的推断。

托马斯.贝叶斯（ThomasBayes，1702－1761）是长老会的牧师。他对这个问题产生浓厚的兴趣，并且对这个问题进行认真的研究，期间，他写了一篇文章来回答Bernoulli 的问题，提出了后来以他的名字命名的公式：贝叶斯公式。

但是，直到贝叶斯死后才由他的朋友Richard Price 在 1763 年发表了这篇文章，对Bernoulli 的问题提供了回答。这篇文章标志着贝叶斯统计的产生。但贝叶斯统计的思想在开始时并没有得到重视。后来，Laplace 本人重新发现了贝叶斯公式，而且阐述得比贝叶斯更为清晰。由于贝叶斯统计对于概率的观点过于主观，与当时的主流统计观点相左，此外也很难应用当时严谨的数学理论解释。

**2、贝叶斯公式**

贝叶斯可以由结果A来推测各个原因Bi的概率



二、**贝叶斯统计和频率统计的比较**

**1、二者利用的信息**

（1）频率学派的观点：统计推断是根据样本信息对总体分布或总体的特征数进行推断，这里用到两种信息：总体信息和样本信息。

（2）贝叶斯学派的观点：除了上述两种信息以外，统计推断还应该使用第三种信息：**先验信息。**

人们在试验之前对要做的问题在经验上和资料上总是有所了解的，这些信息对统计推断是有益的。先验信息，即是抽样(试验)之前有关统计问题的一些信息。一般说来，先验信息来源于经验和历史资料。这些先验信息不仅可以减少样本容量，而且在很多情况还可以提高统计精度；而经典统计学派忽略了这些信息。

1. 总体信息：总体分布或所属分布族提供给我们的信息
2. 样本信息：从总体抽取的样本提供给我们的信息
3. 先验信息：在抽样之前有关统计推断的一些信息

**2.关于θ的理解**

假设Ⅰ：随机变量X有一个密度函数p(x;θ)，其中θ是一个参数，不同的θ对应不同的密度函数，故从贝叶斯观点看，p(x;θ)是在给定θ后的一个条件密度函数，因此记为p(x|θ)更恰当一些。在贝叶斯统计中记为p(x|θ)，它表示在随机变量θ给定某个值时，总体指标X的条件分布。这个条件密度能提供我们的有关的θ信息就是总体信息。

假设Ⅱ：当给定θ后，从总体p(x|θ)中随机抽取一个样本X1，X2...Xn，该样本中含有θ的有关信息。这种信息就是样本信息。

假设Ⅲ：从贝叶斯观点来看，未知参数θ是一个随机变量。而描述这个随机变量的分布可从先验信息中归纳出来，这个分布称为先验分布，其密度函数用（θ）表示。

贝叶斯统计在重视使用总体信息和样本信息的同时，还注意先验信息的收集、挖掘和加工，使它数量化，形成先验分布，参加到统计推断中来，以提高统计推断的质量。

贝叶斯学派认为：任一未知量θ都可看作随机变量，可用一个概率分布去描述，这个分布称为先验分布；在获得样本之后，总体分布、样本与先验分布通过贝叶斯公式结合起来得到一个关于未知量θ新的分布即后验分布；任何关于θ的统计推断都应该基于θ的后验分布进行。

相反，经典统计学派却把未知量θ就简单看成一个未知参数，来对它进行统计推断。

表 1频率统计与贝叶斯统计的比较

|  |  |  |
| --- | --- | --- |
|  | **频率统计** | **贝叶斯统计** |
| **P值** | 通常表示显著水平 | 关于参数的概率 |
| **认识论** | 只从特定的样本中获取知识 | 承认过往研究结果并以此为基础结合新样本更新知识 |
| **样本要求** | 通常大样本 | 各种样本量 |
| **参数的性质** | 固定值 | 未知随机量 |
| **总体参数的特点** | 具体的值 | 不确定的分布 |
| **估计区间的意义** | 从总体中抽取无数个样本，其中95%的样本包含总体均值 | 区间包含95%的总体值 |

贝叶斯缩小了范围、指明了方向

**举例说明：**

我把手机放在了家里的某个地方。我可以使用仪器底座上的电话定位器来定位电话，当我按下电话定位器时，电话会开始发出哔哔声。

问题：我应该搜索我家的哪个区域？

**频率推理**

我能听到手机发出的哔哔声。我还有一个心理猜测模型，即根据声音来确定区域。因此，在听到哔哔声后，我推断出我家的区域，我必须遍历搜索来找到手机。

**贝叶斯推理**

我能听到手机发出的哔哔声。现在，除了帮助我识别声音来自哪个区域的心理模型之外，我还知道过去经常放错电话的位置。因此，我结合我的推论使用了哔哔声和我之前关于我过去放错电话的位置的先验信息，以确定我必须搜索以找到手机的区域。

从这个例子，可以看出两者的区别，贝叶斯推理比频繁推理应用了更多的经验，缩小了范围、避免盲目。

**贝叶斯统计的作用过程**

极小概率问题：日出问题

“不管发生了什么，太阳第二天仍然会照常升起”，你对这个真理的坚信来源于你从来没碰见过一天，太阳没有升起。但是...

想象一下，有一天早上你醒来，太阳决定休息一天。这不仅会（最有可能）破坏你的一天并搞砸你的生物钟，这也会直接改变你对太阳升落的感觉。你不再坚信太阳永远会第二天升起来的真理了！更有可能预测到第二天太阳也不会升起。或者说，你对太阳将再次休息一天的期望将会比以前高很多。

贝叶斯统计的作用过程就是：我们根据新证据改变了对事件发生概率的先有的、固有的看法。这是所有贝叶斯统计数据的关键。

更数学的描述一下，贝叶斯规则：

贝叶斯的规则告诉我们，我们必须从一些关于事件发生可能性的固有概率开始（事前）。我们称之为先验概率。逐渐地，随着我们获得新的观察和证据，我们查看证据，决定我们当前立场的可能性基础上更新我们的信念。这种更新的信念称为后验概率（事后）。

回到我们的日出问题，我们每天都观察到太阳升起，每当它发生时我们都会更确定它会在第二天再次升起。但是，如果有一天我们发现太阳没有上升，这将根据新的证据对我们的后验概率产生巨大影响。

这在数学上以下面的形式表达，起初看起来令人生畏但可以被抽象：我们更新的信念是基于我们最初的信念和基于我们当前信念（可能性）呈现的新证据。有多少新的证据，我们的信念有有多少可能是正确的。如果最初信念是太阳明天不上升的概率是百万分之一，如果某一天（仅仅是如果）太阳没有照常升起，那么我的信念错误的可能性非常高，后验概率会更新以预测它是更有可能再次发生。

参考：

**<https://baijiahao.baidu.com/s?id=1627429240976795329&wfr=spider&for=pc>**

**<https://blog.csdn.net/u012308776/article/details/47211653>**

**<http://blog.sina.com.cn/s/blog_7fb03f7d0102x7qf.html>**

**三、在机器学习中的应用**

**1、机器学习的概念及定义**

维基百科上的定义

机器学习是人工智能的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。

机器学习有下面几种定义：

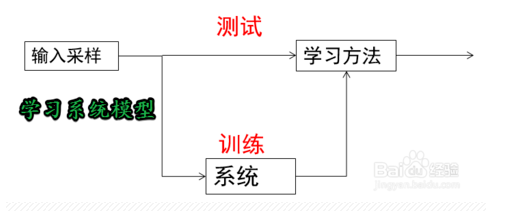
机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。

机器学习是对能通过经验自动改进的计算机算法的研究。

机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。

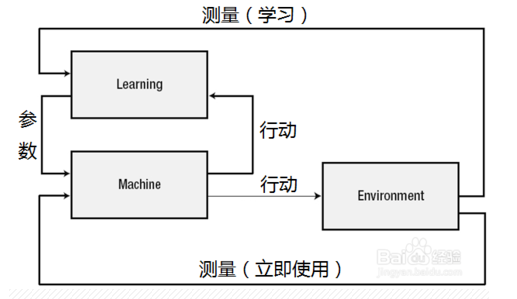
一种经常引用的英文定义是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

 机器学习是计算机科学领域中的一个领域，现有的数据被用来预测或响应未来的数据。它与模式识别、计算统计、人工智能等领域有着密切的联系。 机器学习可以使用机器（计算机和软件）从已知数据中挖掘意义，从而给机器付予学习环境的能力。



**2、机器学习模型**

机器学习是将测量转化为将来操作的参数的过程，机器从环境中获取信息后，输入系统的数据分为两个部分，一个用于对机器产生即时响应，另一部分为输入进行学习。例如，在飞机自动驾驶中，对飞机高度的测量，通常不直接用于控制，而是用来帮助选择其它的飞行参数，如速度 、俯仰角等。



**3、贝叶斯统计在现实场景中应用：**

1. 检测垃圾电子邮件
2. 将新闻分为科技、政治、体育等类别
3. 判断一段文字表达积极的情绪还是消极的情绪
4. 用于人脸检测软件

参考：

<https://blog.csdn.net/sinat_30353259/article/details/80932111>

<https://blog.csdn.net/qq_34993631/article/details/79176274>

<https://blog.csdn.net/shb_derek1/article/details/76907066>

<https://blog.csdn.net/springyh/article/details/79525301>

<https://blog.csdn.net/R1uNW1W/article/details/79126170>