原创 极大似然估计与贝叶斯估计

2016-10-11 14:04:58 jim_刘 阅读数 33171 ☆收藏 更多

序言

本序言是对整体思想进行的一个概括。若没有任何了解，可以先跳过，最后回来看看；若已有了解，可以作为指导思想。
极大似然估计与贝叶斯估计是统计中两种对模型的参数确定的方法，两种参数估计方法使用不同的思想。前者来自于频率派，认为参数是固定的，我们要是根据已经掌握的数据来估计这个参数；而后者属于贝叶斯派，认为参数也是服从某种概率分布的，已有的数据只是在这种参数的分布下产生的。所以，上，极大似然估计就是假设一个参数 θ，然后根据数据来求出这个θ. 而贝叶斯估计的难点在于$p\left(\theta\right)$ 需要人为设定，之后再考虑结合MAP （maximum posterior）方法来求一个具体的θ.
所以极大似然估计与贝叶斯估计最大的不同就在于是否考虑了先验，而两者适用范围也变成了：极大似然估计适用于数据大量，估计的参数能够较好的反况；而贝叶斯估计则在数据量较少或者比较稀疏的情况下，考虑先验来提升准确率。

预知识

为了更好的讨论，本节会先给出我们要解决的问题，然后给出一个实际的案例。这节不会具体涉及到极大似然估计和贝叶斯估计的细节，但是会提出问题于后续方法理解。

问题前提

首先，我们有一堆数据$D=\{x_1,x_2,\ldots,x_n\}$，当然这些数据肯定不是随便产生的，我们就假设这些数据是以含有未知参数θ 某种概率形式（如Bernc 0-1分布）分布的。我们的任务就是通过已有的数据，来估计这个未知参数θ。估计这个参数的好处就在于，我们可以对外来的数据进行预测。

问题实例

假设一个抛硬币实验，我们之前不知道这些硬币是不是正反均匀的，也许硬币正反不等，假设正面向上设为1的概率为ρ，反面向上设为0为$（1-\rho）$.我次实验，得到两次正面，一次反面，即序列为$'110'$。这里，$D=(1,1,0)$，$\theta=\rho$。

符号说明

这里给出一些符号表示。可看到不理解时过来查看。

符号	含义
D	已有的数据(data)
θ	要估计的参数(parameter)
$p(\theta)$	先验概率(prior)
$p(\theta\|D)$	后验概率(posterior)
$p(D)$	数据分布(evidence)
$p(D\|\theta)$	似然函数(likelihood of θ w.r.t. D)
$p(x,\theta\|D)$	已知数据条件下的x,θ概率

方法介绍

这一节将会详细阐明极大似然估计和贝叶斯估计，要注意到两种方法在面对未知参数θ 时采用的不同态度。

极大似然估计

模型推导

极大似然估计法认为参数是固有的，但是可能由于一些外界噪声的干扰，使数据看起来不是完全由参数决定的。没关系，数学家们觉得，虽然有误差存在**在这个数据给定的情况下，找到一个概率最大的参数就可以了。**那问题其实就变成了一个条件概率最大的求解，即求使得$p(\theta|D)$ 最大的参数θ，形式化

$$\arg\max_{\theta} p(\theta|D)$$

而根据条件概率公式有

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

因为我们在极大似然估计中假设 θ 是确定的，所以$p(\theta)$就是一个常数。$p(D)$ 同样是根据已有的数据得到的，也是确定的，或者我们可以把其看作是对整个概率的一个归一化
候，求解公式(1) 就变成了求解

$$\arg\max_{\theta} p(D|\theta)$$

的问题。
(3) 式中的$p(D|\theta)$ 就是似然函数，我们要做的就是求一个是似然最大的参数，所以称为极大似然估计。
想求解这个问题，需要假设我们的数据是相互独立的。$D = \{x_1, x_2, x_3, \ldots, x_n\}$，这时候有

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta),$$

一般对(4)式取对数求解对数极大似然，就可以把连乘变成求和，然后求导取极值点就是要求的参数值，不在此赘述。

实例

为了便于理解，我们以之前的抛硬币实验作为实例。
回到当时我们一开始抛硬币实验，$D = (1,1,0)$, $\theta = \rho$ 的话，我们可以得到

$$\begin{aligned} p(D|\theta) &= p(x_1|\rho)p(x_2|\rho)p(x_3|\rho) \\ &= p(1|\rho)p(1|\rho)p(0|\rho) \\ &= \rho * \rho * (1-\rho) \end{aligned}$$

然后使用对数极大似然估计就可以得到参数ρ 的值了。

贝叶斯估计

考虑到这节对先验概率(prior)这个概念用的次数比较多，我们首先先介绍先验与后验概率是什么，怎么得到；其次会介绍贝叶斯估计模型的推导过程；最
例子来加深理解。

先验概率、后验概率

先验概率(prior)与后验概率(posterior)简称为**先验**和**后验**。这两个概念其实是来自于贝叶斯定理，相信学过概率论的一定有所了解。在此试作简单介绍。
之前提到的先验概率到底是什么呢？，毫无疑问必须得与放在一起来介绍。一个先一个后，我们肯定是针对同一个事物才有先后之分，如果针对两个事物
没有意义了么？那这个共同的对象，就是我们的参数θ。后验概率是指掌握了一定量的数据后我们的参数分布是怎么样的，表示为$p(\theta|D)$；那先验就是在
数据后我们的参数怎么分布。

看到这里，你可能会问：如果连数据都没有，我怎么知道我的参数是怎么分布的？你提出这个问题，就说明你是一个赤裸裸的频率派学家，你需要通过数
的参数！而这并不是贝叶斯派的考虑，贝叶斯估计最重要的就是那个先验的获得。虽然你这次的一组数据，比如说扔三次硬币产生的序列是（110）这样
是其实我根据我历史的经验来看，一枚硬币正反面其实很有可能是按照均匀分布来的，只不过可能因为你**抛得次数少**了所以产生了不是均匀分布的效果。
虑我以往的经验在里面。

你可能又会问：那你这个均匀分布不就是完全猜来的嘛，你怎么知道我这次是不是一样的硬币呢？没错！就是**“猜来的”**。先验在很多时候完全是假设，
有的数据是否吻合先验猜想，所以这里的猜很重要。还要注意，先验一定是与数据无关的，你不能看到了数据再做这些猜想，一定是**没有任何数据之前**你
参数的先验概率。

有个这部分知识，我们可以开始推导贝叶斯估计模型了。

模型推导

还是继续上面的模型，注意公式(2) 其实是一个很概括的模型，既没有对概率形式以及概率参数进行定义，也没有运用到参数固定与否的思想，所以公式
用于贝叶斯模型，我们仍然想对该式进行处理得出我们的贝叶斯估计方法。照抄下来(2) 式为

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

此时，这里面除了分母可以看作是一个归一化因子外，其余均是概率分布的函数。也就是说，无法再像极大似然估计那样将先验概率($p(\theta)$)看作一个常量。这时候就需要考虑
验概率了。我们这次把分母也展开来看看，根据全概率公式[1]得到

$$p(D) = \int_{\theta} p(D|\theta)p(\theta)d\theta.$$

我们来把这个式子(4)

$$p(D|\theta)=\prod_{i=1}^{n}p(x_i|\theta)$$

和式子(6)一起带入(2)式，得到

$$p(\theta|D)=\frac{(\prod_{i=1}^{n}p(x_i|\theta))p(\theta)}{\int_\theta(\prod_{i=1}^{n}p(x_i|\theta))p(\theta)d\theta}$$

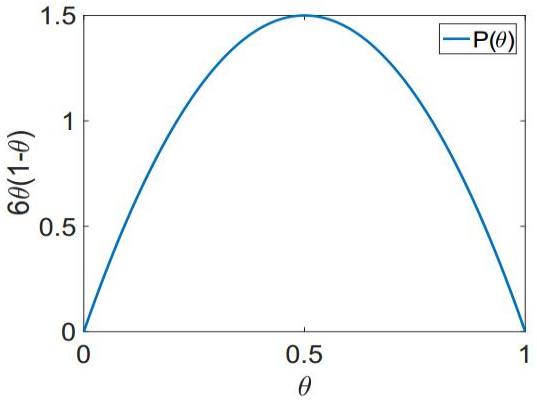
至此，我们就完成了对贝叶斯估计模型的推到过程。有人会问，怎么就完成了？还有那么长一段公式，我们怎么计算啊？其实仔细看看(7)式，其实这些是知道的，我们就通过下面的实例来详述。

实例

式(7)中的符号有先验，根据之前对先验的介绍，这是在没有数据之前我们就已经知道的函数了。知道是什么意思？不妨还是在那个抛硬币试验中，我们$\theta(\rho)$的先验概率是服从

$$f_\rho(\rho)=6\rho(1-\rho)$$

概率分布的。如图



然后$(\prod_{i=1}^{n}p(x_i|\theta))$也已经知道是$\rho*\rho*(1-\rho)$了。这时要的事情，其实就是把所有已知的全都一股脑带进去就可以了。有人问，已知概率分布怎么知道概率，我想这个概率论的书上找找。

但是，其实做到这一步，我们会发现虽然解决了问题，但是又会带来新的问题，因为在解决这一类贝叶斯估计的问题的时候，我们让参数以某种概率密度就会导致在计算过程中不可避免的高复杂度，人们为了计算上的方便，就提出不再是把所有的后验概率$p(\theta|D)$都找出来，而是仍然采用类似于极大似然想，来**极大后验概率(Maximum A Posterior)**，得到这种简单有效的叫做**MAP**（前面英文的首字母）的算法。下面我们再一步步介绍一下**MAP**。

极大后验概率(MAP)

虽然本节独自成为一节，但是其实是隶属于贝叶斯估计的，属于贝叶斯估计里面的一个trick，放弃一点的准确性，来极大提升算法性能。所以，这个部分模型，只能算是算法。
MAP（Maximum A Posterior）的理论依据是绝大部分情况下，参数值最有可能出现在概率最大点附近。为了说清楚MAP的来龙去脉，本节将首先介绍叶斯估计的参数进行预测，然后分析直接使用之前得到的后验概率有什么不好，最后介绍MAP算法做的工作。

使用贝叶斯估计的参数做预测

前一节中，我们通过贝叶斯估计得到了后验概率$p(\theta|D)$。那么这个后验概率能用来做什么呢？当然，就比如我们一直在说的那个例子，得到了数据$D=$还想预测第四次得到的结果什么是什么怎么办？我们当然就需要计算$p(1|D)$和$p(0|D)$看看谁大谁小，哪个更有可能发生。这里，为了泛化，我们将问化一下为

已知数据$D=(x_1,x_2,\ldots,x_n)$，预测新的数据x的值。

这个问题还有很多细节，比如先验概率，后验概率，数据分布等一些细节，因为前面已经介绍过了，这里为了突出重点，不再重复。在此需要关注的是，的数据的值，其实就是能够在**已知数据D的情况下，找到数据的数学期望**[2]。即求

$$E(x|D)=\int_x xp(x|D)dx.$$

也就是我们需要求$p(x|D)$，这该怎么办？其实这个式子比较迷惑人的点就在于，它内藏了一个参数，也就是x的分布其实与参数是有关的，但是又参数θ是服从某种概率分布
所有可能的情况都考虑就得到了

$$p(x|D) = \int_{\theta} p(x,\theta|D)d\theta$$

这一式子。
接下来还是运用基本的条件概率公式

$$p(x,\theta|D) = p(x|\theta,D)p(\theta|D).$$

对这一句公式的解释就是，x和θ在已知数据D的条件下的概率，等于x在已知θ和数据D的条件下的概率乘θ在已知数据D的条件下的概率。为什么我要费这个心来说这个，
了方便大家理解这个多维条件概率符号的含义，另一方面更重要的是右边式子的第一项$p(x|\theta,D)$可这样

$$p(x|\theta,D) = p(x|\theta)$$

化简。为什么？因为我们从数据里面得到的东西对一个新的数据来说，其实只是那些参数，所以对x而言，θ就是D，两者是同一条件。
那么(10)式就变成了[3]

$$p(x|D) = \int_{\theta} p(x,\theta|D)d\theta = \int_{\theta} p(x|\theta)p(\theta|D)d\theta.$$

$p(x|\theta)$是已知的(*例如在我们的问题里面可以是*$p(1|\rho)$*或者*$p(0|\rho)$)；$p(\theta|D)$也是已知的，我们在贝叶斯估计中已经通过(7)式求出来了。所以这个式子完全就是一个只含有
入(9)式完全可以计算出来数学期望。但是！这里面我忽略了一个事实，这里面存在什么困难呢？下面会帮助大家分析。

贝叶斯估计中的一个困难

还是回到(12)式，这里面的困难是参数是随机分布的，我们需要考虑到每一个可能的参数情况然后积分，这种数学上的简单形式，其实想要计算出来需
算。那我们不妨退而求其次，我找一个跟你差不多效果的后验概率，然后就只计算这个后验带入计算。那么什么样的后验概率和对所有可能的θ积分情况
想法就是，**找一个θ能够最大化后验概率**，怎么才能最大化后验概率呢？

MAP算法

其实最大化后验概率还是老一套，最大化(7)式，对(7)式观察发现，其实分母只是一个归一化的因子，并不是θ的函数。真正有效的其实就是要最大化我
于是使用

$$\theta_{MAP} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta)p(\theta)$$

这其实与极大似然估计形式上很相似，但是主要区别在于运用了一个先验概率在这个极大化里面。参数都已经计算出来了，其他过程，其实还是按照极大似然来做就行了，不
斯一样对所有可能的参数情况都考虑在求积分了。

总结

全文对比分析了极大似然估计和贝叶斯估计，在进行参数估计的过程中，极大似然估计是想让似然函数极大化，而考虑了MAP算法的贝叶斯估计，其实
概率极大化。主要区别在于估计参数中，一个考虑了先验一个没有考虑先验，主要区别看(3)，(13)式。

参考文献

1. 贝叶斯定理
   https://zh.wikipedia.org/wiki/%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%AE%9A%E7%90%86 ↩
2. Andrew’ s notes (note5)
   http://v.163.com/special/opencourse/machinelearning.html ↩
3. Pattern Recognition and Machine Learning
   http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.p

文章最后发布于: 2016-10-11