贝叶斯方法

By xudong

内容提要

贝叶斯公式
参数估计
- 极大似然估计
- 贝叶斯估计

贝叶斯分类器
贝叶斯分类器和其他分类器的比较

贝叶斯公式

这个大题目下包括三部分的内容，分别是贝叶斯公式的由来，对贝叶斯公式的解释，以及一个小例子。首先来看贝叶斯公式的由来。一上来就是公式可能有些不友好，不过也没办法，在我的能力范围内这是最顺畅的叙述方式了。

$$P(\theta|x) = \frac{P(\theta;x)}{P(x)} = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)}$$

以上两个等式中使用了条件概率的定义和全概率公式，把连等式中的第一项和最后一项拿出来就得到了贝叶斯公式。

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)}$$

下面来解释这个公式的含义。公式中θ表示一个事件，例如说扔一次硬币的结果，$P(\theta)$被称为先验，“先”表示实验之前，先验是指获得数据之前对于事件θ发生概率的预估，是对事件θ的最初认识。x表示实验数据，$P(\theta|x)$被称为后验，“后”表示实验之后，后验是指获得数据之后对于事件θ发生概率的估计，是获得数据之后对事件θ的重新认识。贝叶斯公式就是将最初的认识$P(\theta)$，转换成重新认识$P(\theta|x)$的机器，数据x决定了加工之后的结果。

之前的学习体会告诉我，一个简单又能一窥理论全貌的例子远比宏大完善的理论更加友好，所以用一个例子来说明上面叙述的理论。例子是这样的，有两个盛有饼干的碗，第一个碗中有一半圆饼干和一半方饼干，第二个碗中所有的都是圆饼干。现在随机的把一个碗给小明，小明从碗中拿出一个饼干，这个饼干是圆形的。问小明手中的碗是第一个的概率。

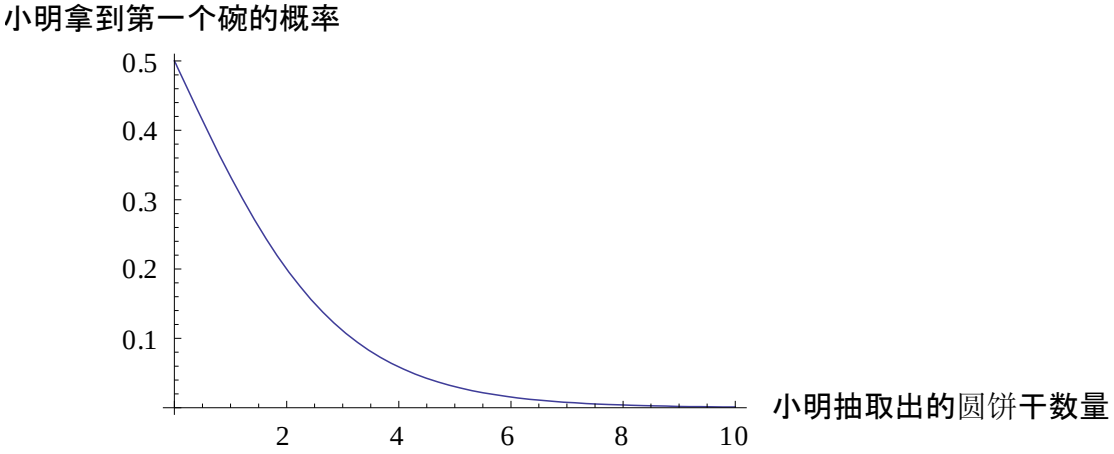
下面通过贝叶斯公式来计算小明拿到第一个碗的概率。first 表示小明拿到第一个碗，second 表示小明拿到第二个碗。r 是小明抽到圆饼干的数量，s 是小明抽到方饼干的数量。

$$P(\text{first}|r=1) = \frac{P(r=1|\text{first})P(\text{first})}{P(r=1|\text{first})P(\text{first}) + P(r=1|\text{second})P(\text{second})}$$

其中$P(r=1|\text{first}) = \frac{1}{2}$，$P(r=1|\text{second}) = 1$，$P(\text{first}) = P(\text{second}) = \frac{1}{2}$

由以上公式计算可知小明拿到第一个碗的概率是 1/3。

贝叶斯公式的一个特点是随着数据的增加，可以不断的调整估计。下图显示了随着小明抽取的圆饼干数量的增加，小明拿第一个碗的概率估计的变化。



还有另外一个关于贝叶斯方法的应用，简单有趣，最关键的是还十分有用。 话说 1968 美国一艘核潜艇意外沉没了，美国军方展开了大面积的搜索，花费了很大的人力物力还是没能找到失事的核潜艇。后来军方听取了一个数学家的建议，使用了一种基于贝叶斯方法的搜救方案，结果当然是成功的找到了失事潜艇。下面简要叙述一下具体的搜救方案，军方的专家们在两方面贡献了他们的背景知识。首先军方将整个海域划分成很多的子区域，军方中那些资深的舰长对每片海域出事的概率提供一个先验的估计。之后军方的搜救专家估计出如果潜艇的确在这片海域失事，搜救队能够发现失事潜艇的概率。这样每片海域都对应着一个潜艇的失事概率，每搜索完一个海域所有海域对应的失事概率会相应的发生变化。每搜索完一个海域，所有海域中失事概率最大的海域将是下一个被搜索的对象。

下面通过贝叶斯公式给出搜索完一个海域之后，该海域和其他海域失事概率的调整方案。p表示搜救之前在该搜救海域潜艇失事的概率，p'表示搜索之后该海域潜艇失事的概率，q 表示如果潜艇在该海域失事，搜救能够发现的概率。

$$p' = \frac{p(1-q)}{(1-p)+p(1-q)} = p\frac{1-q}{1-pq} < p.$$

r表示搜救之前其他海域潜艇失事的概率，r'表示搜救之后其他海域潜艇失事的概率。

$$r' = r\frac{1}{1-pq} > r.$$

【注】两个例子来源于 http://en.wikipedia.org/wiki/Bayesian_inference

参数估计

考虑这样一个问题，一枚硬币，抛掷出现正面的概率为 p，出现反面的概率为 1-p，但是参数 p 未知。为了估计参数 p 的取值，进行 10 次随机试验，出现了 3 次正面，7 次反面。现在我们已经获取了实验数据，问题是如何才能通过实验数据估计出参数 p 呢？下面介绍两

种方法，分别是极大似然估计和贝叶斯估计。

【极大似然估计】

极大似然估计的想法和统计物理中导出最概然分布想法是相同，这个想法就是我们所看到的，就是最可能发生的。以这个想法为基础就能够建立极大似然估计的方法。极大似然估计包括两个步骤，第一步写出实验数据 x 发生的概率$P(x|\theta)$，概率表示中含有待估的未知参数θ；第二步极大化目标函数$P(x|\theta)$，使得目标函数$P(x|\theta)$取极值的$\hat{\theta}$，就是参数θ的极大似然估计。下面以硬币的问题例具体说明这两个步骤。

第一步写出实验数据 x 发生的概率$P(x|\theta)$。参数θ在硬币问题中指的是出现正面的概率 p，实验数据 x 指的是 10 次实验中出现了 3 次正面和 7 次反面。如果实验中出现 h 次正面，t 次反面，那么出现该实验结果出现的概率为

$$P(x|\theta) = p^h(1-p)^t$$

【注】以上概率表达式假设正反面出现的次序确定

第二步极大化$P(x|\theta)$，得到参数θ的极大似然估计$\hat{\theta}$。由于极大化$P(x|\theta)$等价于极大化$\mathrm{Log}[P(x|\theta)]$，可以通过求解$\mathrm{Log}[P(x|\theta)]$的最大值来简化求解过程。

$$\mathrm{Log}[P(x|\theta)] = h\mathrm{Log}[p] + \mathrm{tLog}[1-p]$$

极值条件为

$$\frac{\partial \mathrm{Log}[P(x|\theta)]}{\partial p} = 0$$

将$\mathrm{Log}[P(x|\theta)] = h\mathrm{Log}[p] + \mathrm{tLog}[1-p]$代入极值条件得

$$\frac{\partial \mathrm{Log}[P(x|\theta)]}{\partial p} = \frac{h}{p} - \frac{t}{1-p} = 0$$

由等式$\frac{h}{p} - \frac{t}{1-p} = 0$可以求解出使得$\mathrm{Log}[P(x|\theta)]$最大的参数 p

$$\hat{p} = \frac{h}{h+t}$$

这就是硬币正面出现概率的极大似然估计。将具体的实验结果代入以上公式计算这枚硬币出现的概率为 p=3/10。

可能有人会想，为啥要搞的这么麻烦，$\hat{p} = h/(h+t)$是多么的直白明了，即使不使用极大似然的方法仍然能够写出$\hat{p} = h/(h+t)$。对于硬币的问题，我们的确可以更快捷的凭借直觉给出参数 p 的估计，那是因为参数 p 有明确的含义，p 表示概率，我们很容易想到使用频率来估计概率。但是有的问题中参数θ没有明确的含义，我们就很难通过直觉来得到参数θ的估计量$\hat{\theta}$。简单来说我们的直觉能够在待估参数有明确含义的时候提供便捷，但是当待估参数没有明确含义的时候我们的直觉只能靠边站，事情交由极大似然估计来解决。

极大似然估计有几个很好的性质，这里只列出极大似然估计的性质，详细内容大家可以参见《统计学全教程》9.4~9.8 节。

- 极大似然估计是相合的
- 极大似然估计是渐进正态的
- 极大似然估计是渐近最优的

【贝叶斯估计】

还是回到硬币的问题，我们通过极大似然估计得到硬币出现正面的概率是 3/10，但是生活经验告诉我们硬币正反面出现的概率相等都是 1/2。到底我们应该相信那个结果呢？一种好的方法就是将生活经验和实验数据两个因素综合在一起考虑，贝叶斯估计很好的做到了

这一点。

贝叶斯估计可以分为三个步骤来实现。第一步确定先验，第二步写出似然函数并计算后验，第三步根据后验计算贝叶斯估计。下面通过硬币的例子来说明贝叶斯估计的实现步骤。

第一步确定先验，我们使用的先验分布是p~Beta$[\alpha,\beta]$，Beta$[\alpha,\beta]$具体是这个样子

$$f(p)=\frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}p^{\alpha-1}(1-p)^{\beta-1}$$

不过知道他的含义更加重要，Beta$[\alpha,\beta]$相当于实验之前已经进行了$\alpha+\beta$次扔硬币的实验，出现了α次正面和β次反面。

第二步写出似然函数，并计算后验。

$$P(\theta|x)\propto P(x|\theta)P(\theta)$$

$$P(x|\theta)P(\theta)=p^h(1-p)^t\frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}p^{\alpha-1}(1-p)^{\beta-1}$$

添加归一化系数之后我们就能够得到后验分布

$$P(\theta|x)=\frac{\Gamma[\alpha+\beta+h+t]}{\Gamma[\alpha+h]\Gamma[\beta+t]}p^{\alpha+h-1}(1-p)^{\beta+t-1}$$

第三步根据后验计算贝叶斯估计。贝叶斯估计为后验均值。

$$\hat{\theta}=\int\theta P(\theta|x)\,d\theta$$

将后验的具体表达代入可得

$$\hat{p}=\int p\frac{\Gamma[\alpha+\beta+h+t]}{\Gamma[\alpha+h]\Gamma[\beta+t]}p^{\alpha+h-1}(1-p)^{\beta+t-1}\,dp$$

计算得

$$\hat{p}=\frac{\alpha+h}{\alpha+\beta+h+t}$$

以上我们结束了贝叶斯估计的三个步骤，代入具体的数据$\alpha=200$，$\beta=200$，$h=3$，$t=7$计算得

$$p=\frac{\alpha+h}{\alpha+\beta+h+t}=\frac{203}{410}=0.495$$

我们的先验知识对结果产生了很大的影响，不添加先验时极大似然估计的结果是 p=3/10，添加先验之后，较少的实验数据只对先验做出微小的调整，贝叶斯估计的结果是$p=0.495$。可以看出样本较少时先验对结果产生重要的影响，但随着样本量的增加先验的影响逐渐减弱，并且贝叶斯估计的结果趋近极大似然估计的结果。这个结论不仅仅对于硬币问题成立，对于所有的贝叶斯估计，随着样本量的增加先验的影响逐渐减弱，贝叶斯估计趋近极大似然估计。这个性质很像人的思维过程，人们总是根据生活现实修正原有想法，原有的想法如果和现实相一致，这些想法将得到加强，否则原有的想法会被削弱，取而代之的是更加接近事实的想法，总之人们的想法会随着经验的积累更加贴近事实，这就是随着样本量的增加，先验的作用逐渐减弱的过程。豆瓣上有一本书叫做《Bayesian Brain》，好像说的就是这件事情，不过我还没有看过。听过这样一个说法，人感受的信息中只有 20%的信息来源于外界环境，剩余的 80%是人脑根据已存储的信息重构的。

《贝叶斯脑》http://www.douban.com/subject/2507536/

有点跑题了，回到正题，对以上的讨论做一下总结，以上的讨论有两点是重要的

- 如果样本量小，先验知识又是可获得的，贝叶斯估计能够将先验知识和样本信息整合起来获得更好的效果。
- 如果样本量较大，先验产生的作用很小，可以忽略。贝叶斯估计趋近极大似然估计，只

反应样本信息

下面再简单叙述两个例子来说明一下贝叶斯方法的应用。

第一个是豆瓣评分，如果某项内容有很多人评分，那评分的结果就能够体现用户对这项内容的评价，但是如果一开始评分用户很少的话，几个用户评分结果很难体现大多数人对这项内容的评价。最极端的情况，如果一项内容只有一个人评分，这项内容很容易变成最好或者最坏，这样的结果十分的片面不准确。豆瓣处理这个问题的方式有点偷懒，干脆不给评价数量少于 10 的内容给出评价，虽然这种方法十分赖皮，但这个赖皮方法很可能是争议最少成本最低的方法了。对于评分的这个问题，我们可以尝试使用贝叶斯方法来解决。假设现在布拉德皮特拍摄了一部新电影，为了使用贝叶斯方法来给这部电影评分，我们首先需要构造一个先验。这个先验包含两部分的内容 1）先验中有多少人评分 2）先验中这些人评分结果是什么？再次强调一下，先验中的评分人数和结果并非真实的用户评价而是在布拉德皮特之前所有电影评价的基础上合理假设的。例如先验中评分的人数可以假设为布拉德皮特之前电影平均评价人数的 1/10，先验中的评分结果直接使用布拉德皮特之前所有电影的平均得分。随着评价用户的增加，先验对评价结果产生的影响逐渐下降，最终的评价结果可能高于先验也可能低于先验，这将由给出评价的豆瓣 er 决定。

另外一个例子的思路和实现方法和豆瓣评分完全相同，只是使用的场合发生了变化。这个例子的场景是贝叶斯网的参数估计。为了构建贝叶斯网，很重要的一件事情就是估计贝叶斯网中的条件概率，也就是要估计，给定事件 A 发生事件 B 发生的概率，我们可以用给定事件 A 发生，事件 B 发生的频率来估计条件概率$P(B|A)$。假设$P(B|A) = 0.08$，如果样本量很大的话，天下太平估计$P(B|A)$不会出现什么问题，但是如果样本量很少，比如说事件 A 只发生了 5 次，这就出现麻烦了，因为$P(B|A) = 0.08$，最可能的情况就是 A 发生了 5 次，B 一次也没发生。使用频率估计的结果是$P(B|A) = 0/5 = 0$，这是一个向下偏误的结果，而且出现 0 对贝叶斯网推断也产生非常不利的影响。为了解决这个问题我们使用解决豆瓣评分问题相同的方法，添加先验，最终的估计变成了如下形式

$$\frac{n_B + mp}{n_A + m}$$

n_A，n_B分别表示事件 A，B 发生的次数，p 表示对条件概率$P(B|A)$预估计。m，是一个正数，m 越大先验对结果产生的影响越大。

上面两个例子再次说明了当样本量很小时，可以使用贝叶斯估计结合先验知识改善估计的质量。

贝叶斯分类器

我们再回过头来考虑小明的例子，现在问题不是计算概率，而是判断小明拿到的是第一个碗还是第二个碗。原来概率计算的问题变成了分类问题，在这个问题中我们需要根据小明抽取出的饼干的颜色来推断小明拿到的是第一个碗还是第二个碗，推断的依据是什么呢？能够想到的最直接的方法就是比较给定抽取结果的情况下，小明拿到第一个碗的概率和小明拿到第二个碗的概率，并且认为概率较大的结果是真正发生的结果。这就是贝叶斯分类器的想法。

为了叙述简单，我们假设总共只有两类，根据贝叶斯公式我们可以写出给定数据的条件下，第一类的概率和第二类的概率。

$$P(Y = 1|x) = \frac{\pi_1 P(x|Y = 1)}{\pi_1 P(x|Y = 1) + \pi_2 P(x|Y = 2)}$$

$$P(Y=2|x)=\frac{\pi_2 P(x|Y=2)}{\pi_1 P(x|Y=1)+\pi_2 P(x|Y=2)}$$

分类问题转化成比较这两个条件概率大小的问题，可以看出比较这两个条件概率的大小，因为两个条件概率的分母都是相同的，只需要比较$\pi_1 P(x|Y=1)$，$\pi_2 P(x|Y=2)$足够了。π_1，π_2就是数据中第一类和第二类所占的百分比，很容易计算。难点在于如何计算$P(x|Y=1)$，$P(x|Y=2)$，这两个条件概率。

如果数据是连续变量，这里的条件概率应该替换为条件概率密度。直接估计概率密度难度比较大，一般的解决办法是假设条件概率满足某种分布，然后利用数据估计分布中的待定参数。如果我们假设条件概率满足高斯分布，那我们就得到了高斯分类器。