> 在贝叶斯学派的观点中，先验概率、后验概率以及共轭分布的概念非常重要。而在机器学习中，我们阅读很多资料时也要频繁地跟他们打交道。所以理清这些概念很有必要。

欢迎关注白马负金羁的博客 http://blog.csdn.net/baimafujinji，为保证公式、图表得以正确显示，强烈建议你从该地址上查看原版博文。本博客主要关注方向包括：数字图像处理、算法设计与分析、数据结构、机器学习、数据挖掘、统计分析方法、自然语言处理。

贝叶斯定理：一个例子

其实我们在之前介绍朴素贝叶斯分类器时就介绍过它，如果你有点忘了，这里就通过一个例子来帮你回忆一下。

假设有一所学校，学生中60%是男生和40%是女生。女生穿裤子与裙子的数量相同；所有男生穿裤子。现在有一个观察者，随机从远处看到一名学生，因为很远，观察者只能看到该学生穿的是裤子，但不能从长相发型等其他方面推断被观察者的性别。那么该学生是女生的概率是多少?

用事件 G 表示观察到的学生是女生，用事件 T 表示观察到的学生穿裤子。于是，现在要计算的是条件概率 $P(G|T)$ ，我们需要知道：

- $P(G)$ 表示一个学生是女生的概率。由于观察者随机看到一名学生，意味着所有的学生都可能被看到，女生在全体学生中的占比是 40% ，所以概率是 $P(G)=0.4$ 。注意，这是在没有任何其他信息下的概率。这也就是先验概率。后面我们还会详细讨论。
- $P(B)$ 是学生不是女生的概率，也就是学生是男生的概率，这同样也是指在没有其他任何信息的情况下，学生是男生的先验概率。 B 事件是 G 事件的互补的事件，于是易得 $P(B)=0.6$ 。
- $P(T|G)$ 是在女生中穿裤子的概率，根据题目描述，女生穿裙子和穿裤子各占一半，所以 $P(T|G)=0.5$ 。这也就是在给定 G 的条件下，T 事件的概率。
- $P(T|B)$ 是在男生中穿裤子的概率，这个值是1。
- $P(T)$ 是学生穿裤子的概率，即任意选一个学生，在没有其他信息的情况下，该名学生穿裤子的概率。根据全概率公式 $P(T)=\sum_{i=1}^{n}P(T|A_i)P(A_i)=P(T|G)P(G)+P(T|B)P(B)$ ，计算得到 $P(T)=0.5\times 0.4+1\times 0.6=0.8$。

根据贝叶斯公式

$$P(A_i|T)=\frac{P(T|A_i)P(A_i)}{\sum_{i=1}^{n}P(T|A_i)P(A_i)}=\frac{P(T|A_i)P(A_i)}{P(T)}$$

基于以上所有信息，如果观察到一个穿裤子的学生，并且是女生的概率是

$$P(G|T)=\frac{P(T|G)P(G)}{P(T)}=0.5\times 0.4\div 0.8=0.25.$$

先验概率（Prior probability）

在贝叶斯统计中，先验概率分布，即关于某个变量 X 的概率分布，是在获得某些信息或者依据前，对 X 之不确定性所进行的猜测。这是对不确定性（而不是随机性）赋予一个量化的数值的表征，这个量化数值可以是一个参数，或者是一个潜在的变量。

先验概率仅仅依赖于主观上的经验估计，也就是事先根据已有的知识的推断。例如， X 可以是投一枚硬币，正面朝上的概率，显然在我们未获得任何其他信息的条件下，我们会认为 $P(X)=0.5$；再比如上面例子中的，$P(G)=0.4$。

在应用贝叶斯理论时，通常将先验概率乘以似然函数（Likelihood Function）再归一化后，得到后验概率分布，后验概率分布即在已知给定的数据后，对不确定性的条件分布。

似然函数（Likelihood function）

似然函数（也称作似然），是一个关于统计模型参数的函数。也就是这个函数中自变量是统计模型的参数。对于观测结果 $\mathbf{x}$ ，在参数集合 θ 上的似然，就是在给定这些参数值的基础上，观察到的结果的概率 $\mathcal{L}(\theta)=P(\mathbf{x}|\theta)$ 。也就是说，似然是关于参数的函数，在参数给定的条件下，对于观察到的 $\mathbf{x}$ 的值的条件分布。

似然函数在统计推断中发挥重要的作用，因为它是关于统计参数的函数，所以可以用来对一组统计参数进行评估，也就是说在一组统计方案的参数中，可以用似然函数做筛选。

你会发现，“似然”也是一种“概率”。但不同点就在于，观察值 $\mathbf{x}$ 与参数 θ 的不同的角色。概率是用于描述一个函数，这个函数是在给定参数值的情况下的关于观察值的函数。例如，已知一个硬币是均匀的（在抛落中，正反面的概率相等），那连续10次正面朝上的概率是多少？这是个概率。

而似然是用于在给定一个观察值时，关于描述参数的函数。例如，如果一个硬币在10次抛落中正面均朝上，那硬币是均匀的（在抛落中，正反面的概率相等）概率是多少？这里用了概率这个词，但是实质上是“可能性”，也就是似然了。

后验概率（Posterior probability）

后验概率是关于随机事件或者不确定性断言的条件概率，是在相关证据或者背景给定并纳入考虑之后的条件概率。后验概率分布就是未知量作为随机变量的概率分布，并且是在基于实验或者调查所获得的信息上的条件分布。“后验”在这里意思是，考虑相关事件已经被检视并且能够得到一些信息。

后验概率是关于参数 θ 在给定的证据信息 X 下的概率，即 $P(\theta|X)$ 。若对比后验概率和似然函数，似然函数是在给定参数下的证据信息 X 的概率分布，即 $P(X|\theta)$ 。二者有如下关系：

- 我们用 $P(\theta)$ 表示概率分布函数，用 $P(X|\theta)$ 表示观测值 X 的似然函数。后验概率定义为 $P(\theta|X)=\frac{P(X|\theta)P(\theta)}{P(X)}$，注意这也是贝叶斯定理所揭示的内容。
- 鉴于分母是一个常数，上式可以表达成如下比例关系（而且这也是我们更多采用的形式）：$Posterior\ probability \propto Likelihood \times Prior\ probability$

Gamma 函数

Gamma函数 $\Gamma(x)$ 定义为

$$\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-t}dt$$

通过分部积分法，可以很容易证明Gamma函数具有如下之递归性质

$$\Gamma(x+1)=x\Gamma(x)$$

也是便很容易发现，它还可以看做是阶乘在实数集上的延拓，即

$$\Gamma(x)=(x-1)!$$

在此基础上，我们还可以定义Beta函数如下

$$\mathbf{B}(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Beta函数的另外一种定义形式为（注意这两种定义是等价的）

$$\mathbf{B}(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}dt$$

Beta 分布

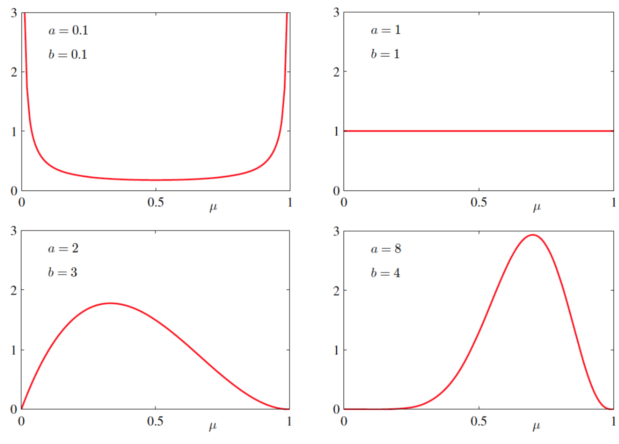
之所以提到Gamma函数，那是因为在定义Beta分布时我们会用到它。Beta分布的概率密度函数（PDF）定义为：

$$Beta(\theta|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

或

$$Beta(\theta|a,b)=\frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

可见，Beta分布有两个控制参数 a 和 b，而且当这两个参数取不同值时，Beta分布的PDF图形可能会呈现出相当大的差异。



Beta 分布的均值和方差分别有下面两式给出

$$E[\theta] = \frac{a}{a+b}$$
$$var[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$$

共轭分布

Conjugate Prior Definition:
A family F of prior distribution P(\theta) is conjugate to a likelihood P(Data | \theta) if the posterior P(\theta | Data) is also in F.

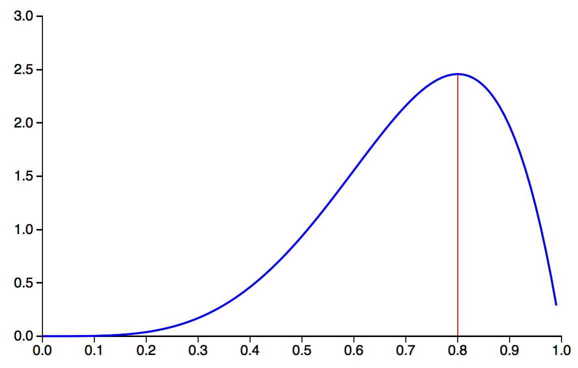
我们还是从一个例子讲起。假如你有一个硬币，它有可能是不均匀的，所以投这个硬币有 θ 的概率抛出Head，有 $(1-\theta)$ 的概率抛出Tail。如果抛了五次这个硬币，有三次是Head，有两次是Tail，这个 θ 最有可能是多少呢？如果你必须给出一个确定的值，并且你完全根据目前观测的结果来估计 θ，那么显然你会得出结论 $\theta = \frac{3}{5}$。

但上面这种点估计的方法显然有漏洞，这种漏洞主要体现在实验次数比较少的时候，所得出的点估计结果可能有较大偏差。大数定理也告诉我们，在重复实验中，随着实验次数的增加，事件发生的频率才趋于一个稳定值。一个比较极端的例子是，如果你抛出五次硬币，全部都是Head。那么按照之前的逻辑，你将估计 θ 的值等于 1。也就是说，你估计这枚硬币不管怎么投，都朝上！但是按正常思维推理，我们显然不太会相信世界上有这么厉害的硬币，显然硬币还是有一定可能抛出Tail的。就算观测到再多次的Head，抛出Tail的概率还是不可能为0。

前面介绍的贝叶斯定理或许可以帮助我们。在贝叶斯学派看来，参数 θ 不再是一个固定的值了，而是满足一定的概率分布！回想一下前面介绍的先验概率和后验概率。在估计 θ 时，我们心中可能有一个根据经验的估计，即先验概率，$P(\theta)$。而给定一系列实验观察结果 X 的条件下，我们可以得到后验概率为

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

在上面的贝叶斯公式中，$P(\theta)$ 就是个概率分布。这个概率分布可以是任何概率分布，比如高斯分布，或者刚刚提过的 Beta 分布。下图是Beta(5,2)的概率分布图。如果我们将这个概率分布作为 $P(\theta)$，那么我们在还未抛硬币前，便认为 θ 很可能接近于0.8，而不大可能是个很小的值或是一个很大的值。换言之，我们在抛硬币前，便估计这枚硬币更可能有0.8的概率抛出正面。



虽然 $P(\theta)$ 可以是任何种类的概率分布，但是如果使用Beta 分布，会让之后的计算更加方便。我们接着继续看便知道这是为什么了。况且，通过调节 Beta 分布中的 a 和 b，你可以让这个概率分布变成各种你想要的形状！Beta 分布已经很足够表达我们事先对 θ 的估计了。

现在我们已经估计好了 $P(\theta)$ 为一个 Beta 分布，那么 $P(X|\theta)$ 是多少呢？其实就是个二项（Binomial）分布。继续以前面抛5次硬币抛出3次Head的观察结果为例，$X = \text{抛5次硬币3次结果为}Head\text{的事件}$，则 $P(X|\theta) = C_2^5\theta^3(1-\theta)^2$。

贝叶斯公式中分母上的 $P(X)$ 是个Normalizer，或者叫做边缘概率。在 θ 是离散的情况下，$P(X)$ 就是 θ 为不同值的时候，$P(X|\theta)$ 的求和。例如，假设我们事先估计硬币抛出正面的概率只可能是0.5或者0.8，那么 $P(X) = P(X|\theta=0.5) + P(X|\theta=0.8)$，计算时分别将 $\theta = 0.5$ 和 $\theta = 0.8$ 代入到前面的二项分布公式中。而如果我们采用 Beta 分布，θ的概率分布在[0,1]之间是连续的，所以要用积分，即

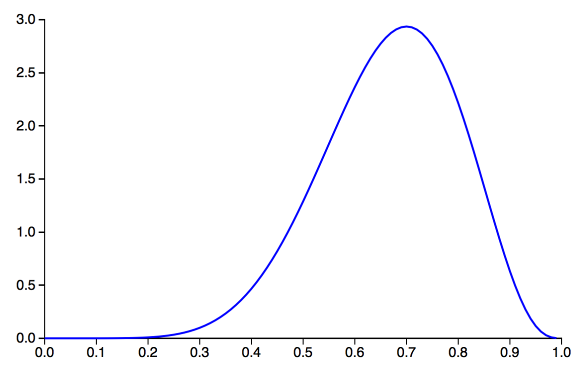
$$P(X) = \int_0^1 P(X|\theta)P(\theta)d\theta$$

下面的证明就告诉我们：$P(\theta)$ 是个 Beta 分布，那么在观测到“$X = \text{抛5次硬币中出现3个}head$”的事件后，$P(\theta|X)$ 依旧是个 Beta 分布！只是这个概率分布的形状因为观测的事件而发生了变化。

$$\begin{aligned}
P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int_0^1 P(X|\theta)P(\theta)d\theta} \\
&= \frac{C_2^5\theta^3(1-\theta)^2\frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 C_2^5\theta^3(1-\theta)^2\frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta} \\
&= \frac{\theta^{(a+3-1)}(1-\theta)^{(b+2-1)}}{\int_0^1 \theta^{(a+3-1)}(1-\theta)^{(b+2-1)}d\theta} \\
&= \frac{\theta^{(a+3-1)}(1-\theta)^{(b+2-1)}}{\mathbf{B}(a+3,b+2)} \\
&= Beta(\theta|a+3,b+2)
\end{aligned}$$

因为观测前后，对 θ 估计的概率分布均为 Beta 分布，这就是为什么使用 Beta 分布方便我们计算的原因了。当我们得知 $P(\theta|X) = Beta(\theta|a+3, b+2)$后，我们就只要根据 Beta 分布的特性，得出 θ 最有可能等于多少了。（即 θ 等于多少时，观测后得到的 Beta 分布有最大的概率密度）。

例如下图，仔细观察新得到的 Beta 分布，和上一图中的概率分布对比，发现峰值从0.8左右的位置移向了0.7左右的位置。这是因为新观测到的数据中，5次有3次是head（60%），这让我们觉得，θ 没有0.8那么高。但由于我们之前觉得 θ 有0.8那么高，我们觉得抛出head的概率肯定又要比60%高一些！这就是Bayesian方法和普通的统计方法不同的地方。我们结合自己的先验概率和观测结果来给出预测。



如果我们投的不是硬币，而是一个多面体（比如骰子），那么我们就要使用 Dirichlet 分布了。使用Dirichlet 分布之目的，也是为了让观测后得到的posterior probability依旧是 Dirichlet 分布。关于 Dirichlet 分布的话题我们会在后续的文章中继续介绍。

到此为止，我们终于可以引出“共轭性”的概念了！后验概率分布（正比于先验和似然函数的乘积）拥有与先验分布相同的函数形式。这个性质被叫做共轭性（Conjugacy）。共轭先验（conjugate prior）有着很重要的作用。它使得后验概率分布的函数形式与先验概率相同，因此使得贝叶斯分析得到了极大的简化。例如，二项分布的参数之共轭先验就是我们前面介绍的 Beta 分布。多项式分布的参数之共轭先验则是 Dirichlet 分布，而高斯分布的均值之共轭先验是另一个高斯分布。

总的来说，对于给定的概率分布 $P(X|\theta)$，我们可以寻求一个与该似然函数，即$P(X|\theta)$， 共轭的先验分布 $P(\theta)$，如此一来后验分布 $P(\theta|X)$ 就会同先验分布具有相同的函数形式。而且对于任何指数族成员来说，都存在有一个共轭先验。

参考文献

[1] 以上内容部分引自“胖胖小龟宝”在http://bbs.pinggu.org/上的帖子

[2] Pattern Recognition And Machine Learning, Christopher Bishop

[3] 抛硬币的例子来自http://maider.blog.sohu.com/306392863.html

- 上一篇 我的LaTeX秘籍（不断更新中）
- 下一篇 蒙特卡洛采样之拒绝采样（Reject Sampling）