理解共轭先验

原文作者：TCB

原文链接：http://lesswrong.com/lw/5sn/the_joys_of_conjugate_priors/

翻译：赵常凯

2013.1.31

对于应用贝叶斯理论的人来说，当你观察一个事件 x ,你预估计并给出其内部参数 θ ，表示你对于事件 x 发生的置信程度。

如果你熟悉贝叶斯方法，当你每次观测到新的 x 数据时你就会更新你预先给出的参数 θ。那么就要问了，你新观测到的 x 样本点对于你改变样本参数 θ 的影响有多大？这就取决与你一开始对参数 θ 的确定程度。如果你给出的参数 θ 基于你经过上千上万次认真实验得到的因此你很确定你的参数值，那么单一的新数据不会对参数有多大影响。但是如果你的参数 θ 的估计仅仅是从一个不可靠的朋友那听来的，那么新数据对于你重新估计参数值 θ 的影响就会大很多。

当然，当你重新估计 θ 的时候，你也要重新估计你新参数值的确定程度（置信程度）。换个方式说，你就是要计算 θ 的可能值的新概率分布。新概率分布为 $P(\theta|x)$,其计算可以使用贝叶斯法则：

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

$P(x|\theta)$ 表示以预估 θ 为参数的 x 概率分布，可以直接求得。$P(\theta)$ 是已有原始的 θ 概率分布。不同设定参数的精确度决定了你对 θ 的置信程度。所以分子部分可以直接计算求得。分母部分的计算很棘手。对于任意分布形式，积分计算可能会有很多困难。

也许你并不是想找出参数 θ 的整个分布。你只是想求得一个最优值来做预测。如果这是你的目标，那你就可以选取分布 $P(\theta|x)$ ，求 θ 的值使得 $P(\theta|x)$ 最大，作为新的参数。由于我们已经获得了 $P(\theta|x)$ 的形式，所以可以得到对于参数的置信程度。（可以理解为最大似然法求参数）

实际中，利用最大似然法求参数 θ 通常很困难。因为有局部最优解的存在，还有优化问题中的一

些普遍问题。对于足够简单的分布，可以利用 EM 算法保证参数收敛到一个局部最优解。但是对于复杂得多的分布，这个方法就变得力不从心，这就要利用近似算法。所以要尽量保证 $P(x|\theta)$ 和 $P(\theta)$ 简单。$P(x|\theta)$ 分布的选择是模型选择的问题，选择复杂的模型可以更好的反应数据的深层形式，但也会增加更多的时间和内存开销。

假设在确定 $P(\theta)$ 分布形式之前我们先选择模型的形式。那么如何确定 $P(\theta)$ 的最佳的分布形式？注意每次你观测一个新数据的时候，你就要计算一次上面等式。这样在观察数据的过程中，你就要乘上许多不同的概率分布。如果 $P(\theta)$ 分布没有选择好，$P(\theta)$ 会很快变得非常麻烦。

聪明的你会发现，选取 $P(\theta)$ 作为 $P(x|\theta)$ 分布的共轭先验。如果 $P(x|\theta)$ 乘以 $P(\theta)$ 然后归一化结果后其形式和 $P(\theta)$ 的形式一样，那么我们就说 $P(\theta)$ 共轭于 $P(x|\theta)$。

注：$P(x|\theta)$ 我们也称作似然函数。先验概率 $P(\theta)$ 和似然函数的乘积，然后归一化得到后验概率 $P(\theta|x)$。共轭先验的定义为：如果后验概率分布和先验概率分布有相同的形式（如同为指数族分布），则后验概率分布和先验概率分布统称共轭分布。那么先验概率 $P(\theta)$ 称为似然函数的共轭先验。

考虑一个离散情况的例子：投掷一个非均匀硬币（正反面概率不相等），可以使用参数为 θ 伯努利模型，那么结果 x 的分布形式为：（关于伯努利分布可以参考其他资料）

$$P(x|\theta)=\theta^{x}(1-\theta)^{1-x}$$

其共轭先验为 beta 分布，具有两个参数 α 和 β，我们称之为超参数（hyperparameters）。简单解释就是，这两个参数决定了我们的 θ 参数。

Beta 分布形式为：

$$P(x|\alpha,\beta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

同样 θ 为硬币为正面的概率。取值范围为 0 到 1。所以这个方程是归一化的方程。

假如你观察一次投硬币 x 事件然后要更新关于参数 θ 的置信度。Beta 函数的分母是一个归一化测常数，计算 $P(\theta|x)$ 的时候可以忽略它，只要计算完后再归一化即可。

$$
\begin{aligned}
P(\theta \mid x) &\propto P(x \mid \theta)P(\theta) \\
&\propto (\theta^{x}(1-\theta)^{1-x})(\theta^{\alpha-1}(1-\theta)^{\beta-1}) \\
&= \theta^{x+\alpha+1}(1-\theta)^{(1-x)+\beta-1}
\end{aligned}
$$

归一化这个等式后会得到另一个 beta 分布，就是伯努利分布的共轭先验。

如果对二项分布熟悉，beta 分布的分子部分和二项分布非阶乘部分很相似，归一化后得到的 beta 分布为：

$$
P(\theta \mid \alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}
$$

注：二项分布的形式为：

$$
Bin(m \mid N,\theta)=\binom{N}{m}\theta^{m}(1-\theta)^{N-m}=\frac{N!}{(N-m)!m!}\theta^{m}(1-\theta)^{N-m}
$$

也可以写作：

$$
Bin(\alpha \mid \alpha+\beta,\theta)=\binom{\alpha+\beta}{\alpha}\theta^{\alpha}(1-\theta)^{\beta}
$$

可以看出，beta 分布和二项式分布几乎是一样的。他们最大的不同点是，beta 分布是带有预设定参数 α 和 β 关于 θ 的函数。而二项分布是带有预设定参数 θ 和 $\alpha+\beta$ 关于 α 的函数。很显然 beta 分布共轭于二项式分布。

另外一个区别是：beta 分布用伽马函数作为归一化系数，而二项分布使用阶乘系数。回忆伽马函数只是将系数向实数域的一个扩展。这样就允许 α 和 β 为任意正实数。而二项式的系数只能定义为任意正整数。关于更多关于二项分布，beta 分布，gama 函数的内容见链接 http://www.mhtl.uwaterloo.ca/courses/me755/web_chap1.pdf。

现在想一下这些问题，共轭先验有什么意义？它仅仅是我们的一种数学计算工具吗？答案显然不是，它有更深的意义对于 beta 分布的形式。

考虑上面的内容，如果你已经观察了很多数据，那么再观察另外一个数据对于你对模型的认识并

不会改变多少。如果，你仅仅一开始观测了少量数据，那么观察另外的单一数据对于你的模型参数置信度影响就会很大。你可以通过共轭先验的形式获得这个直觉上的判断。

考虑投硬币的例子，我们设 α 和 β 为得到正面和反面的次数，将其作为 beta 分布的参数，实验有两种情况：第一种我们投 10 次，3 次正面 7 次反面。第二种投 10000 次，3000 次正面，7000 次反面。从这个 beta 分布中很容易得到两种情况下各自“声称正面概率为 30%”的置信度的不同。

（如果我们对一个模型没有任何先验知识，我们可以将 beta 分布的两个参数 α 和 β 都等于 1，beta 分布变成均匀分布。或者将两个系数都等于 N+1。将这个两个超参数设为小于 1 的值，那么模型就具有了“负数据”，可以让我们避免对于真实数据中带有噪音的参数 θ 过度拟合。）

注：上面一段有异议，关于两个参数 α 和 β 都等于 1，表示模型没有先验知识是有争议的。Haldane 先验 alpha = beta = 0 and Jeffreys' 先验模型, alpha = beta = 0.5。

概括一下，beta 分布是伯努利分布和二项式分布的共轭先验。在做贝叶斯理论相关计算时非常有用和有效。也可使用实数和虚数的先验数据。其他共轭先验如狄利克雷分布是多项式分布的共轭先验，原理是相似的。

关于共轭先验是一个重要的概念，笔者在读 PRML 这本书时会很多相关的概率分布知识，对于理解这本书起着关键的作用。

交流关于模式识别，机器学习的内容可以加我 QQ：417267003

翻译水平有限，望多指教。

关于共轭先验的关系可以看这里。

http://www.johndcook.com/conjugate_prior_diagram.html#binomial