贝叶斯学习中基于贝叶斯判别分析的先验分布选取*)

胡振宇[1] 林士敏[1] 陆玉昌[2]

(广西师范大学计算机科学系 桂林541004)[1] (清华大学计算机科学与技术系 北京100084)[2]

Choosing a Suitable Prior for Bayesian Learning Based on Bayesian Discrimination

HU Zhen-Yu[1] LIN Shi-Min[1] LU Yu-Chang[2]

(Department of Computer Science, Guangxi Normal Unviersity, Guilin 541004)[1]

(Department of Computer Science and Technology, Tsinghua Univeristy, Beijing 100084)[2]

Abstract In this paper we propose an experimental method to choose a prior distribution. Different from many researchers, who offered lots of principles that separated from sample information, we consider it a Bayesian discrimination problem combining with the sample information. We introduce the concept of Posterior belief about prior distributions. With the well-known Bayes theorem we give out a formula to calculate it and propose a method to discriminate a prior between prior distributions-- Highest Posterior Belief (HPB). We also show that under certain condition, the HPB method is identical with the ML-II method.

Keywords Machine learning, Prior distribution, Bayesian discrimination, Prior belief, Posterior belief

先验分布的选取是贝叶斯学习乃至整个贝叶斯方法中的一个基本问题。文[1]指出，在一般条件下，当样本数据足够多时，先验的选取对结果的影响很小，任何的先验都将得出大致一样的结果。也就是说，贝叶斯学习具有先验鲁棒性。但是，确实存在个别情况，某些先验会导致后验的不相容。况且，相容性是当样本数据趋于无穷多时的渐近特性，即使后验是相容的，当样本数据较少时，先验的选取还是能对后验产生明显的影响。

假设有一些可供选用的先验分布，我们面临的问题是如何在其中选出一个最为合适的先验。本文把这个先验选取问题作为贝叶斯判别问题，将训练样本的取值情况加以考虑，提出后验信念的概念，并用贝叶斯公式给出其计算方法。根据后验信念大小提出了选取先验的极大后验信念法(Highest Posterior Belief，HPB)。本文给出一个类似于 Neyman-Pearson 引理的公式以用于实际判别，并给出消除先验信念影响的贝叶斯因子，证明了选择先验的 II 型极大似然法就是当先验信念相等时的极大后验信念法。最后给出一个应用例子。

1. 先验分布的后验信念

用 $X^{(n)}$ 来表示训练数据，其含有 n 个 i. i. d. (独立同分布)的观察数据 $(X_1,\cdots,X_n)$，$P(\cdot)$ 表示在参数 θ 的控制下 $X^{(n)}$ 的概率。$\mu_i(i=0,\cdots,m)$ 表示 θ 的可能的先验分布。用 $\pi(\mu_i)(i=0,\cdots,m)$ 来表示我们对每个先验 $\mu_i(i=0,\cdots,m)$ 的先验信念(我们将其看成是先验分布的函数，称之为信念)。那么，参数 θ 的先验分布可表示为：

$$\mu(\theta)=\sum\pi(\mu_i)\mu_i(\theta) \tag{1}$$

用 $\pi_n(\mu_i)(i=0,\cdots,m)$ 来表示对每个先验的修正后的信念(即关于每个先验分布的后验概率)。将抽样看成是如下的过程：首先，根据先验信念选择一个 θ 的先验分布；其次，从其先验分布中选择 θ 的一个值；最后根据 $(\theta,X^{(n)})$ 的联合分布抽取样本。根据贝叶斯公式，有：

$$\begin{aligned}\pi(\mu_i|X^{(n)}) &= \frac{\int p(X^{(n)}|\mu_i,\theta)P(\mu_i,\theta)d\theta}{p(X^{(n)})}\\ &= \frac{\int p(X^{(n)}|\mu_i,\theta)\pi(\mu_i)\mu_i(\theta)d\theta}{p(X^{(n)})}\\ &= \frac{\pi(\mu_i)\int p(X^{(n)}|\mu_i,\theta)\mu_i(\theta)d\theta}{p(X^{(n)})}\end{aligned} \tag{2}$$

其中 $p(X^{(n)})=\sum\int p(X^{(n)}|\mu_i,\theta)P(\mu_i,\theta)d\theta$

记号 $\pi(\mu_i|X^{(n)})$ 表示关于一个先验的后验信念。(2)式给出了先验信念、后验信念和样本信息之间的关系，它反映了我们在获得了样本数据后对某个先验的信念，或者说是样本数据对我们的先验信念的支持程度。

2. 基于最大后验信念的先验分布的选取

从上节的讨论易知，一个关于参数 θ 的最合适的先验分布可以通过使后验信念最大化的方法来获得。这样一个最合适的先验分布称之为最大后验信念(Highest Posterior Belief，HPB)先验。

定义1 使(2)式的值最大的先验分布称为最大后验信念先验。

为了使问题更清楚地表述，下面介绍后验信念比的概念。

定义2 记 $\alpha_0=\pi_0\int p(X^{(n)}|\mu_0,\theta)\mu_0(\theta)d\theta$，$\alpha_1=\pi_1\int p(X^{(n)}|\mu_1,\theta)\mu_1(\theta)d\theta$，比值 α_0/α_1 称为 $\mu_0(\theta)$ 对 $\mu_1(\theta)$ 后验信念比，π_0/π_1 称为先验信念比。由(2)式，有：

$$\frac{\alpha_0}{\alpha_1}=\frac{\pi_0\int p(X^{(n)}|\mu_0,\theta)\mu_0(\theta)d\theta}{\pi_1\int p(X^{(n)}|\mu_1,\theta)\mu_1(\theta)d\theta} \tag{3}$$

假设仅有两个可以考虑的先验 $\mu_1(\theta)$ 和 $\mu_0(\theta)$，根据最大后验

*)智能技术与系统国家重点实验室开放课题(99002)资助。胡振宇 硕士，主要研究数据采掘、信息安全。林士敏 教授，主要研究机器学习、知识发现。陆玉昌 教授，主要研究智能系统、数据采掘。

先验法，如果要选用 $\mu_1(\theta)$ 而不是 $\mu_0(\theta)$ 就必须有 $\alpha_0/\alpha_1 < 1$，或

$$\frac{\int p(X^{(n)}|\mu_0,\theta)\mu_0(\theta)d\theta}{\int p(X^{(n)}|\mu_1,\theta)\mu_1(\theta)d\theta} < \frac{\pi_1}{\pi_0} \tag{4}$$

此即为 Neyman-Pearson 引理的另一种形式。

定义3 对于训练数据 X，将下面的表达式

$$l(\mu) = \int f(x|\mu,\theta)\mu(\theta)d\theta \tag{5}$$

看作是关于先验 μ 的函数，称之为先验似然(函数)。

将 $l(\mu) = \int f(x|\mu,\theta)\mu(\theta)d\theta$ 称为先验似然(函数)的直觉理由是：如果 $\mu(\theta)$ 能使 $l(\mu) = \int f(x|\mu,\theta)\mu(\theta)d\theta$ 较大，它是真正的先验的可能性就较大，因为对于较大的 $l(\mu) = \int f(x|\mu,\theta)\mu(\theta)d\theta$，数据 X 发生的可能性更高。式(4)表明：如果要拒绝 $\mu_0(\theta)$，先验似然比必须超过一定的界限。

定义4 因子

$$B = \frac{posterior\_belief\_ratio}{prior\_belief\_ratio} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} \tag{6}$$

称为先验贝叶斯因子。

先验信念比和后验信念比这两种比率相除，可能会减弱先验信念的影响，突出数据的影响。先验贝叶斯因子既依赖于数据又依赖于对先验的先验信念。从这个角度看，贝叶斯因子 B 是数据 X 对先验分布 $u_0(\theta)$ 的支持程度。因为由(3)式可得

$$B = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} = \frac{\int p(X^{(n)}|\mu_0,\theta)\mu_0(\theta)d\theta}{\int p(X^{(n)}|\mu_1,\theta)\mu_1(\theta)d\theta} \tag{7}$$

它恰好是先验似然 $\mu_0(\theta)$ 和 $\mu_1(\theta)$ 的比率，它部分地消除了 $\mu_0(\theta)$ 和 $\mu_1(\theta)$ 的先验信念的影响。

显然，如果对于各个先验分布的先验信念都相等，那么式(3)即蜕变为式(7)。由此得到一个确定先验分布的极大先验似然方法：

定理1 若对于各个先验分布的先验信念都相等，则使得先验似然达到最大值的 $\mu(\theta)$ 即为最大后验信念先验。

用上述方法获得的先验分布也称为 ML-II 先验(二型极大似然先验)，它类似于参数估计的 MLE 的方法。在给定先验分布的形式时，可用此方法方便地求得合适的先验。

例1 设 $X \sim N(\theta,1)$，主观地假设 θ 的先验的均值为 0，先验四分位点在 ±1 正态分布。则 $C(0,1)$(用 μ_0 表示)或 $N(0,2.19)$(用 μ_1 表示)被认为是符合条件的先验。假若仅考虑 μ_1 和 μ_0，哪一个更合适呢?注意到 $l(\mu_1)$ 为 $N(0,3.19)$，而

$$l(\mu_0) = \int (2\pi)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x-\theta)^2\} \cdot \frac{1}{\pi(1+\theta^2)}d\theta$$

表1给出在 x 取不同的值时 $l(\mu_0)$ 和 $l(\mu_1)$ 的值。当 x 较小时，μ_1 和 μ_0 的差别不大；当 $x=4.5$ 时，μ_0 的可能性两倍于 μ_1；当 $x=6.0$ 时，μ_0 的可能性十倍于 μ_1；当 $x=10$ 时，毫无疑问 μ_1 是错误的。表明当 $x \geqslant 6$ 时，数据对 μ_0 的支持远大于对 μ_1 的支持，由此我们应该采用 μ_0 作为最合适的先验分布。

表1 μ 先验似然值

x	0	4.5	6.0	10
$l(\mu_1)$	0.22	0.0093	0.00079	3.5×10^{-8}
$l(\mu_0)$	0.21	0.018	0.0094	0.0032

表2给出当分别给 μ_1 和 μ_0 赋予0.3及0.7的先验信念时，它们之间的后验信念。

表2 μ 后验信念值(1)

x	0	4.5	6.0	10
$\pi(\mu_1\|X)$	0.310	0.181	0.0348	4.687×10^{-6}
$\pi(\mu_0\|X)$	0.690	0.819	0.9650	0.9999953

表3给出当分别给 μ_1 和 μ_0 赋予0.7及0.3的先验信念时，它们之间的后验信念

表3 μ 后验信念值(2)

x	0	4.5	6.0	10
$\pi(\mu_1\|X)$	0.7097	0.5466	0.1639	2.55×10^{-5}
$\pi(\mu_0\|X)$	0.2903	0.4534	0.8361	0.94745

由表3可知，即使赋予 μ_1 以较大的先验信念，当样本取值较大时，仍然对它不利。这一点从图1也能看出来。

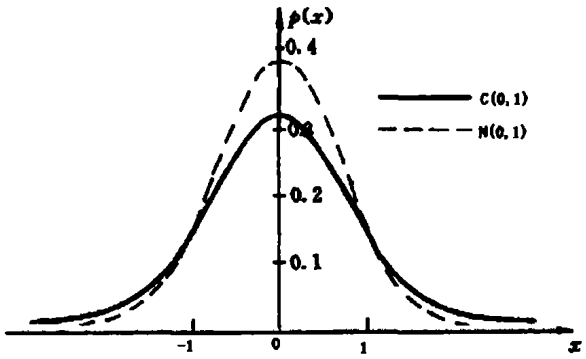

图1 C(0,1)与 N(0,1)的密度曲线

结束语 先验分布的研究对于贝叶斯学习具有重大的理论意义和实践意义。许多学者提出了选择先验的不同的方法或策略。HPB 先验不仅反映了先验信念对一个先验分布的支持，也反映了样本数据(或试验数据)对该先验分布的支持，既避免了单纯的主观偏向，又在一定程度上排除了样本数据带来的偶然性影响，为合理地选择先验分布提供了一种方便、简单的方法。

参考文献

1 林士敏，王双成，陆玉昌. 贝叶斯方法的学习机制与问题求解. 清华大学学报，2000，40(9)：61～64

2 Diaconis P，Freedman D. On The Consistency of Bayes Estimates. Ann. Statist，1986，14：1～26

3 Diaconis P，Freedman D. On The Consistency of Bayes Estimates of location. Ann. Statist，1986，14：69～87

4 Ghosal S，Ghosh J K，Samanta T. On Convergence of Posterior Distributions. Ann. Statist，1995，23：2145～2152

5 Ghosh J K，Ghosal S，Samanta T. Stability and Convergence of Posterior in Non-Regular Problems. Statistical Decision Theory and Related Topics，1994，5：183～199