文章编号: 1001-0920(2008)05-0589-04

二值 probit 回归模型的坍缩变分贝叶斯推断算法

卿湘运, 王行愚, 牛玉刚
(华东理工大学 信息科学与工程学院, 上海 200237)

摘 要: 给出了二值 probit 回归模型的坍缩变分贝叶斯推断算法. 此算法比变分贝叶斯推断算法能更逼近对数边缘似然, 得到更精确的模型参数后验期望值. 如果两个算法得到的分类错误一致, 则该算法的迭代次数较变分法明显减少. 仿真实验结果验证了所提出算法的有效性.

关键词: 二值 probit 模型; 变分贝叶斯推断; 坍缩变分贝叶斯推断
中图分类号: TP393 **文献标识码**: A

Collapsed variational Bayesian inference algorithm for binary probit regression model

QING Xiang-yun, WANG Xing-yu, NIU Yu-gang
(College of Information Science and Technology, East China University of Science and Technology, Shanghai 200237, China. Correspondent: WANG Xing-yu, E-mail: xywang@ecust.edu.cn)

Abstract: A collapsed variational Bayesian inference algorithm for binary probit regression model is proposed. The algorithm is a better approximation of the alogrithm of the marginal likelihood than standard variational Bayesian inference algorithm. It also can achieve more accurate posterior expectation of model parameters than the latter. If the classification error rates of two algorithms are close, the iteration number of the proposed algorithm is siginificantly decreased. Simulation experiments show the effectiveness of the algorithm.

Key words: Binary probit regression model; Variational Bayesian inference; Collapsed variational Bayesian inference

1 引 言

二值 probit 回归模型作为一个标准的回归与分类算法, 在统计决策、模式识别与数据挖掘中有着广泛的应用. 此模型单就分类精度而言, 一般比支持向量机等先进算法要差, 但实现简单, 容易理解, 且对分类结果的不确定性可用概率度量[1]. 在对此非线性回归模型进行参数估计时, 采用最大似然估计可能导致模型和概率估计产生偏差, 故常应用贝叶斯方法, 对模型参数赋予适当的先验分布. 利用马尔可夫链蒙特卡罗(MCMC)推断此贝叶斯模型参数, 虽可能得到参数的逼近无偏估计[2], 但面临的主要问题是: 1) 需进行大量抽样, 特别在高维数据情况下进行多变量高斯分布抽样的计算量较大; 2) 难以决定马尔可夫链何时收敛. 因此利用变分贝叶斯推断算法(VB)可加快参数的逼近推断速度[3]. 但变分贝叶斯算法可能收敛到局部最优. 文献[4]比较了应用上述两种方法对二值 probit 回归模型求解参数后验分布的差异. 在某些情况下, 变分法得到的参数分布与真正的参数分布相差很大, 而 MCMC 则能得到较好的逼近结果. 所以在进行实际的贝叶斯推断时, 应对变分法作深入地研究与测试, 才能将其作为可靠的计算工具.

本文利用坍缩变分贝叶斯算法(CVB)推断模型参数, 尽量消除由于模型隐变量相关而导致的逼近误差, 取得更精确的推断结果. 在几个模拟和真实的数据集上的实验结果显示了该方法的有效性.

2 二值 probit 回归模型

已知 N 个 p 维样本数据 $x_i \in R^p$ (在本文中为行矢量) 及二值随机变量 y_i, $1 \leq i \leq N$. 引进 N 个辅助变量 z_i, 有

收稿日期: 2007-03-26; **修回日期**: 2007-09-24.
基金项目: 国家 973 计划项目(2002CB312203); 国家自然科学基金项目(60674089); 上海市重点学科研究项目(B504).
作者简介: 卿湘运(1977—), 男, 湖南双峰人, 博士生, 从事智能控制与理论、机器学习的研究; 王行愚(1944—), 男, 上海人, 教授, 博士生导师, 从事智能控制与理论、模式识别的研究.

$$y_i = \begin{cases} 1, & z_i > 0; \\ 0, & \text{其他}. \end{cases} \tag{1}$$

模型的层次先验分布为

$$z_i = x_i\theta + \varepsilon_i,\ \varepsilon_i \sim N(0,1),\ \theta \sim N(0,V),$$

其中 θ 为 p 维回归系数列矢量,服从均值为 0,协方差矩阵为 V 的多变量高斯分布. 显然,在给定 θ 的情况下,y_i 为独立贝努利随机变量,且

$$p(y_i = 1 \mid z_i) = \Phi(z_i), \tag{2}$$

其中 Φ 表示均值为 0,方差为 1 的标准正态分布累积分布函数. 对于此模型,容易得到 Gibbs 抽样方法. θ 的全条件分布为

$$\theta \mid z \sim N(M,\Lambda), \tag{3}$$

$$M = \Lambda X^{\mathrm{T}} z,\ \Lambda = (V^{-1} + X^{\mathrm{T}}X)^{-1},$$

其中 X 为 N 个样本组成的 $N \times p$ 维观测数据矩阵.

每个辅助变量 z_i 的全条件分布为

$$z_i \mid \theta, x_i, y_i \propto \begin{cases} N(x_i\theta,1)\, I(z_i > 0), & y_i = 1; \\ N(x_i\theta,1)\, I(z_i \le 0), & \text{其他}. \end{cases} \tag{4}$$

其中 $I()$ 为指示函数,括号里的条件满足则为 1,否则为 0. 实际上 z_i 服从截断标准正态分布,当 y_i 为 1 时,z_i 在标准正态分布 0 处左截断;当 y_i 为 0 时,z_i 在标准正态分布 0 处右截断.

3 变分贝叶斯推断算法

设在贝叶斯层次模型中所有超参数集为 ζ,所有中间隐变量集为 η,观测变量为 X,变分推断方法就是力图找到一个具有因子分解形式的 $Q(\eta)$ 逼近后验分布 $p(\eta \mid X,\zeta)$. 因此,最小化 $Q(\eta)$ 和 $p(\eta \mid X,\zeta)$ 的 Kullback-Leibler 散度目标函数为

$$D(Q(\eta) \,||\, p(\eta \mid X,\zeta) = E_Q[\ln Q] - E_Q[\ln p(\eta \mid X,\zeta)] + \ln p(X \mid \zeta).$$

应用 Jensen 不等式,最大化对数边缘似然的下界 $L(Q,\zeta)$,得

$$\ln p(X \mid \zeta) \ge L[Q,\zeta] = E_Q[\ln p(\eta \mid X,\zeta)] - E_Q[\ln Q],$$

进而可得到后验 $p(\eta \mid X,\zeta)$ 的逼近分布 $Q(\eta)$.

对于二值 probit 回归模型,设 $D := \{x_1, \cdots, x_N, y_1, \cdots, y_N, V\}$,其目标函数为

$$L(D) = \iint Q(z,\theta) \times \ln \frac{p(\theta \mid 0,V)\left\{\prod_{i=1}^{N} p(z_i \mid x_i\theta,1)\right\}\prod_{i=1}^{N} p(y_i \mid z_i)}{Q(z,\theta)} \mathrm{d}z\mathrm{d}\theta.$$

假定中间变量 θ 与 z_i 独立,得

$$Q(z,\theta) = \left(\prod_{i=1}^{N} Q(z_i)\right) Q(\theta) = \left(\prod_{i=1}^{N} Q(z_i \mid \tilde{u}_i,\tilde{\sigma}_i^2)\right) Q(\theta \mid \tilde{M},\tilde{\Lambda}).$$

则目标函数为

$$L(D) = E_Q[\ln p(\theta \mid 0,V)] + \sum_{i=1}^{N} E_Q[\ln p(z_i \mid x_i\theta,1)] + \sum_{i=1}^{N} E_Q[\ln p(y_i \mid z_i)] - E_Q[\ln Q]. \tag{5}$$

最大化此目标函数,可分别对 θ 和 z_i 求导,得

$$Q(z_i \mid \tilde{\mu}_i,\tilde{\sigma}_i^2) \propto \begin{cases} N(\tilde{\mu}_i,\tilde{\sigma}_i^2)\, I(z_i > 0), & y_i = 1; \\ N(\tilde{\mu},\tilde{\sigma}_i^2)\, I(z_i < 0), & \text{其他}; \end{cases}$$

$$\tilde{\mu}_i = x_i\langle\theta\rangle,\ \tilde{\sigma}_i^2 = 1.$$

其中 $\langle\theta\rangle$ 为 θ 的变分逼近期望值 $E_Q(\theta)$. 而 z_i 的变分逼近期望值则为

$$\langle z_i\rangle = \begin{cases} \tilde{\mu}_i + \tilde{\sigma}_i\Psi(-\tilde{\mu}_i/\tilde{\sigma}_i)/\Phi(\tilde{\mu}_i/\tilde{\sigma}_i), & y_i = 1; \\ \tilde{\mu}_i - \tilde{\sigma}_i\Psi(-\tilde{\mu}_i/\tilde{\sigma}_i)/\Phi(-\tilde{\mu}_i/\tilde{\sigma}_i), & \text{其他}. \end{cases} \tag{6}$$

其中 $\Psi()$ 表示均值为 0,方差为 1 的标准正态分布概率密度函数.

因为

$$Q(\theta \mid \tilde{M},\tilde{\Lambda}) \propto N(\tilde{M},\tilde{\Lambda}),$$

$$\tilde{M} = \tilde{\Lambda}X^{\mathrm{T}}\langle z\rangle,\ \tilde{\Lambda} = (V^{-1} + X^{\mathrm{T}}X)^{-1},$$

故 θ 的变分逼近期望值为

$$\langle\theta\rangle = \tilde{\Lambda}X^{\mathrm{T}}\langle z\rangle. \tag{7}$$

此模型各中间变量的先验与后验分布构成共轭指数对,因此也可根据 MCMC 抽样式(3) 和(4) 直接得到式(6) 和(7). 迭代执行式(6) 和(7),直至目标函数(5) 不再增加或迭代停止条件满足.

文献[4] 发现,应用上述变分贝叶斯算法求得此模型的后验参数非常不精确,且在迭代过程中参数取值变化区间小,限制了变分贝叶斯算法得到更精确的参数估计值. 其原因是:1) 此算法实际上有"迫零"特性,强调逼近模型分布的拖尾部分,因而趋向低估参数的方差[5];2) θ 与 z_i 在二值 probit 回归模型中是高度相互依赖的,而 VB 方法忽略了这些依赖关系,致使目标函数不能非常紧地逼近对数边缘似然. 原因 1) 是变分法性质所决定的,无法克服;但对于原因 2),则可得到更紧的逼近.

4 坍缩变分贝叶斯推断算法

下面对二值 probit 回归模型推导出一种新的 VB 算法,以更紧地逼近对数边缘似然目标函数. 假定各 z_i 相互独立,但 z_i 与 θ 则相互依赖. 首先对 θ 积分,即对 z_i 求边缘概率密度函数,再对 z_i 求变分推断参数. 文献[6] 称这种将部分隐变量首先积分再进行变分推断的方法为坍缩变分贝叶斯(CVB) 推

断, 并应用于文本主题聚类的 LDA(Latent Dirichlet Allocation) 模型,取得了较变分贝叶斯推断更理想的结果.

z 的边缘分布为

$$p(z \mid V, x) = \int_{\theta} p(z \mid \theta, X)\, p(\theta \mid V)\, \mathrm{d}\theta =$$

$$\int_{\theta} (2\pi)^{-N/2} \times \exp[-(z - X\theta)^{\mathrm{T}}(z - X\theta)/2] \times (2\pi)^{-D/2} \times |V|^{-1/2} \times \exp[-(\theta^{\mathrm{T}} V^{-1}\theta)/2]\, \mathrm{d}\theta =$$

$$(2\pi)^{-(N+D)/2} \times |V|^{-1/2} \times \int_{\theta} \exp\{-[z^{\mathrm{T}} z + \theta^{\mathrm{T}}(X^{\mathrm{T}} X + V^{-1})\theta - z^{\mathrm{T}} X\theta - \theta^{\mathrm{T}} X^{\mathrm{T}} z]/2\}\, \mathrm{d}\theta.$$

若 $\Lambda = (X^{\mathrm{T}} X + V^{-1})^{-1}$,则上式中

$$p(z \mid V, X) = (2\pi)^{-(N+D)/2} \times |V|^{-1/2} \times \int_{\theta} \exp\{-[z^{\mathrm{T}} z - z^{\mathrm{T}} X\Lambda X^{\mathrm{T}} z + (\theta - \Lambda X^{\mathrm{T}} z)^{\mathrm{T}} \Lambda^{-1} (\theta - \Lambda X^{\mathrm{T}} z)]/2\}\, \mathrm{d}\theta =$$

$$(2\pi)^{-N/2} \times |I_p + V X^{\mathrm{T}} X|^{-1/2} \times \exp\{-[z^{\mathrm{T}}(I_N - X\Lambda X^{\mathrm{T}}) z]/2\},$$

其中 I_p 表示 $p \times p$ 维单位矩阵. 若 $H = I_N - X\Lambda X^{\mathrm{T}}$, 则

$$p(z \mid V, X) = (2\pi)^{-N/2} \times |I_p + V X^{\mathrm{T}} X|^{-1/2} \times \exp[-(z^{\mathrm{T}} H z)/2].$$

因此目标函数为

$$L_{\mathrm{CVB}}(D) = \int_{z} Q(z) \ln \frac{p(z \mid V, X) \prod_{i=1}^{N} p(y_i \mid z_i)}{Q(z)}\, \mathrm{d}z. \quad (8)$$

因为 CVB 较 VB 对变分后验的假设条件更弱, 故

$$L_{\mathrm{VB}}(D) \leq L_{\mathrm{CVB}}(D),$$

CVB 比 VB 能得到更好的逼近下界. 对每个 z_i 最大化目标函数,即对 z_i 求导,由此可得

$$Q(z_i \mid \mu_i, \sigma_i^2) \propto \begin{cases} N(\mu_i, \sigma_i^2)\, I(z_i > 0), & y_i = 1; \\ N(\mu_i, \sigma_i^2)\, I(z_i < 0), & \text{其他}; \end{cases}$$

$$\mu_i = -[H(i,:)\langle z\rangle - H(i,i)\langle z_i\rangle]/H(i,i);$$

$$\sigma_i = (1/H(i,i))^{1/2}.$$

其中 $H(i,:)$ 表示 H 的第 i 行. 故 z_i 的变分逼近期望值为

$$\langle z_i\rangle = \begin{cases} \mu_i + \sigma_i \Psi(-\mu_i/\sigma_i)/\Phi(\mu_i/\sigma_i), & y_i = 1; \\ \mu_i - \sigma_i \Psi(-\mu_i/\sigma_i)/\Phi(-\mu_i/\sigma_i), & \text{其他}. \end{cases} \quad (9)$$

θ 的变分逼近期望值为

$$\langle\theta\rangle = \Lambda X^{\mathrm{T}} \langle z\rangle. \quad (10)$$

H 和 ΛX^{T} 可在迭代之前计算好,对比 VB 方法, 其主要计算区别在于计算 $\tilde{\mu}_i$ 与 μ_i. CVB 每次迭代实际上可不计算 $\langle\theta\rangle$,因此省略掉两者相似的计算部分,则 VB 的计算量为 $(2Np + p^2)T$,CVB 的计算量为 $N(N-1)T$,其中 T 为两算法的迭代次数. 一般来说,当样本数远大于维数时,CVB 的计算量要比 VB 方法大,因为 CVB 的每个 z_i 类似 Gibbs 抽样方法由其余所有的 z_i 产生,只是 CVB 方法不需要抽样.

5　实验结果

5.1　仿真数据

本仿真研究采用文献[4] 的仿真数据生成模型,说明应用 CVB 算法能得到较精确的参数逼近值,而 VB 方法则不能. 考虑一个最简单的二值 probit 回归模型, $p = 2$, $N = 100$, $x_i = (1, q_i)$, $\theta = (\alpha, \beta)$. q_i 服从均值为 0,方差为 1 的标准正态分布,随机产生 100 个数, $\alpha = 1$, $\beta = 5$. 故 $p(y_i = 1 \mid q_i) = \Phi(1 + 5q_i)$,如图 1 所示.

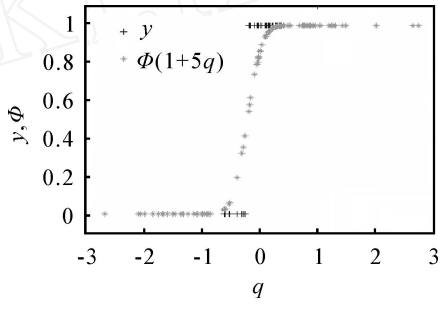

图 1　仿真数据 y_i 与 $\Phi(1 + 5q_i)$

设定先验协方差矩阵 $V = 5I_p$. 根据 z_i 的先验分布随机产生迭代初始值,并根据 z_i 计算 θ 的迭代初始值. 设在第 t 次迭代 $\mathrm{Dist}(t) = \|\theta_t - \theta_{t-1}\|_2$,其迭代停止标准为[7]: 迭代次数大于 2 000 或 $|(\mathrm{Dist}(t) - \mathrm{Dist}(t-1))/\mathrm{Dist}(t-1)| < 0.001$. 结果如图 2 所示.

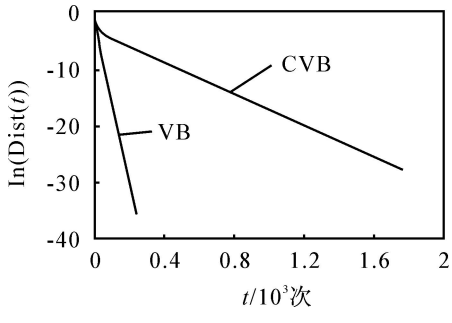

图 2　仿真数据 CVB 与 VB 的收敛比较

对于 VB, 参数推断结果 $\langle\alpha\rangle = 0.56$, $\langle\beta\rangle = 1.78$, 共迭代 237 次;对于 CVB, 参数推断结果 $\langle\alpha\rangle = 1.31$, $\langle\beta\rangle = 5.20$,迭代 1 779 次. 根据不同的随机初始值重复实验 20 次,参数后验推断结果一致,仅收敛次数稍微不同. 由此结果可知,VB 算法虽然很

快收敛,但陷入局部最优,其参数逼近结果与真实参数值有较大差距,从而也验证了文献[4]的结论. CVB 算法迭代次数尽管较 VB 算法多,但其参数逼近结果与真实参数值很接近,说明 CVB 算法的逼近效果更理想.

5.2 真实数据集

对于真实数据集,虽可首先应用 MCMC 方法得到参数比较精确的后验逼近,再分别检验 VB 和 CVB 算法的逼近精度,但 MCMC 的收敛难以判断,所以本文用分类错误率指标来衡量两算法的性能差异. 本实验从 UCI 机器学习库中选取 4 个数据集,均为二分类问题,对各特征不作预处理及统计检验. 设定先验协方差矩阵 $V = 100 I_p$,其迭代停止标准与仿真数据采取的标准一致. 对每个数据集,随机选取 70 %的数据作为训练数据,剩余 30 %的数据作为分类测试数据,如此随机抽取 20 次,分别应用 VB 与 CVB 推断算法得到的平均分类错误率、迭代次数结果及各数据集属性如表 1 所示.

表 1 4 个数据集及分类结果

指标	数据集			
	ion	sonar	liver	pima
观测数据个数 N	351	208	345	768
维数 p	33	60	6	8
VB 平均分类错误率 / %	25.9 ±4.6	36.8 ±7.7	33.3 ±4.9	31.4 ±2.4
VB 平均迭代次数	377	374	686	700
CVB 平均分类错误率 / %	17.4 ±4.0	26.7 ±4.0	33.5 ±5.0	31.8 ±2.3
CVB 平均迭代次数	932	648	62	42
SVM 平均分类错误率 / %	14.2 ±2.6	22.2 ±4.2	33.3 ±4.6	23.0 ±1.7

从表 1 可知,对于 ion 和 sonar 两个数据集,虽然 CVB 比 VB 的迭代次数多,但分类错误率要低得多,意味着 CVB 比 VB 能得到更准确的模型参数后验推断值,与仿真数据的结论较吻合. 而对于 liver 和 pima 两个数据集,CVB 与 VB 算法得到的分类错误率接近,但 CVB 的迭代次数却只有 VB 算法的 10 %左右,说明若两算法得到同样的参数后验推断值,则 CVB 比 VB 算法的迭代次数明显减少. 从表 1 也可以看出,probit 回归模型的分类精度比支持向量机还要差一些.

6 结 论

本文给出了二值 probit 回归模型的坍缩变分贝叶斯参数推断算法. 在每次迭代增加较少计算量的条件下,坍缩变分贝叶斯要么比变分贝叶斯得到更精确的参数后验推断结果,要么在都得到较一致的参数后验推断结果时,坍缩变分贝叶斯比变分贝叶斯的迭代次数明显减少. 该算法也可推广到 logistic 回归模型及概率支持向量机模型的逼近推断中[8].

参考文献(References)

[1] Figueiredo M A T. Adaptive sparseness for supervised learning[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2003, 25(9): 1150-1159.

[2] Albert J, Chibs S. Bayesian analysis of binary and polychotomous response data[J]. J of the American Statistical Association, 1993, 88(442): 669-679.

[3] Beal M. Variational algorithms for approximate Bayesian inference[D]. London: London University, 2003.

[4] Consonni G, Marin J M. Mean-field variational approximate Bayesian inference for latent variable models[J]. Computational Statistics and Data Analysis, 2007, 52(2): 790-798.

[5] Minka T. Divergence measures and message passing [R]. Cambridge: Microsoft Research Limited, 2005.

[6] Yee W T, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent dirichlet allocation[C]. Advances in Neural Information Processing Systems. Cambridg: MIT Press, 2006: 1353-1360.

[7] Qi Y, Jaakkola T S. Parameter expanded variational Bayesian methods[C]. Advances in Neural Information Processing Systems. Cambridg: MIT Press, 2006: 1097-1104.

[8] Holmes C C, Held L. Bayesian auxiliary variable models for binary and multinomial regression [J]. Bayesian Analysis, 2006, 1(1): 145-168.