科技部補助專題研究計畫成果報告
期末報告

從即時貝氏學習方法到群組學習

計 畫 類 別 ： 個別型計畫
計 畫 編 號 ： MOST 104-2118-M-004-005-
執 行 期 間 ： 104年08月01日至105年07月31日
執 行 單 位 ： 國立政治大學統計學系

計 畫 主 持 人： 翁久幸

計畫參與人員： 碩士班研究生-兼任助理人員：張文嘉
碩士班研究生-兼任助理人員：蔡儀君
碩士班研究生-兼任助理人員：黃國展
碩士班研究生-兼任助理人員：簡于閔

報 告 附 件 ： 出席國際學術會議心得報告

中 華 民 國 105 年 09 月 11 日

中 文 摘 要 ： 線上(即時)機器學習係指內那些依順序逐個逐個處理資料觀測值的機器學習演算方法，在這種學習方法中，各個觀測值在被處理過後即可刪去，不需保留，因此,線上演算法需要的記憶空間較少。加上這類演算法一般較為簡單，容易執行，對於處理大量且即時的資料具有相當優勢。近年來由於網際網路發達產生許多大量且即時的資料，因而使即時演算法益加受到關注。本研究討論網路產品評比資料之即時統計分析方法,應用於實際資料之情況良好。
研究成果可以有實際之應用。

中 文 關 鍵 詞 ： 貝氏分析, 線上(即時)機器學習

英 文 摘 要 ： Online learning refers to learning methods that process data one-by-one. Since the data point can be removed after being processed, online methods require less memory and are advantageous when dealing with very large real-time data. This project studies online statistical inference for Internet product ratings data. The proposed method is applied to two real datasets. The results are satisfactory.

英 文 關 鍵 詞 ： Bayesian inference, online machine learning

Final report for project: From online to batch learning of a Bayesian method

1 Introduction

The present project explored Woodroofe-Stein's identity for Bayesian learning. We extend the moment matching in Weng and Lin (2011) for models of ranked data to item response theory models. Both are latent ability models, and they often model the outcomes by normal or logistic distributions. We obtain online algorithms to adjust the parameters in certain ordinal IRT models designed for Likert-type data, and demonstrate the effectiveness of the proposed algorithm through two real datasets. Our experiments show that the proposed method works well. This is joint work with Dr. Coad.

2 Item response models

Item response models have been widely used in modeling scored data from educational tests; see [1].

Example 1: Basic IRT. The basic one-parameter IRT model is for analyzing dichotomously scored test data. It models the probability of a correct response to a test item as a function of the examinee's ability and the item's difficulty. Let the item response variable Y_{ij} be 0 or 1, corresponding to whether the response to the jth test item taken by the ith individual is correct or not, θ_i represent the ability of the ith individual, and β_j represent the difficulty of the jth test item. The model has the form

$$P(Y_{ij} = 1|\theta_i, \beta_j) = F(\theta_i - \beta_j), \tag{1}$$

where $F(\cdot)$ is a c.d.f. from a continuous distribution. When $F(\cdot)$ is the standard logistic c.d.f., (1) is the Rasch model [5]; and when $F(\cdot)$ is the standard normal c.d.f., (1) is called the Normal Ogive (or Probit) model.

Example 2: Ordinal IRT. [6] introduced the graded response model to analyze ordered polytomous data. Let Y_{ij} denote the score of the ith individual on item j, θ_i the proficiency of the ith individual, β_j the discrimination parameter for test item j, $\delta_{j,c}$ the item response parameter for test item j and category c, where $c = \{0, 1, ..., C_j\}$. The model specifies the probability of the ith individual responding in category c or higher on the test item j as

$$P(Y_{ij} \geq c|\beta_j, \theta_i, \delta_{j,c}) = F(\beta_j(\theta_i - \delta_{j,c})),$$

where $F(\cdot)$ is the c.d.f. of a logistic or a normal distribution. In this model, the number of categories C_j for item j can be different across j. For Likert-type data in which the categories are $\{1, 2, ..., C\}$ for all items, [4] proposed a modified graded response model, which resolved the item response parameter $\delta_{j,c}$ into the item location parameter α_j and the category threshold parameter d_{c-1}, where $d_0 = -\infty$; that is,

$$P(Y_{ij} \geq c|\beta_j, \theta_i, \alpha_j, d_{c-1}) = F(\beta_j(\theta_i + \alpha_j - d_{c-1})). \tag{2}$$

The resulting probability of category c is

$$P(Y_{ij} = c|\beta_j, \theta_i, a_j, d_c, d_{c+1}) = F(\beta_j(\theta_i + a_j - d_c)) - F(\beta_j(\theta_i + a_j - d_{c+1}))$$

[2] proposed an ordinal IRT model to fit online product ratings data. Let $Y_{ij} \in \{1, 2, ..., C\}$ denote the rating of item i by rater j. Typically a rater may only rate a small proportion of products. So, many Y_{ij} are missing. They assume that the observed Y_{ij} is determined by an unobserved variable Y^*_{ij}:

$$Y_{ij} = c \Leftrightarrow Y^*_{ij} \in (\gamma_{c-1}, \gamma_c], \tag{3}$$

where the γ_c are cutpoints, $\gamma_0 = -\infty$, $\gamma_C = \infty$, and assume that Y^*_{ij} is parameterized as

$$Y^*_{ij} = \alpha_j + \beta_j\theta_i + \epsilon_{ij},\ \epsilon_{ij} \overset{iid}{\sim} N(0, 1). \tag{4}$$

In (4), α_j captures the center location of rater j's rating, β_j measures the discriminating ability of rater j, and θ_i represents the latent quality of item i. By (3), the probability of observing Y_{ij} in category c or higher is

$$P(Y_{ij} \geq c|\alpha_j, \beta_j, \theta_i, \gamma_{c-1}) = \Phi(\beta_j\theta_i + \alpha_j - \gamma_{c-1}). \tag{5}$$

So, the probability of category c is

$$P(Y_{ij} = c|\alpha_j, \beta_j, \theta_i, \gamma_{c-1}, \gamma_c) = \Phi(\beta_j\theta_i + \alpha_j - \gamma_{c-1}) - \Phi(\beta_j\theta_i + \alpha_j - \gamma_c).$$

That ϵ_{ij} in (4) follows either the normal or the logistic distribution. We observe that model (5) closely resembles (2), but they differ in the parameterization of the item location and category threshold parameters.

3 The approximation method

Let ϕ and Φ denote the density and distribution function of a standard normal variable, and let $\phi(x|\mu, \sigma)$ denote the density of the normal distribution with mean μ and standard deviation σ.

The proposed method is based on the moment equations obtained from a version of Stein's identity. The famous Stein's lemma [7] concerns the expectation of a normally distributed random variable. It is of interest primarily because of its applications to the James-Stein estimator [3] and to empirical Bayes methods. In the context of setting sequential confidence levels, [10] studied integrable expansions for posterior distributions and developed a variant of Stein's identity. This identity concerns the expectation with respect to a "nearly normal distribution" in the sense: $p(\boldsymbol{z}) = \phi(\boldsymbol{z})f(\boldsymbol{z})$, where f is a real-valued almost differentiable function defined on R^p. [9] used this identity to obtain a Bayesian approximate moment-matching method, and referred to it as the Woodroofe-Stein identity to distinguish it from Stein's lemma.

Here we describe only some moments equations from this identity. For a detailed account of the identity, we refer readers to [8, Section 2.2] and [9, Corollary 2]. Let $\boldsymbol{\psi}^* = (\psi_1^*, \ldots, \psi_p^*)'$ be a suitably normalized vector of the parameter $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)'$. Suppose that the posterior density of $\boldsymbol{\psi}^*$ can be written as

$$C\phi(\boldsymbol{\psi}^*)f(\boldsymbol{\psi}^*), \tag{6}$$

where C is the normalizing constant. Further suppose that f is a twice continuously differentiable function. An application of the Woodroofe-Stein identity gives the following:

$$E(\boldsymbol{\psi}^*) = E\left(\frac{\nabla f(\boldsymbol{\psi}^*)}{f(\boldsymbol{\psi}^*)}\right), \tag{7}$$

$$E(\psi_i^*\psi_q^*) = \delta_{iq} + E\left[\frac{\nabla^2 f(\boldsymbol{\psi}^*)}{f(\boldsymbol{\psi}^*)}\right]_{iq}, \quad i, q = 1, \ldots, k, \tag{8}$$

where $\delta_{iq} = 1$ if $i = q$ and 0 otherwise, and $[\cdot]_{iq}$ indicates the (i, q) entry of a matrix.

From (7) and (8), the mean and variance of ψ_i^* are

$$E(\psi_i^*) = E\left(\frac{\partial \log f(\boldsymbol{\psi}^*)}{\partial \psi_i^*}\right), \tag{9}$$

$$Var(\psi_i^*) = E((\psi_i^*)^2) - E(\psi_i^*)^2 = 1 + E\left[\frac{\partial^2 f/\partial(\psi_i^*)^2}{f}\right] - \left[E\left(\frac{\partial \log f}{\partial \psi_i^*}\right)\right]^2. \tag{10}$$

4 Online inference of IRT models for Likert-type data

For the basic one-parameter IRT models (1), if the item response function F is taken as the standard logistic or normal c.d.f., then it is not difficult to derive real-time parameter adjustment by modifying Algorithms 1 and 3 in [9].

Our primary interest here are online inference for models (2) and (5), designed for Likert-type data. We assume that each α_j follows $N(\mu_{\alpha_j}, \sigma^2_{\alpha_j})$, each β_j follows $N(\mu_{\beta_j}, \sigma^2_{\beta_j})$, and each θ_i follows $N(\mu_{\theta_i}, \sigma^2_{\theta_i})$ with all parameters mutually independent. Now define the normalized quantities

$$\alpha_j^* = \frac{\alpha_j - \mu_{\alpha_j}}{\sigma_{\alpha_j}}, \quad \beta_j^* = \frac{\beta_j - \mu_{\beta_j}}{\sigma_{\beta_j}}, \quad \theta_i^* = \frac{\theta_i - \mu_{\theta_i}}{\sigma_{\theta_i}}. \tag{11}$$

The posterior distribution of $(\alpha_j^*, \beta_j^*, \theta_i^*)$ given $y_{ij} = c$ is

$$p(\alpha_j^*, \beta_j^*, \theta_i^* | y_{ij}) \propto \phi(\alpha_j^*, \beta_j^*, \theta_i^*) f(\alpha_j^*, \beta_j^*, \theta_i^*), \tag{12}$$

where f is the likelihood based on data $y_{ij} = c$. Note that (12) is of the form (6). Therefore, one can apply (9), (10) and (11) to derive expressions of approximations.

Below we present the sequential update rule when f in (12) is from (5) and comments on how similar procedures can be applied when f is as (2).

For model (5), the posterior distribution of $(\alpha_j^*, \beta_j^*, \theta_i^*)$ given $y_{ij} = c$ is (12) with

$$f(\alpha_j^*, \beta_j^*, \theta_i^*) = \Phi(\beta_j \theta_i + \alpha_j - \gamma_{c-1}) - \Phi(\beta_j \theta_i + \alpha_j - \gamma_c). \tag{13}$$

The proposed estimates for the posterior means and variances are

$$\tilde{\mu}_{\alpha_j} = \mu_{\alpha_j} + \left(\frac{\sigma^2_{\alpha_j}}{\nu}\right) \cdot \Omega\left(\frac{\mu_x}{\nu}, \frac{a}{\nu}\right), \quad \tilde{\sigma}^2_{\alpha_j} = \sigma^2_{\alpha_j} \left\{1 - \left(\frac{\sigma_{\alpha_j}}{\nu}\right)^2 \cdot \Delta\left(\frac{\mu_x}{\nu}, \frac{a}{\nu}\right)\right\}, \tag{14}$$

$$\tilde{\mu}_{\beta_j} = \mu_{\beta_j} + \left(\frac{\sigma^2_{\beta_j}\mu_{\theta_i}}{\nu}\right) \cdot \Omega(\frac{\mu_x}{\nu}, \frac{a}{\nu}), \quad \tilde{\sigma}^2_{\beta_j} = \sigma^2_{\beta_j} \left\{1 - \left(\frac{\sigma_{\beta_j}\mu_{\theta_i}}{\nu}\right)^2 \cdot \Delta\left(\frac{\mu_x}{\nu}, \frac{a}{\nu}\right)\right\}, \tag{15}$$

$$\tilde{\mu}_{\theta_i} = \mu_{\theta_i} + \left(\frac{\sigma^2_{\theta_i}\mu_{\beta_j}}{\nu}\right) \cdot \Omega(\frac{\mu_x}{\nu}, \frac{a}{\nu}), \quad \tilde{\sigma}^2_{\theta_i} = \sigma^2_{\theta_i} \left\{1 - \left(\frac{\sigma^2_{\theta_i}\mu_{\beta_j}}{\nu}\right)^2 \cdot \Delta\left(\frac{\mu_x}{\nu}, \frac{a}{\nu}\right)\right\}, \tag{16}$$

where

$$\mu_x = \mu_{\beta_j}\mu_{\theta_i} + \mu_{\alpha_j} - \gamma_{c-1} \qquad \text{and} \qquad a = \gamma_c - \gamma_{c-1}, \tag{17}$$

$$\begin{aligned} \Omega(\mu_x, a) &= \frac{\phi(\mu_x) - \phi(\mu_x - a)}{\Phi(\mu_x) - \Phi(\mu_x - a)}, \\ \Delta(\mu_x, a) &= \frac{\mu_x\phi(\mu_x) - (\mu_x - a)\phi(\mu_x - a)}{\Phi(\mu_x) - \Phi(\mu_x - a)} + \left(\frac{\phi(\mu_x) - \phi(\mu_x - a)}{\Phi(\mu_x) - \Phi(\mu_x - a)}\right)^2, \end{aligned} \tag{18}$$

$$\nu = \sqrt{1 + \sigma^2_{\alpha_j} + \sigma^2_{\beta_j}\mu^2_{\theta_i} + \sigma^2_{\theta_i}\mu^2_{\beta_j}}. \tag{19}$$

For the unknown cutpoints $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{C-1})^T$, we propose to estimate them through the distribution of y_{ij}^* in (4) and the relation $\{y_{ij} = c\} \Leftrightarrow \{y_{ij}^* \in (\gamma_{c-1}, \gamma_c]\}$ in (3).

For the modified graded response model (2), the posterior distribution of $(\alpha_j^*, \beta_j^*, \theta_i^*)$ given $y_{ij} = c$ is (12) with

$$f(\alpha_j^*, \beta_j^*, \theta_i^*) = \Phi(\beta_j(\theta_i + \alpha_j - d_{c-1})) - \Phi(\beta_j(\theta_i + \alpha_j - d_c)). \tag{20}$$

The parameter update can be derived similarly. To see how, first note that an observed Y_{ij} from model (2) can be determined by an unobserved variable

$$Y_{ij}^\dagger = \beta_j(\alpha_j + \theta_i) + \epsilon_{ij},\ \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{21}$$

and the relation $Y_{ij} = c \Leftrightarrow Y_{ij}^\dagger \in (\beta_j d_{c-1}, \beta_j d_c]$; or, equivalently, $Y_{ij} = c \Leftrightarrow Y_{ij}^\dagger/\beta_j \in (d_{c-1}, d_c]$. Therefore, by approximating the distribution of $Y_{ij}^\dagger/\beta_j$, we can estimate the threshold parameters $(d_1, ..., d_{C-1})$ in the same manner as for $\boldsymbol{\gamma}$. The sequential update for the posterior means and variances of α_j, β_j, and θ_i can be derived analogously.

5 Concluding remarks

Online methods are necessary when large amount of data arrive in streams data and real-time parameter adjustment is needed. We have developed an algorithm for Bayesian online parameter estimation for IRT models with Likert-type data. We have also compared our real-time estimation method with the offline MCMC methods via the package Ratings. Though sacrificing some accuracy, in general, our proposed method achieves a good performance, but with considerably less computational time. Thus, for situations where faster approximate methods are desirable, our proposed method can be a useful alternative to offline methods. That said, we have to point out some limitations of our method. First, with only mean and variance updates, our method can not provide estimates of quantities of interest, which are easily obtainable by MCMC methods. Second, there is a lack of theoretical analysis on the discrepancies between the offline method and the proposed online one. We leave it as future work.

References

[1] R. J. De Ayala. *The Theory and Practice of Item Response Theory*. Guilford Publications, 2013.

[2] D. E. Ho and K. M. Quinn. Improving the presentation and interpretation of online ratings data with model-based figures. *The American Statistician*, 62(4):279–288, 2008.

[3] W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–379, 1961.

[4] E. Muraki. Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, 14(1):59–71, 1990.

[5] G. Rasch. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume IV, pages 321–333. Univ. California Press, Berkeley, 1961.

[6] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 71(4):1–100, 1969.

[7] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9:1135–1151, 1981.

[8] R. C. Weng. A Bayesian Edgeworth expansion by Stein's Identity. *Bayesian Analysis*, 5(4):741–764, 2010.

[9] R. C. Weng and C.-J. Lin. A Bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12:267–300, 2011.

[10] M. Woodroofe. Very weak expansions for sequentially designed experiments: linear models. *Ann. Statist.*, 17:1087–1102, 1989.

Report on attending useR! 2016 conference, June 27 - 30, 2016, Stanford University, Palo Alto, California.

The annual useR! conference is the main meeting of the international R user and developer community. This year the conference was held at the campus of Stanford University in Stanford, CA. The conference is being organized with support from the Department of Statistics, Stanford University and the Stanford Libraries.

On Monday 2016.6.27, I attended two tutorials. One is Machine Learning Algorithmic Deep Dive, which provides a review of the implementations of several machine learning algorithms. The other is Bayesian Inference Using R Interfaces to Stan, which introduced Bayesian inference using Hamiltonian Markov Chain Monte Carlo as implemented in Stan. Hamiltonian Monte Carlo method is rather than, and may outperform the existing methods in many cases.

I also benefit a lot from several invited talks and contributed sessions. I attended the following invited talks: Richard Becker's Forty Years of S, which described the creation of S at Bell Labs in 1970s; Donald Knuth's Literate Programming, describing his work developing Tex in the 1980s, Hadley Wickham's Towards a Grammar of Interactive Graphics, and Daniela Witten's Flexible and Interpretable Regression Using Convex Penalties, which discussed Fused Lasso Additive Models (FLAM) using piecewise constant functions.

My poster presentation was on Wednesday 2:30-3:30. My work is on Bayesian inference for product ratings data using R. It was great to have a bunch of attendees to discuss with me about my work. Among them, one is from French having co-developed JAG and asking about particle filtering; one is from Google at Mountain View, who is interested in analyzing product ratings; one is from University of Washington, who commented on producing R package, etc.

It is a great conference for R users to meet people who use and develop R tools, and learn more about how statistics and R are used in a variety of applications in the big data era.

During these days, I met some friends from industries and academics. Having chats with them inspired me and encouraged me to keep on moving. It was really a fruitful trip.

科技部補助計畫衍生研發成果推廣資料表

日期:2016/09/11

科技部補助計畫	計畫名稱: 從即時貝氏學習方法到群組學習
	計畫主持人: 翁久幸
	計畫編號: 104-2118-M-004-005- 學門領域: 其他應用統計

無研發成果推廣資料

104年度專題研究計畫成果彙整表

計畫主持人：翁久幸	**計畫編號：**104-2118-M-004-005-
計畫名稱：從即時貝氏學習方法到群組學習	

成果項目					量化	單位	質化 （說明：各成果項目請附佐證資料或細項說明，如期刊名稱、年份、卷期、起訖頁數、證號...等）
國內	學術性論文	期刊論文			0	篇	
		研討會論文			1		
		專書			0	本	
		專書論文			0	章	
		技術報告			0	篇	
		其他			0	篇	
	智慧財產權及成果	專利權	發明專利	申請中	0	件	
				已獲得	0		
			新型/設計專利		0		
		商標權			0		
		營業秘密			0		
		積體電路電路布局權			0		
		著作權			0		
		品種權			0		
		其他			0		
	技術移轉	件數			0	件	
		收入			0	千元	
國外	學術性論文	期刊論文			0	篇	
		研討會論文			1		
		專書			0	本	
		專書論文			0	章	
		技術報告			0	篇	
		其他			0	篇	
	智慧財產權及成果	專利權	發明專利	申請中	0	件	
				已獲得	0		
			新型/設計專利		0		
		商標權			0		
		營業秘密			0		
		積體電路電路布局權			0		
		著作權			0		
		品種權			0		
		其他			0		

	技術移轉	件數	0	件	
		收入	0	千元	
參與計畫人力	本國籍	大專生	0	人次	
		碩士生	3		
		博士生	0		
		博士後研究員	0		
		專任助理	0		
	非本國籍	大專生	0		
		碩士生	0		
		博士生	0		
		博士後研究員	0		
		專任助理	0		
其他成果 （無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。）					

科技部補助專題研究計畫成果自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現（簡要敘述成果是否具有政策應用參考價值及具影響公共利益之重大發現）或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估
   ■達成目標
   □未達成目標（請說明，以100字為限）
   　□實驗失敗
   　□因故實驗中斷
   　□其他原因
   說明：

2. 研究成果在學術期刊發表或申請專利等情形（請於其他欄註明專利及技轉之證號、合約、申請及洽談等詳細資訊）
   論文：□已發表　■未發表之文稿　□撰寫中　□無
   專利：□已獲得　□申請中　■無
   技轉：□已技轉　□洽談中　■無
   其他：（以200字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性，以500字為限）
   近年來由於網際網路發達產生許多大量且即時的資料，因而使即時演算法益加受到關注。本研究討論網路產品評比資料之即時統計分析,研究成果可以有實際之應用。

4. 主要發現
   本研究具有政策應用參考價值：■否　□是，建議提供機關
   （勾選「是」者，請列舉建議可提供施政參考之業務主管機關）
   本研究具影響公共利益之重大發現：■否　□是
   說明：（以150字為限）