Sparse Bayesian Classifiers for Text Categorization - Online Learning

Wen-Hua Ju
whju@avaya.com
David Madigan
madigan@stat.rutgers.edu

December 1, 2002

1 Introduction

We consider the problem of online learning for the sparse Bayesian classifiers described in the companion report "Sparse Bayesian Classifiers for Text Categorization - Batch Learning." The goal of online learning in the sparse Bayesian classification context is to sequentially update the posterior distribution of the model parameters as each new labeled example arrives. The Bayesian paradigm supports this operation in a natural fashion; starting from the prior, the first example produces a posterior distribution incorporating the evidence from the first example. This then becomes the prior distribution awaiting the arrival of the second example, and so on. In practice, however, except in those cases where the posterior distribution has the same mathematical form as the prior distribution, some form of approximation is required to carry out the sequential updating.

Various Monte Carlo algorithms exist to perform online Bayesian learning (see, for example, Ridgeway and Madigan, 2002). These algorithms produce Monte Carlo samples from the posterior distributions of interest. This enables "fully Bayesian" prediction via numerical integration of predictions with respect to the posterior. We may pursue this approach in later work.

For now, as with the batch learning work, we are focusing on the so-called "plug-in" approach that estimates the mode of the posterior distribution (i.e., the mostly likely values for the model parameters given the data) and bases predictions on this one value. This approach is computationally more efficient and may even result in more accurate predictions in certain circumstances (Smith, 1999, Ju, Madigan, and Scott, 2002).

There exists a substantial literature concerned with online learning of maximum likelihood estimators and posterior modes. One branch of this literature has developed online EM algorithms, analagous to the batch EM algorithm. Relevant references include Titterington (1984), Neal and Hinton (1998), and Sato and Ishii (2000). A second branch focuses on so called quasi-Bayes methods. The basic idea is to approximate the posterior at each stage with something tractable (like a Gaussian), update this tractable distribution using Bayes theorem, and then re-approximate the posterior. Relevant references include Smith and Makov (1978), Chai, Chieu, and Ng (2002), and Opper (1998). We plan to investigate both approaches.

2 Proposed Algorithm

We have developed a more direct approximation that is specific to the particular sparse Bayesian model we are currently using. In what follows we will describe this approach to online learning. We are currently implementing this algorithm and are preparing to run some small-scale experiments.

Our proposed approach uses a quadratic approximation to the log-likelihood for each observation. The posterior distribution of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ is:

$$[\boldsymbol{\beta}, \boldsymbol{\tau} | y_1, \ldots, y_t] \propto \left(\prod_{i=1}^{t} [y_i | \boldsymbol{\beta}] \right) [\boldsymbol{\beta} | \boldsymbol{\tau}][\boldsymbol{\tau}]$$

$$= \left(\prod_{i=1}^{t} \{ y_i \Phi(\mathbf{x}_i \boldsymbol{\beta}) + (1 - y_i)(1 - \Phi(\mathbf{x}_i \boldsymbol{\beta})\} \right) \frac{|\Sigma^{-1}|^{1/2}}{(2\pi)^{d/2}} \exp(-\frac{1}{2} \boldsymbol{\beta}^T \Sigma^{-1} \boldsymbol{\beta}) \gamma^d \exp(-\gamma \sum_{i=1}^{d} \tau_i).$$

On the log scale, we have:

$$
\begin{aligned}
&\log[\boldsymbol{\beta}, \boldsymbol{\tau}|y_1, \ldots, y_t] \\
&\propto \sum_{i=1}^{t} \log\left(y_i \Phi(\mathbf{x}_i\boldsymbol{\beta}) + (1-y_i)(1-\Phi(\mathbf{x}_i\boldsymbol{\beta})\right) + \frac{1}{2}\sum_{i=1}^{d} \log \tau_i^{-1} - \frac{1}{2}\sum_{i=1}^{d} \beta_i^2 \tau_i^{-1} - \gamma \sum_{i=1}^{d} \tau_i \\
&\approx \left(\sum_{i=1}^{t} a_i(\mathbf{x}_i\boldsymbol{\beta})^2 + b_i(\mathbf{x}_i\boldsymbol{\beta}) + c_i\right) + \frac{1}{2}\sum_{i=1}^{d} \log \tau_i^{-1} - \frac{1}{2}\sum_{i=1}^{d} \beta_i^2 \tau_i^{-1} - \gamma \sum_{i=1}^{d} \tau_i
\end{aligned}
$$

where $a_i(\mathbf{x}_i\boldsymbol{\beta})^2 + b_i(\mathbf{x}_i\boldsymbol{\beta}) + c_i$ approximates $\log \Phi(\mathbf{x}_i\boldsymbol{\beta})$ when $y_i = 1$ and approximates $\log(1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}))$ when $y_i = 0$, $i = 1, \ldots, t$. In either case the approximation uses a Taylor expansion around $\mathbf{x}_i\hat{\boldsymbol{\beta}}_{i-1}$ where $\hat{\boldsymbol{\beta}}_{i-1}$ estimates the posterior mode given $y_1, \ldots, y_{i-1}$ (see Appendix A). Continuing:

$$
\begin{aligned}
&\log[\boldsymbol{\beta}, \boldsymbol{\tau}|y_1, \ldots, y_t] \\
&\propto \sum_{i=1}^{t} \left(\sum_{j=1}^{d}\sum_{k=1}^{d} a_i x_{ij} x_{ik} \beta_j \beta_k + \sum_{j=1}^{d} b_i x_{ij} \beta_j\right) + \frac{1}{2}\sum_{i=1}^{d} \log \tau_i^{-1} - \frac{1}{2}\sum_{i=1}^{d} \beta_i^2 \tau_i^{-1} - \gamma \sum_{i=1}^{d} \tau_i \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \psi_{ij} \beta_i \beta_j + \sum_{i=1}^{d} \phi_j \beta_j + \frac{1}{2}\sum_{i=1}^{d} \log \tau_i^{-1} - \frac{1}{2}\sum_{i=1}^{d} \beta_i^2 \tau_i^{-1} - \gamma \sum_{i=1}^{d} \tau_i \qquad (1)
\end{aligned}
$$

where:

$$\psi_{ij} = \sum_{l=1}^{t} a_l x_{li} x_{lj}, \quad i, j = 1, \ldots, d$$

and

$$\phi_j = \sum_{l=1}^{t} b_l x_{lj}, \quad j = 1, \ldots, d.$$

The coefficients ψ_{ij} and ϕ_j are updated in an online fashion as data accumulate.

Denote by Q_t the approximate log-posterior of (1). Then:

$$\frac{\partial Q_t}{\partial \beta_i} = 2\sum_{j=1}^{d} \psi_{ij}\beta_j + \phi_i - \beta_i \tau_i^{-1}, \quad i = 1, \ldots, d,$$

and

$$\frac{\partial Q_t}{\partial \tau_i} = -\frac{1}{2}\tau_i^{-1} + \frac{1}{2}\beta_i^2 \tau_i^{-2} - \gamma, \quad i = 1, \ldots, d.$$

Letting $\frac{\partial Q_t}{\partial \beta_i} = 0$ and solving for τ_i, we have:

$$\tau_i = \frac{\beta_i}{\phi_i + 2\sum_{j=1}^{d} \psi_{ij}\beta_j}.$$

Substituting the above into $\frac{\partial Q_t}{\partial \tau_i} = 0$ results in a system of d equations:

$$\phi_i - 2\sum_{j=1}^{d} \psi_{ij}\beta_j - \beta_i(\phi_i + 2\sum_{j=1}^{k} \psi_{ij}\beta_j)^2 + 2\beta_i\gamma = 0$$

where $i = 1, \ldots, d$. This system does not seem to admit a closed form solution for $\boldsymbol{\beta}$.

We are considering two approaches to maximizing (1). One approach to to use a Newton-Raphson procedure directly. Since the posterior mode will change ever more slowly as data accumulate, this should require fewer iterations as t gets large.

A second approach is to alternately maximize (1) with respect to $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$. In particular, denote the current estimate of $\boldsymbol{\tau}$ by $\hat{\boldsymbol{\tau}}$. Then fixing $\boldsymbol{\tau}$ at $\hat{\boldsymbol{\tau}}$, choose the next estimate of $\boldsymbol{\beta}$ by solving:

$$2\sum_{j=1}^{d} \psi_{ij}\beta_j + \phi_i - \beta_i\hat{\tau}_i^{-1} = 0, \quad i = 1, \ldots, d,$$

leading to:

$$\hat{\boldsymbol{\beta}} = -\left(2\Psi - \Upsilon\right)^{-1}(\phi_1, \ldots, \phi_d)^T$$

where Ψ is a $d \times d$ symmetric matrix with ijth element $\psi_{ij}, i, j = 1, \ldots, d$ and Υ is a diagonal matrix with diagonal elements $\hat{\tau}_i^{-1}, i = 1, \ldots, d$. Next update $\boldsymbol{\tau}$ by setting:

$$\hat{\hat{\tau}}_i = \frac{-1 + \sqrt{1 + 8\gamma\beta_i^2}}{4\gamma}, i = 1, \ldots, d.$$

Appendix

Here we show the Taylor expansions for the quadratic approximations to the likelihood. To simplify notation, let $c = \mathbf{x}_i\boldsymbol{\beta}$ and $\hat{c} = \mathbf{x}_i\hat{\boldsymbol{\beta}}_{i-1}$. The standard normal probability density function is denoted by ϕ. First consider the case where $y_i = 1$:

$$\begin{aligned}\log\Phi(c) &\approx \log\Phi(\hat{c}) + (c-\hat{c})\frac{\phi(\hat{c})}{\Phi(\hat{c})} + \frac{(c-\hat{c})^2}{2}\left(\frac{\phi'(\hat{c})}{\Phi(\hat{c})} - \left(\frac{\phi(\hat{c})}{\Phi(\hat{c})}\right)^2\right) \\ &\propto \frac{\phi(\hat{c})}{\Phi(\hat{c})}c + \frac{1}{2}\left(\frac{\phi'(\hat{c})}{\Phi(\hat{c})} - \left(\frac{\phi(\hat{c})}{\Phi(\hat{c})}\right)^2\right)c^2 - \hat{c}\left(\frac{\phi'(\hat{c})}{\Phi(\hat{c})} - \left(\frac{\phi(\hat{c})}{\Phi(\hat{c})}\right)^2\right)c\end{aligned}$$

so that:

$$a_i = \frac{1}{2}\left(\frac{\phi'(\hat{c})}{\Phi(\hat{c})} - \left(\frac{\phi(\hat{c})}{\Phi(\hat{c})}\right)^2\right)$$

and

$$b_i = \frac{\phi(\hat{c})}{\Phi(\hat{c})} - \hat{c}\left(\frac{\phi'(\hat{c})}{\Phi(\hat{c})} - \left(\frac{\phi(\hat{c})}{\Phi(\hat{c})}\right)^2\right).$$

Now consider the case where $y_i = 0$:

$$\begin{aligned}
\log(1-\Phi(c)) &\approx \log(1-\Phi(\hat{c})) - (c-\hat{c})\frac{\phi(\hat{c})}{1-\Phi(\hat{c})} - \frac{(c-\hat{c})^2}{2}\left(\frac{\phi'(\hat{c})}{1-\Phi(\hat{c})} + \left(\frac{\phi(\hat{c})}{1-\Phi(\hat{c})}\right)^2\right) \\
&\propto \frac{-\phi(\hat{c})}{1-\Phi(\hat{c})}c - \frac{1}{2}\left(\frac{\phi'(\hat{c})}{1-\Phi(\hat{c})} + \left(\frac{\phi(\hat{c})}{1-\Phi(\hat{c})}\right)^2\right)c^2 + \hat{c}\left(\frac{\phi'(\hat{c})}{1-\Phi(\hat{c})} + \left(\frac{\phi(\hat{c})}{1-\Phi(\hat{c})}\right)^2\right)c
\end{aligned}$$

so that:

$$a_i = -\frac{1}{2}\left(\frac{\phi'(\hat{c})}{1-\Phi(\hat{c})} + \left(\frac{\phi(\hat{c})}{1-\Phi(\hat{c})}\right)^2\right)$$

and

$$b_i = -\frac{\phi(\hat{c})}{1-\Phi(\hat{c})} + \hat{c}\left(\frac{\phi'(\hat{c})}{1-\Phi(\hat{c})} + \left(\frac{\phi(\hat{c})}{1-\Phi(\hat{c})}\right)^2\right).$$

Acknowledgements

We are grateful to Bill DuMouchel and Colin Mallows for helpful discussions.

References

Chai, K.M.A., Chieu, K.L., and Ng, H.T. (2002). Bayesian online classifiers for text classification and filtering ages.In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 97 - 104.

Ju, W., Madigan, D., and Scott, S. (2002). On sparse Bayesian classifiers. `http://www.stat.rutgers.edu/ madigan/PAPERS/sparse3.pdf`

Neal, R.M. and Hinton, G.E. (1998). A new view of the EM algorithm that justifies incremental, sparse and other variants. In: *Learning in Graphical Models*, M.I. Jordan (Editor). Kluwer Academic Publishers, 355–368.

Opper, M. (1998). A Bayesian approach to online learning. In: *On-Line Learning in Neural Networks*, D. Saad (Editor, Cambridge University Press.

Ridgeway, G. and Madigan, D. (2002). Bayesian analysis of massive datasets via particle filters In *Proceedings of KDD-02, The Eighth International Conference on Knowledge Discovery and Data Mining*, xxx-xxx.

Sato, M. and Ishii, S. (2000). On-line EM algorithm for the normalized gaussian network. *Neural Computation*, **12**, 407–432.

Smith, A.F.M. and Makov, U. (1978). A quasi-Bayes sequential procedure for mixtures. *Journal of the Royal Statistical Society (Series B)*, **40**, 106–112.

Smith, R.L. (1999). Bayesian and frequentist approaches to parametric predictive inference (with discussion). In: *Bayesian Statistics 6*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, pp. 589-612.

Titterington, D.M. (1984). Recursive parameter estimation using incomplete data, *Journal of the Royal Statistical Society (Series B)*, **46**, 257–267.