Computational Statistics and Data Analysis 73 (2014) 163–176

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Mean field variational Bayesian inference for support vector machine classification

Jan Luts [a], John T. Ormerod [b,*]

[a] School of Mathematical Sciences, University of Technology, Sydney Broadway 2007, Australia
[b] School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia

ARTICLE INFO

Article history:
Received 28 August 2013
Accepted 31 October 2013
Available online 17 December 2013

Keywords:
Approximate Bayesian inference
Variable selection
Missing data
Mixed model
Markov chain Monte Carlo

ABSTRACT

A mean field variational Bayes approach to support vector machines (SVMs) using the latent variable representation on Polson and Scott (2012) is presented. This representation allows circumvention of many of the shortcomings associated with classical SVMs including automatic penalty parameter selection, the ability to handle dependent samples, missing data and variable selection. We demonstrate on simulated and real datasets that our approach is easily extendable to non-standard situations and outperforms the classical SVM approach whilst remaining computationally efficient.

1. Introduction

Support vector machines (SVMs) and its variants remain one of the most popular classification methods in machine learning and have been successfully utilized in many applications. Such applications include image classification, speech recognition, cancer diagnosis, natural language processing, forecasting, bio-informatics and as such these methods are likely to remain popular for many years to come. The strengths of SVMs derive from its formulation as an elegant convex optimization problem (involving few tuning parameter) which can be efficiently solved and whose solution only depends on a subset of the input samples, called support vectors.

Despite such popularity standard SVMs suffer from several shortcomings. Section 10.7 of Hastie et al. (2009) summarizes these as: (i) natural handling of data of mixed type, (ii) handling of missing values (iii) robustness to outliers in the input space (iv) insensitive to monotonic transformations of inputs (v) computational scalability to large sample sizes, (vi) inability to deal with irrelevant inputs and (vii) interpretability. To this list we would add (viii) the inability to deal with the correlation within samples. In this paper we aim to address (ii), (vi) and (viii).

This paper is not the first to consider these problems. Missingness has been considered by Smola et al. (2005), Pelckmans et al. (2005) and Nebot-Troyano and Belanche-Muñoz (2010). Dealing with irrelevant inputs via variable/feature selection in SVMs has been considered by many authors including Weston et al. (2000), Tipping (2001), Guyon et al. (2002), Zhu et al. (2003), Gold et al. (2005) and Chu et al. (2006). On the other hand, very few papers consider the modification of SVMs to handle the dependent or non-identically distributed data. Notable exceptions include Dundar et al. (2007), Lu et al. (2011), Pearce and Wand (2009) and Luts et al. (2012). However, these problems are dealt with isolation and using different approaches, rather than in a unified manner and it is difficult to see how these approaches could be adapted to multiple complications, e.g., missingness and variable selection.

* Corresponding author. Tel.: +61 2 9351 5883; fax: +61 2 9351 4534.
E-mail addresses: john.ormerod@sydney.edu.au, jtormerod@hotmail.com (J.T. Ormerod).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.10.030

In the paper we follow the earlier work of Boser et al. (1992), Bishop and Tipping (2000), Gao and Wong (2005) and Polson and Scott (2011) who propose various latent variable representations of the SVM loss function and reformulate the problem in a (pseudo-) Bayesian framework. This provides a unified approach which releases SVMs from many of the above problems including allowing efficient penalty parameter selection, correlation within samples, variable selection and missing data via well developed Bayesian methodology. Typically such Bayesian models are fit via Markov chain Monte Carlo (MCMC) methods. Unfortunately, MCMC methods can be notoriously slow when applied to large or complex models and can be rendered unsuitable in applications where speed is essential. These situations are precisely the same situations where SVMs are typically popular.

Our approach to this problem is to apply the mean field variational Bayes (VB) methods to the models we propose. The main advantage of this approach is a streamlined and computationally efficient framework for handling to many of the problems associated with the classical SVM approach. In tandem with these algorithms we also develop Gibbs sampling approaches to these methods to facilitate comparisons with an "exact" approach to these models.

In Section 2 we provide the framework for our approach. In Section 3 we consider various extensions including automatic penalty parameter selection, group correlations, variable selection and missing predictors respectively. In Section 4 we show how our approach offers several computational advantages over the classical SVM approach. In Section 5 we conclude. Appendices contain details of our MCMC samplers.

Notation

The notation $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that $\boldsymbol{x}$ has a multivariate normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. If x has an inverse gamma distribution, denoted as $x \sim \text{IG}(A, B)$, then it has density $p(x) = B^A\Gamma(A)^{-1}x^{-A-1}\exp(-B/x)$, $x, A, B > 0$. If x has an inverse Gaussian distribution, denoted as $x \sim \text{Inverse-Gaussian}(\mu, \lambda)$ with mean μ and variance μ^3/λ, then it has density

$$p(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\left\{-\frac{\lambda(x-\mu)^2}{2x\mu^2}\right\}, \quad x, \mu, \lambda > 0.$$

If x has a generalized inverse Gaussian distribution, denoted as $x \sim \text{GIG}(\gamma, \psi, \chi)$, then it has density

$$p(x) = \frac{(\psi/\chi)^{\gamma/2}}{2K_\gamma(\sqrt{\psi\chi})}x^{\gamma-1}\exp\left\{-\frac{1}{2}\left(\frac{\chi}{x} + \psi x\right)\right\}, \quad x, \psi, \chi > 0, \ \gamma \in \mathbb{R},$$

where $K_\gamma(\cdot)$ is a modified Bessel function of the second kind. If $\boldsymbol{x}$ is a vector of length d then $\text{diag}(\boldsymbol{x})$ is the $d \times d$ diagonal matrix whose diagonal elements are $\boldsymbol{x}$. If $\boldsymbol{X}$ is a $d \times d$ matrix then $\text{dg}(\boldsymbol{X})$ is the vector of length d comprising of the diagonal elements of $\boldsymbol{X}$. The jth column of a matrix $\mathbf{X}$ is denoted as $\mathbf{X}_j$.

2. Methodology

In this section we present a VB approach to a Bayesian SVM classification formulation for binary classification problems. After introducing Bayesian SVMs and VB methodology we describe the latent variable SVM representation of Polson and Scott (2011) which gives rise to our basic VBSVM approach.

2.1. Bayesian support vector machines

Consider a training set $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ represents an input vector and $y_i \in \{-1, +1\}$ the corresponding class label. SVMs can be formulated in terms of finding a hyperplane that separates the observations with $y_i = 1$ from those with $y_i = -1$ with the largest minimal separating distance or margin. In general such a hyperplane does not exist and the problem needs to be reformulated as a trade-off between the size of the margin and infringements caused by points being on the wrong side of the hyperplane (for more details see for example Vapnik, 1998 or Chapter 12 of Hastie et al., 2009). This optimization problem amounts to finding $\boldsymbol{\beta} \in \mathbb{R}^p$ which minimizes

$$\min_{\boldsymbol{\beta}} \mathcal{J}(\boldsymbol{\beta}) = \left\{\sum_{i=1}^n (1 - y_i\boldsymbol{x}_i^T\boldsymbol{\beta})_+\right\} + \alpha\|\boldsymbol{\beta}\|^2, \tag{1}$$

where α is a positive penalty parameter (the choice of which we will discuss later) and $x_+ = \max(0, x)$. Larger values of α serve to shrink the fitted values of the $\boldsymbol{\beta}$ coefficients. The above problem can be reformulated as a convex quadratic programming problem and can be solved using a variety of efficient methods (for example Chapter 7 of Cristianini and Shawe-Taylor, 2000). This results in the classification rule $\text{sign}(\boldsymbol{x}_i^T\boldsymbol{\beta})$ for input vector $\boldsymbol{x}_i$.

The terms $(1 - y_i\boldsymbol{x}_i^T\boldsymbol{\beta})_+$ in (1) are referred to as the hinge loss of the data and using a logarithmic scoring rule interpretation (Bernardo, 1979) can be interpreted as negative conditional log-likelihoods. This has motivated the Bayesian SVM formulations where

$$p\ell(y_i|\boldsymbol{\beta}) = \exp\left\{-(1 - y_i\boldsymbol{x}_i^T\boldsymbol{\beta})_+\right\}, \quad 1 \le i \le n, \text{ and } \boldsymbol{\beta} \sim N\left(\mathbf{0}, \tfrac{1}{2}\alpha^{-1}\boldsymbol{I}_p\right), \tag{2}$$