Expert Systems With Applications 69 (2017) 277–287

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Fast sampling methods for Bayesian max-margin models

Wenbo Hu, Jun Zhu*, Bo Zhang

Dept. of Comp. Sci. & Tech., State Key Lab of Intell. Tech. & Sys., TNList Lab, Center for Bio-Inspired Computing Research, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 27 March 2016
Revised 8 September 2016
Accepted 16 October 2016
Available online 17 October 2016

Keywords:
Inference
Stochastic MCMC
Subgradient MCMC
Bayesian max-margin models
Approximate detailed balance

ABSTRACT

Bayesian max-margin models have shown superiority in various practical applications, such as text categorization, collaborative prediction, social network link prediction and crowdsourcing, and they conjoin the flexibility of Bayesian modeling and predictive strengths of max-margin learning. However, Monte Carlo sampling for these models still remains challenging, especially for applications that involve large-scale datasets. In this paper, we present the stochastic subgradient Hamiltonian Monte Carlo (HMC) methods, which are easy to implement and computationally efficient. We show the approximate detailed balance property of subgradient HMC which reveals a natural and validated generalization of the ordinary HMC. Furthermore, we investigate the variants that use stochastic subsampling and thermostats for better scalability and mixing. Using stochastic subgradient Markov Chain Monte Carlo (MCMC), we efficiently solve the posterior inference task of various Bayesian max-margin models and extensive experimental results demonstrate the effectiveness of our approach.

1. Introduction

Bayesian max-margin (BMM) models have been shown to be very effective in many real-world applications, such as text analysis (Zhu, Ahmed, & Xing, 2012), collaborative prediction (Xu, Zhu, & Zhang, 2012), social network link prediction (Zhu, 2012) and crowdsourcing (Tian & Zhu, 2015). Such BMM models conjoin the advantages of the discriminative max-margin learning and flexible Bayesian models, and they achieve the best of the both worlds: obtaining the flexibility from a Bayesian model and meanwhile doing discriminative max-margin learning, through a newly-developed unified Bayesian inference framework, regularized Bayesian inference (RegBayes) (Zhu, Chen, & Xing, 2014).

In order to deal with large-scale datasets, developing effective and scalable inference methods is a crucial problem for Bayesian max-margin models, which is becoming a norm in many application areas. Previous variational-approximation-based inference methods are raised to solve the BMM models with mean-field assumptions on posterior distributions (Zhu et al., 2012). When the BMM models use nonparametric Bayesian priors, such variational methods need to adopt the model truncation to finish the variational approximation (Xu, Zhu, & Zhang, 2013; Zhu, Chen, & Xing, 2011). Moreover, in such inference scheme, solving support vector machine (SVM) subproblems is time-consuming, which motivated the further developments of the Gibbs classifier formulation and the data augmentation-based Gibbs sampler (Xu et al., 2013; Zhang, Zhu, & Zhang, 2014; Zhu, Chen, Perkins, & Xing, 2014).

In Bayesian inference, if we use a conjugate prior (w.r.t a given likelihood), we can easily derive the close-form posterior (Gelman, Carlin, Stern, & Rubin, 2014). However, the BMM models are usually non-conjugate due to the non-smoothness of the hinge loss, which is often involved in an unnormalized pseudo-likelihood. The straightforward Gibbs sampler is not applicable due to the non-conjugacy. With a newly discovered data augmentation technique (Polson & Scott, 2011), the augmented Gibbs sampler achieves accurate posterior sampling and is truncation-free for nonparametric BMM models (Xu et al., 2013; Zhang et al., 2014). However, the Gibbs samplers with data augmentation are not efficient either in high-dimensional spaces as they often involve inverting large matrices (Polson & Scott, 2011). Moreover, the benefit of introducing extra variables would be counteracted in the view of the extra computation on dealing with the extra sampling variables (Roberts & Stramer, 2002).

In this paper, we present the subgradient-based Hamiltonian Monte Carlo (HMC) methods for BMM models, which directly draw samples from the original posterior instead of the augmented one. After adopting some mild conditions of the posterior functions, we show the approximate detailed balance property for subgradient HMC methods. Then using stochastic subgradient estimation (Robbins & Monro, 1951; Welling & Teh, 2011), we further

* Corresponding author.
E-mail addresses: hwb13@mails.tsinghua.edu.cn (W. Hu), dcszj@mail.tsinghua.edu.cn (J. Zhu), dcszb@mail.tsinghua.edu.cn (B. Zhang).

http://dx.doi.org/10.1016/j.eswa.2016.10.036
0957-4174/

develop the stochastic subgradient MCMC for fast computation. By annealing the discretization stepsizes properly, our stochastic subgradient MCMC methods approximately converge to the target posteriors of basic Bayesian SVM fairly efficiently. To apply stochastic subgradient MCMC on two different types of BMM models with latent variables, we design two different inference algorithms for latent structure discovery, including a nonparametric Bayesian model. Our stochastic subgradient MCMC can achieve dramatically fast sampling and meanwhile draw accurate posterior samples. We carry out extensive empirical studies on large-scale applications to show the effectiveness and scalability of the presented stochastic subgradient MCMC methods for BMM models.

We note that there have been several previous attempts of using subgradient information in HMC or Langevin Monte Carlo (Neal, 2012; Welling & Teh, 2011), yet our work stands as a first close investigation, in which we give the theoretical guarantee and carry out systematic studies on the stochastic subgradient MCMC for Bayesian max-margin learning.

2. Preliminaries

We first briefly review the Bayesian max-margin models with Gibbs classifiers. Then, we introduce the background knowledge of the inference methods, including Hamiltonian Monte Carlo (HMC) and its extension, as well as stochastic gradient Hamiltonian Monte Carlo.

2.1. Bayesian max-margin models

With the generic framework of *RegBayes* (Zhu, Chen, & Xing, 2014), we can design more flexible Bayesian models by adding proper regularization on the target posterior. Namely, after adding posterior regularization to a functional-optimization-reformulated Bayesian model, a *RegBayes* model generally solves the following problem,

$$\inf_{q(\mathcal{M})\in\mathcal{P}} \mathrm{KL}(q(\mathcal{M})||\pi(\mathcal{M})) - \mathbb{E}_q[\log p(\mathcal{D}|\mathcal{M})] + c\cdot\mathcal{R}(q), \tag{1}$$

where $\mathcal{M}$ denotes the model (parameters); $\mathcal{P}$ is the feasible space of probability distributions $q(\mathcal{M})$; $\mathrm{KL}(q(\cdot)||\pi(\cdot))$ is the KL divergence from the target posterior $q(\mathcal{M})$ to the prior $\pi(\mathcal{M})$; $\mathcal{D}$ is the observation dataset; c is a nonnegative regularization parameter and $\mathcal{R}(q)$ is a well-designed regularization term on q. It is not hard to show that if c equals to 0, the solution of problem (1) is the Bayes posterior $q(\mathcal{M}) \propto \pi(\mathcal{M})p(\mathcal{D}|\mathcal{M})$. If c is not zero, we have an extra dimension of freedom to introduce side information into the inference procedure through the posterior regularization term $\mathcal{R}(q)$. For example, when the regularization $\mathcal{R}$ is defined as a hinge loss in supervised learning tasks, such *Regbayes* models turn out to be Bayesian max-margin models and they successfully incorporate the flexibility of Bayesian models and the max-margin classifiers. This strategy has demonstrated promising performance in various tasks, including text classification and topic extraction (Zhu et al., 2012), social network analysis (Zhu, 2012), and matrix factorization (Xu et al., 2012).

In this paper, we consider two examples of Bayesian max-margin models with latent variables, including *max-margin topic model* (MedLDA) (Zhu et al., 2012) and *infinite SVM* (iSVM) (Zhu et al., 2011). But our methods can be applied to other BMM models. Specifically, MedLDA uses a topic model to find the latent topic representations of the documents and uses a max-margin classifier to do document classification. *Infinite SVM* generally uses a Bayesian nonparametric Dirichlet process prior to describe data multi-modality and meanwhile uses max-margin classifiers to do discriminative tasks. More details of these two examples will be provided along the development of the proposed fast samplers for them.

2.2. BMM models with a Gibbs classifier

In the supervised learning setting, there are generally two types of classifiers that can be used with a Bayesian model to define a BMM model, namely, expected classifiers and Gibbs classifiers. In this part, we give the introduction of the two formulations and analyze the merits of choosing Gibbs classifiers.

Let $\mathcal{D} = \{(x_d, y_d)\}_{d=1}^D$ be a given training set. For each data point $(x_d, y_d) \in \mathcal{D}$, x_d denotes the input features and y_d is the corresponding label, which can be binary or multi-valued. To build a classifier, a Bayesian max-margin model can either use the input features or learn a set of latent features. We use x'_d to denote the features that are fit into a classifier. We consider the linear classifier parameterized by η. Then if the labels are binary, the prediction rule is defined as

$$\hat{y}_d = \mathrm{sgn}\big[f(\eta, x'_d)\big], \quad f(\eta, x'_d) = \eta^\top x'_d, \tag{2}$$

where $\mathrm{sgn}(\cdot)$ is the sign function.

For the above setting, an *expected classifier* learns a posterior distribution $q(\eta)$ in a hypothesis space of classifiers that the q-weighted classifier $\hat{y}_d = \mathrm{sgn}\big(\mathbb{E}_q[f(\eta, x'_d)]\big)$ will have the smallest possible risk, which is typically approximated by the training error $\mathcal{R}_{\mathcal{D}}(q) = \sum_{d=1}^D \mathbb{I}(\hat{y}_d \neq y_d)$, where $\mathbb{I}(\cdot)$ is an indicator function that equals to 1 if predicate holds otherwise 0. We define that $L(y_d, \mathbb{E}_q[f(\eta, x'_d)]) = \max(0, l - y_d\mathbb{E}_q[f(\eta, x'_d)])$ is the hinge loss function with regard to data point d and $l(\geq 1)$ is the cost of making a wrong prediction. Then, we can use the *RegBayes* formulation (Eq. (1)) to define a BMM model with an expected classifier by choosing the loss term $\mathcal{R} = \sum_{d=1}^D L(y_d, \mathbb{E}_q[f(\eta, x'_d)])$. It is known that the hinge loss $\mathcal{R}$ upper bounds the training error $\mathcal{R}_{\mathcal{D}}$.

Alternatively, the *Gibbs classifier* draws a classifier η according to $q(\eta)$ and uses it to do classification, which is proven to have nice generalization performance (Germain, Lacasse, Laviolette, & Marchand, 2009; McAllester, 2003). In the Gibbs classifier, the corresponding loss is the *expected hinge loss*,

$$\mathcal{R}' = \sum_{d=1}^D \mathbb{E}_q[L(y_d, f(\eta, x'_d))]. \tag{3}$$

Since the hinge loss function L is convex, we can show that $\mathcal{R}'$ is an upper bound of $\mathcal{R}$, using Jensen's inequality:

$$\mathbb{E}_q[L(y_d, f(\eta, x'_d))] \geq L(y_d, \mathbb{E}_q[f(\eta, x'_d)]). \tag{4}$$

Then, the *expected hinge loss* $\mathcal{R}'$ is also the upper bound of the expected training error of the Gibbs classifier $\mathcal{R}'(q) \geq \sum_d \mathbb{E}_q[\mathbb{I}(y_d \neq \hat{y}_d)]$. Therefore, the *Gibbs classifier* formulation gives a more relaxed model while at the same time can obtain uncertainty because we draw a single model for each time. In addition, with Gibbs classifiers, truncation-free sampling can be performed for BMM models with Bayesian nonparametric priors, which is more accurate than variational approximation. The BMM models with *Gibbs classifiers* are already shown to have better performance of both classification results and efficiency of the inference algorithms (Xu et al., 2013; Zhang et al., 2014; Zhu, Chen, Perkins, et al., 2014).

2.3. Hamiltonian Monte Carlo

One popular MCMC inference method is Hamiltonian Monte Carlo (HMC), also known as Hybrid Monte Carlo (Neal, 2012). Hamiltonian Monte Carlo is built on the molecular dynamics and the advantage of HMC over random walk Metropolis and Gibbs sampling is proposing a distant move with a high acceptance probability. More recently, the stochastic extensions of HMC are developed for fast sampling.

Formally, we are interested in the posterior distribution $p(\theta|\mathcal{D}) \propto \exp(-U(\theta; \mathcal{D}))$, where θ denotes the variables of inter-