Neural Networks 24 (2011) 1102–1109

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Divergence measures and a general framework for local variational approximation

Kazuho Watanabe [a,*], Masato Okada [b], Kazushi Ikeda [a]

[a] Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan

[b] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, 277-8561, Japan

ARTICLE INFO

Article history:
Received 8 November 2010
Received in revised form 17 May 2011
Accepted 6 June 2011

Keywords:
Local variational approximation
Bayesian learning
Divergence measure
Variational Bayes

ABSTRACT

The local variational method is a technique to approximate an intractable posterior distribution in Bayesian learning. This article formulates a general framework for local variational approximation and shows that its objective function is decomposable into the sum of the Kullback information and the expected Bregman divergence from the approximating posterior distribution to the Bayesian posterior distribution. Based on a geometrical argument in the space of approximating posteriors, we propose an efficient method to evaluate an upper bound of the marginal likelihood. Moreover, we demonstrate that the variational Bayesian approach for the latent variable models can be viewed as a special case of this general framework.

1. Introduction

Bayesian learning is in wide use in many applied data-modeling problems, and it is often accompanied by approximation schemes since it requires intractable computation of posterior distributions. Local variational approximation (LVA), also known as direct site bounding, is a technique to approximate the Bayesian posterior distribution by an analytically computable one. Bishop (2006) provides a tutorial on this scheme in terms of convex duality theory (Jordan, Ghahramani, Jaakkola, & Saul, 1999). Representative applications of this scheme include logistic regression (Bishop, 2006; Jaakkola & Jordan, 2000), the Gaussian process classifier (Gibbs & MacKay, 2000), mixture of experts (Bishop & Svensen, 2003), and sparse linear models (Seeger, 2008, 2009; Wipf & Rao, 2007).

LVA forms lower and/or upper bounds of the unnormalized posterior distribution. The approximation is optimized so as to maximize (minimize) the lower (upper) bound of the normalizing factor, also known as the marginal likelihood of the Bayesian posterior distribution. However, learning algorithms based on LVA have been derived on a case-by-case basis since LVA lacks a general framework and the principle behind it has yet to be described.

In this article, we investigate the relationship between information divergences and LVA to provide a general framework for this scheme. More specifically, we first show that the discrepancy between the log-marginal likelihood, also known as the free energy, and its bound by LVA is expressed as the sum of the Kullback information and the expected Bregman divergence. Decomposing the bounds of the free energy provides a general interpretation of the optimization of LVA in terms of the minimization of the respective divergences. Moreover, it provides an efficient method for computing the lower bound of the free energy by using only the result of the upper bound minimization. We demonstrate how these results apply to practical models by taking two examples. One is for the kernelized logistic regression model, also known as the relevance vector machine. Another is for latent variable models such as the Gaussian mixture model and the hidden Markov model. We show that the so-called variational Bayesian method for the latent variable models can be viewed as a special case of LVA.

The rest of the paper is organized as follows. Section 2 summarizes the general framework of LVA. Section 3 derives equalities for the information divergences related to LVA. Section 4 elaborates on optimization of the approximation and describes an efficient method to combine the lower and upper bounds of the free energy. In Section 5, we demonstrate two applications of LVA to concrete models. Section 6 provides discussions on the comparison of general LVA with other approximation schemes. Section 7 concludes this paper.

2. Local variational approximation

Assume that we are given training examples or observations $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$, where each observation t_i is defined in some domain. Let $\boldsymbol{w} \in R^d$ be the parameter vector, and consider Bayesian learning for a model $p(\boldsymbol{t}|\boldsymbol{w})$.[1] By using the prior distribution $p_0(\boldsymbol{w})$,

[1] The formulation in this paper also applies to discriminative or regression models by simply replacing $p(\boldsymbol{t}|\boldsymbol{w})$ with the conditional distribution $p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w})$ of the outputs $\boldsymbol{t} = \{t_1, \ldots, t_n\}$ given the inputs $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Section 5.1 provides an example of this case.

* Corresponding author.
E-mail address: wkazuho@is.naist.jp (K. Watanabe).

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.06.004

the Bayesian posterior distribution of the parameter $\boldsymbol{w}$ is defined by

$$p(\boldsymbol{w}|\boldsymbol{t}) = \frac{p(\boldsymbol{t}|\boldsymbol{w})p_0(\boldsymbol{w})}{\int p(\boldsymbol{t}|\boldsymbol{w})p_0(\boldsymbol{w})\mathrm{d}\boldsymbol{w}} = \frac{p(\boldsymbol{w},\boldsymbol{t})}{Z}. \tag{1}$$

Then Bayesian learning requires calculating the predictive distribution $p(t|\boldsymbol{t}) = \int p(t|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{t})\mathrm{d}\boldsymbol{w}$ as an estimate of the underlying distribution or, for example, the posterior mean $\int \boldsymbol{w}p(\boldsymbol{w}|\boldsymbol{t})\mathrm{d}\boldsymbol{w}$ as an estimate of the unknown parameter $\boldsymbol{w}$.

However, as is often the case, the normalizing constant of the posterior distribution (1),

$$Z = p(\boldsymbol{t}) = \int p(\boldsymbol{t}|\boldsymbol{w})p_0(\boldsymbol{w})\mathrm{d}\boldsymbol{w},$$

called the marginal likelihood, is analytically intractable, and so is the Bayesian posterior distribution (1). Examples include the case of the logistic regression model (the Bernoulli likelihood) with the Gaussian prior distribution (Jaakkola & Jordan, 2000) and that of the Gaussian model with the sparsity-inducing prior (Seeger, 2008).

Example 1. Consider the following simple example of the pair of the Bernoulli distribution for $t \in \{0, 1\}$ and the Gaussian prior distribution for $w \in R$,

$$p(t|w) = \exp\{tw - \log(1 + \mathrm{e}^w)\} \tag{2}$$

and

$$p_0(w) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-w^2/2}.$$

The Bernoulli distribution has the probability of success ($t = 1$) $\frac{\mathrm{e}^w}{1+\mathrm{e}^w}$. Given the observation t, the integration

$$\int p(t|w)p_0(w)\mathrm{d}w$$

is not analytically tractable, and hence prevents the calculation of the posterior distribution $p(w|t)$. This is mainly due to the existence of the term $-\log(1 + \mathrm{e}^w)$ in the exponent of Eq. (2). We will revisit this example in Sections 3 and 5.1.

LVA forms upper and/or lower bounds of the joint distribution $p(\boldsymbol{w}, \boldsymbol{t})$, denoted by $\overline{p}_{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{t})$ and $\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t})$, respectively. If the bounds satisfying

$$\overline{p}_{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{t}) \geq p(\boldsymbol{w}, \boldsymbol{t}) \quad \text{and} \tag{3}$$

$$\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t}) \leq p(\boldsymbol{w}, \boldsymbol{t}), \tag{4}$$

for all $\boldsymbol{w}$ and $\boldsymbol{t}$, are analytically integrable, then, by normalizing the bounds instead of $p(\boldsymbol{w}, \boldsymbol{t})$, LVA approximates the posterior distribution by

$$p_{\boldsymbol{\lambda}}(\boldsymbol{w}|\boldsymbol{t}) = \frac{\overline{p}_{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{t})}{\overline{Z}(\boldsymbol{\lambda})}, \quad \text{and} \tag{5}$$

$$p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{t}) = \frac{\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t})}{\underline{Z}(\boldsymbol{\xi})}, \tag{6}$$

respectively, where $\overline{Z}(\boldsymbol{\lambda})$ and $\underline{Z}(\boldsymbol{\xi})$ are the normalization constants defined by

$$\overline{Z}(\boldsymbol{\lambda}) = \int \overline{p}_{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{t})\mathrm{d}\boldsymbol{w}$$

and

$$\underline{Z}(\boldsymbol{\xi}) = \int \underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t})\mathrm{d}\boldsymbol{w}.$$

Here, $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ are called the variational parameters, which are introduced to make the bounds adjustable.

The respective approximations are optimized by estimating the variational parameters, $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$, so that $\underline{Z}(\boldsymbol{\xi})$ is maximized and $\overline{Z}(\boldsymbol{\lambda})$ is minimized, since the inequalities

$$\underline{Z}(\boldsymbol{\xi}) \leq Z \leq \overline{Z}(\boldsymbol{\lambda}) \tag{7}$$

hold by definition.

To consider the respective LVAs in terms of information divergence in later sections, let us introduce the free energy,

$$F = -\log Z,$$

following statistical mechanical terminology, and its lower and upper bounds, $\underline{F}(\boldsymbol{\lambda}) = -\log \overline{Z}(\boldsymbol{\lambda})$ and $\overline{F}(\boldsymbol{\xi}) = -\log \underline{Z}(\boldsymbol{\xi})$. By taking the negative logarithms on both sides of Eq. (7), we have

$$\underline{F}(\boldsymbol{\lambda}) \leq F \leq \overline{F}(\boldsymbol{\xi}). \tag{8}$$

Hereafter, we follow the measure of the free energy and adopt the phrases the lower bound maximization ($\underline{F}(\boldsymbol{\lambda})$ maximization) and the upper bound minimization ($\overline{F}(\boldsymbol{\xi})$ minimization) to signify the respective local variational approximations (5) and (6).

3. Divergence measures in LVA

In this section, we derive key equations relating LVA with information divergence. Most existing LVA techniques are based on the convexity of the log-likelihood function or the log-prior (Bishop, 2006; Seeger, 2008). We describe these cases by using general convex functions, ϕ and ψ, and show that the objective functions

$$\overline{F}(\boldsymbol{\xi}) - F = \log \frac{Z}{\underline{Z}(\boldsymbol{\xi})} \geq 0$$

and

$$F - \underline{F}(\boldsymbol{\lambda}) = \log \frac{\overline{Z}(\boldsymbol{\lambda})}{Z} \geq 0$$

to be minimized in the approximations (5) and (6) are decomposable into the sum of the Kullback information and the expected Bregman divergence.

Let ϕ and ψ be twice differentiable real-valued strictly convex functions, and denote by d_ϕ the Bregman divergence associated with the function ϕ (Banerjee, Merugu, Dhillon, & Ghosh, 2005),

$$d_\phi(\boldsymbol{v}_1, \boldsymbol{v}_2) = \phi(\boldsymbol{v}_1) - \phi(\boldsymbol{v}_2) - (\boldsymbol{v}_1 - \boldsymbol{v}_2) \cdot \nabla\phi(\boldsymbol{v}_2) \geq 0, \tag{9}$$

where $\nabla\phi(\boldsymbol{v}_2)$ denotes the gradient vector of ϕ at $\boldsymbol{v}_2$.

Let us consider the case when ϕ and ψ are respectively used to form the following bounds of the joint distribution $p(\boldsymbol{w}, \boldsymbol{t})$,

$$\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t}) = p(\boldsymbol{w}, \boldsymbol{t}) \exp\{-d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi}))\}, \tag{10}$$

$$\overline{p}_{\boldsymbol{\lambda}}(\boldsymbol{w}, \boldsymbol{t}) = p(\boldsymbol{w}, \boldsymbol{t}) \exp\{d_\psi(\boldsymbol{g}(\boldsymbol{w}), \boldsymbol{g}(\boldsymbol{\lambda}))\}, \tag{11}$$

where $\boldsymbol{h}$ and $\boldsymbol{g}$ are vector-valued functions of $\boldsymbol{w}$.[2]

Eq. (10) is interpreted as follows. $\log p(\boldsymbol{w}, \boldsymbol{t})$ includes a term that prevents analytic integration of $p(\boldsymbol{w}, \boldsymbol{t})$ with respect to $\boldsymbol{w}$. If such a term is expressed by the convex function ϕ of some function $\boldsymbol{h}$ transforming $\boldsymbol{w}$, it is replaced by the tangent hyperplane, $\phi(\boldsymbol{h}(\boldsymbol{\xi})) + (\boldsymbol{h}(\boldsymbol{w}) - \boldsymbol{h}(\boldsymbol{\xi})) \cdot \nabla\phi(\boldsymbol{h}(\boldsymbol{\xi}))$, so that $\log \underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t})$ makes a simpler function of $\boldsymbol{w}$, such as a quadratic function. Remember that, if $\log \underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t})$ is quadratic with respect to $\boldsymbol{w}$, $\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{t})$ is analytically integrable by the Gaussian integral.

[2] The functions $\boldsymbol{g}$ and $\boldsymbol{h}$ (also ψ and ϕ) can be dependent on $\boldsymbol{t}$ in this discussion. However, we denote them as if they are independent of $\boldsymbol{t}$ for simplicity. They are actually independent of $\boldsymbol{t}$ in the example of Section 5.1 and in most applications (Bishop, 2006; Seeger, 2008).