International Core Journal of Engineering Volume 6 Issue 12, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

Variational Inference in Bayesian Computation

Ningxin Liu[1,*], Yilan Jiang[2], Jingxuan Cai[3] and Shijie Zhou[4]

[1]School of Natural Science, University of Texas at Austin, Texas 78705, United States;

[2]School of Mathematics, University of California San Diego, California 92092, United States;

[3]School of Mathematics and Science, Jinagsu University, Zhenjiang, Jiangsu 212013, China;

[4]School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China.

*ningxinliu98@gmail.com

Abstract

In Bayesian statistics, there always exists high demand for computational methods to approximate intractable distribution func- tions. Among these Bayesian computing methods, there exists a family of algorithms, known as the Variational Bayesian Inference. This paper mainly discusses the essence of general variational inference and its application to a few simple hidden markov models. We intend to explicitly explain our derivations in detail so that people with basic statistical knowledge also get to understand the approach of variational approxi- mation and other related knowledge.

Keywords

Variational Inference, Hidden Markov Model, Forward-backward algorithm, DIC, WAIC.

1. Introduction

Over the course of the twenty-first century, the demand of Artificial Intelligence such as speech recognition and cognitive services have boomed over the entire world. In order to fulfill the demand, the supply referring to various kinds of computing methods and models cannot be ignored in the field of statistics. This paper will mainly introduce one Bayesian Variational Computing Method and its application in Hidden Markov Models. The major purpose of Variational Inference is to approximate the posterior probability of unobserved variables, make further statistical inferences. Over the past few years, researchers focus more on sampling methods, such as Markov Chain Monte Carlo (MCMC), to approximate the distribution mainly for two reasons. One is that MCMC gets to approximate the exact posterior distribution of the unobserved variables while Variational Inference can only produce an approximation of them. The second is that MCMC algorithms like Gibbs Sampler are easier to implement, while variational computing methods usually require relatively complicated equations. However, Variational Inference has its own charm as it is computationally faster than that of MCMC in most cases. In particular, variational inference works much faster with big data than MCMC.

2. Variational Inference

Variational inference is a family of Bayesian computing techniques, mainly de- signed to approximate the conditional density of latent variables given observed variables. Latent variables, different from observed variables, are something that we cannot directly observe visually and therefore need statistical inferences to draw possible conclusions. Rather than using simulation to approximate in- tractable posterior distribution of latent variables, like Markov Chain Monte Carlo, variational

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

inference tries to cast inference problems as an optimization one and eventually produce a relatively closest estimation to the actual distribution. A problem is described as intractable if it can be solved in theory but any kinds of solutions in practice require too much resources or high time complexity to achieve. One well-known intractable distribution can be observed in a Bayesian mixture of Gaussians model [1] (Blei et al., 2017). Blei et al. argue that the time complexity to compute the marginal distribution of observed variables requires $O(K^n)$ to evaluate K-dimensional integral.

2.1 Definitions

Let $X = x_{1:n} = \{x_1, x_2, ..., x_n\}$ be a set of observed variables and $Z = z_{1:m} = \{z_1, z_2, ..., z_m\}$ be a set of latent variables. We will assume all random variables here are discrete. Continuous random variables can be done in the same way but we will not discuss it explicitly in this paper. It is also worth to note that the number of variables in set X and Z are both denoted as n in other section.

Then $P(X,Z)$ Is joint probability mass function. The ultimate goal is to compute the conditional probability function, $p(Z|X)$. the Baye's rule said,

$$p(Z|X) = \frac{p(Z.X)}{p(X)} \tag{1}$$

The denominator in Equation 1 is often called marginal likelihood. We can calculate $p(X)$ from $p(Z|X)$ by,

$$p(X) = \sum_{Z_{1:n}} p(Z,X) \tag{2}$$

It is often difficult to compute the marginal likelihood in a closed form in many models such as the Bayesian Poisson mixed model (Ormerod and Wand, 2010). In our real lives, there are tons of intractable posterior distribution waiting to be computed. Then let $q(Z) \in Q$ denoted as a candidate approximation for the posterior distribution, $p(Z|X)$ [2] (Blei et al., 2017).

2.2 The Kullback-Leibler Divergence

To find an optimal $q(Z)$ similar to the actual posterior distribution, the first question is how do we know the difference between $q(Z)$ and $p(Z\,X)$. In other words, we want to measure how similar two models are. One typical measurement is called the Kullback-Leibler divergence, also known as the relative entropy. It was first introduced by Solomon Kullback and Richard Leibler in 1951 to respond on their concern on information discrimination. KL divergence is defined intuitively as it is the difference between the expectations of two target functions inside the logarithm with respect to $q(Z)$,

$$KL(q(Z)||p(Z|X) = E_{q(Z)}[log\, q(Z)] - E_{q(Z)}[log\, p(Z|X) = \sum_{Z_{1:n}} q(Z) \frac{q(Z)}{p(Z|X)} \tag{3}$$

Here are some important properties of KL divergence. For any distribution func- tions f, g,

1 It is not symmetric as $KL(f||g) = KL(g||f)$ in general.

2 If f is exactly the same as g, then $KL(f||g) = 0$

3. $KL(f||g) \geq 0$ for any f,g.

Theorem 1 *Jensen's Inequality For any convex function h and a random vari- able X,* $h(E[X]) \leq E[h(x)]$.

Then as $h(x) = -log(x)$ is a convex function

$$KL(f||g) = \sum_Z f(z) log \frac{f(z)}{g(z)} = -\sum_Z f(z)\, log \frac{g(z)}{f(z)}$$

$$= -E \sum_Z f(z) log \frac{g(z)}{f(z)} = -log(\sum_Z f(z) \frac{g(z)}{f(z)} = 0 \tag{4}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

The optimization problem we are trying to resolve here is,

$$q*(Z) = argminq(Z)\in QKL(q(Z)||p(Z|X) \quad (5)$$

where Q is a family of distribution functions we want to specify in the process of variational inference. Notice that even though KL-divergence is useful in many fields such as speech recognition [3] (Hershey et al., 2007), it cannot be directly computed in this case since the KL-divergence equation contains the intractable posterior distribution we want to approximate. The inability to calculate the KL-divergence directly forces us to have another way of measuring the similarity between two distribution functions, which will be introduced in the next subsection.

2.3 The Evidence Lower Bound

The goal here is to compute the posterior distribution function, $P(Z\ X)$, in (1). In (1), this is often more difficult to figure out what $P(X)$, the marginal likeli- hood, is. Therefore, let us try to determine the marginal likelihood by taking the logarithm of it, in which logarithmic functions allows us to simplify logarithms

when their inputs are either a quotient or a product.

$$logp(X) = log(\sum_{Z_{1:n}} p(X,Z)) = log(\sum_{Z_{1:n}} \frac{p(X,Z)}{q(Z)})q(Z)$$
$$= E_{q(Z)}[log\ p(X,Z)] - E_{q(Z)}[log\ q(Z)] \quad (6)$$

The evidence lower bound, abbreviated as ELBO, is derived in (5). The name is intuitive as it is the lower bound of p(X), defined as the evidence.Part 2 of ELBO in (5) is often called Entropy, a concept initially used in thermodynamic system. In this context, we explain it as the average uncertain information we can get from it, while part 1 represent the average amount of information we can get from the latent variables, Z in this case. If we want to maximize the ELBO, then there exists an implicit trade-off between Part 1 and Part 2 since if we get too much information from Z in Part 1, then the uncertainty will be diminished in Part 2.

Note that the ELBO and KL-divergence are closely related,

$$ELBO(q) = E_{q(Z)}[log\ p(X,Z)] - E_{q(Z)}[log\ q(Z)] = E_{q(Z)}[\log p(X)] + E_{q(Z)}[\log p(Z|X)$$
$$-E_{q(Z)}[log\ q(Z)] = log\ p(x) - KL(q(Z)||(Z|X) \quad (7)$$

Therefore, ELBO is the same as negative of KL-divergence plus some constant not depending on $q(Z)$. Recall achieving the goal, we want to minimize KL- divergence, which is the same as maximize the ELBO, according to (6). Therefore, the ELBO equation is our objective function in the optimization.

2.4 The Mean-Field Variational Family and its Effect

Variational Inferences refer to any computing methods that involves posterior distribution approximation through optimization. Among them, the most com- mon one is called the mean-field approximation which assumes that all latent variables are independent and can be partitioned as,

$$q(Z) = \prod_{i=1}^{n} q_i(z_i)\ for\ Z = \{z_1, z_2, \ldots, z_n\} \quad (8)$$

where $q(Z) \in Q$ in which Q is called the Mean-Field Variational Family. Note that each $q_i(z_i)$ can be distinct distribution function. Normally approximation with Mean-Field Variational Family cannot have high similarity with the actual posterior distribution because latent variables are normally

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

dependent. For ex- ample, let z_1 and z_2 be two latent variables and they are positively correlated. Then the mean-field approximation may not make accurate inference as the joint posterior approximation $q(z_1, z_2 X)$ is always a circle under the mean-field assumption while the actual posterior distribution should be an ellipse [4] (Ormerod and Wand, 2010).

Even though the mean-field methods has some shortcoming, they have been widely applied in Neural Computing as framework of ensemble learning [5] (Opper and Winther,1999).The major advantage of the mean-field approximation is that it allows us to study the behavior of high-dimensional distribution functions through a lot of simpler functions.

3. Hidden Markov Model (HMM)

3.1 Introduction of HMM

A Hidden Markov Model (HMM) is a statistical model with the memoryless property (often known as Markov property). The difference between a HMM and a general Markov chain is that a HMM, by its name, contains a sequence of unobserved (or hidden) states. As the simplest dynamic Bayesian network, a hidden markov model has been widely studied in many scientific fields, especially in data science, because of its ability to handle real-world applications

3.2 Characteristics

The two major characteristics of a problem that can be modeled as HMM are as follows.

1 A problem is based on sequences, such as time series, sequence of states.

2 There are two types of data: one type of sequential data is observable, that is, the observation sequence; and the other type of data is unobservable which is the hidden state sequence, referred to as the state sequence.

3.3 Definition

Considering the above description is not precise, below I use accurate mathematical notional to express HMM.

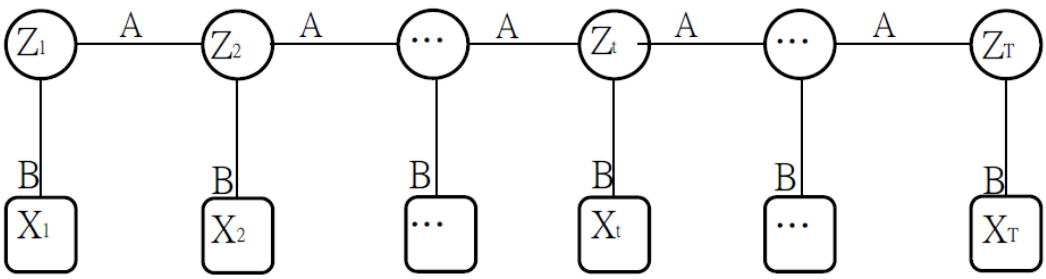

Figure 1 HMM model with hidden states

An HMM model can be determined by the hidden state initial probability distribution π, the state transition probability matrix A and the observed state probability matrix B. π, A determines the sequence of states, and B determines the sequence of observations. Therefore, the HMM model can be represented by a triple λ as follows: λ = (A, B, π)

3.4 Basic questions

There are three basic problems in the HMM models needed to be solved.

1) Evaluate the probability of the observed sequence. That is, given the model $\lambda = (A, B, \pi)$ and the observation sequence $X = (x_1, x_2, ..., x_T)$, calculated in the model. The probability $P(X \lambda)$ of the observed sequence X under λ. The forward and backward algorithm is needed to solve this problem.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

2) Model parameter learning problem. That is, given the observation sequence $X = (x_1, x_2, ..., x_T)$, estimate the parameters of the model $\lambda = (A, B, \pi)$,The conditional probability $P(X\lambda)$ of the observed sequence under the model is maximized. Solving this problem requires the use of EM-based algorithms. About forecast problem. Given the model $\lambda = (A, B, \pi)$ and the observation

4. Bayesian model evaluation

4.1 Introduction

DIC(Deviance Information Criterion), As Spiegelhalter, Best, Carlin and van der Linde[6] (2002) mentioned, is a model selection criterion that would best predict a replicate dataset that has the same structure as that currently observed. It is a Bayesian method for model comparison that can be calculate for many different models. DIC is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by MCMC simulation. DIC is only valid when the posterior distribution is approximately multivariate normal. The first term of DIC is a measure of how well the model fits the data, while the second term is a penalty on the model complexity. Besides, DIC can also be extended to latent variable models by using the variational approximation.

The initial derivation of DIC focused on exponential-family models was written in McGrory and Titterington[7] (2006), which used the variational approximation in a mixture model setting.

4.2 Equation

θ:parameters that appear in the stated sampling distribution y:

For a likelihood $p(y|\theta)$, we define the deviance as

$$D(\theta) = -2 \log L(data|\theta) \tag{9}$$

Authors also suggested that using posterior mean deviance as a measure of fit

$$D = E[D] \tag{10}$$

Then we take Bayesian measures of model dimensionality as (Spiegelhalter et al,2002) said:

$$PD = E\theta|y[d\theta(y, \theta, \tilde{\theta}(y)] = E\theta|y[-2 \log p(y|\theta)] + 2 \log p(y|\tilde{\theta}(y)) \tag{11}$$

Complexity measured by estimate of the 'effective number of parameters' DIC is then define as:

$$DIC = D(\theta) + 2pD = D + pD. \tag{12}$$

4.3 Comparison between DIC, AIC and BIC

DIC is intended as a generalization of Akaike's Information Criterion. For non- hierarchical models with little prior information, p_D should be approximately the true number of parameters. However, unlike AIC, DIC takes prior information into account. The advantage of DIC over AIC and BIC is that DIC is easily calculated form the samples generated by a MCMC simulation.

DIC differs from Bayes factors and BIC in both form and aims. BIC requires specification of the number of parameters, while DIC estimates the effective number of parameters. Since AIC and BIC require calculating the likelihood at its maximum over which is not readily available from the MCMC simulation. DIC can easily calculate by computing as the average of over the samples. DIC cannot provide model averaging procedure just like BIC did.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

5. Variational Inference for Gaussian Hidden Markov models

A hidden Markov model is a process generated by a stationary Markov chain which makes a transition to a different state or stay in the current state. The state sequence of the Markov chain cannot be observed in the HMM and the state in each time point can emit a observation which in turn makes of a distorted version of the state sequence. Given all previous states, the probability of occupying a state z_i at time i only depends on the state at the time-point $i - 1$. And we can define the probability of transition from state z_i to state $z_i + 1$ as a transition matrix

$$\pi = \{\pi_{j1,j2}\},\ 1 \le j_1 j_2 \le K \tag{13}$$

5.1 Assigning the prior distribution

Dirichlet distributions are commonly used as prior distributions in Bayesian statistics as it is the conjugate prior to the categorical distribution and the multinomial distribution. So we take the Dirichlet prior distribution for the transition probability, given by

$$p(\pi) = \prod_{j_1} Dir(\pi_{j1}|\{\alpha^{(0)}_{\pi_{j1,j2}}\}) \tag{14}$$

$\alpha^{(0)}_{\pi_{j1,j2}}$ are hyperparameters. For each state j, we assign univariate Gaussian distribution for emission pdf $p_j(y_i|\phi_j)$ with means μ and variance τ, so that

$$\begin{aligned} p_j(y_i|\phi_j) &= N(y_i|\mu_{j,}\tau_j^{-1})) \\ \phi_j &= (\mu_{j,}, \tau_j) \end{aligned} \tag{15}$$

We then assign precisions as independent gamma prior distribution, and the means are assigned independent univariate Gaussian conjugate prior distributions, given the precisions, so that

$$p(\tau) = \prod_{j=1}^{K} N(\tau_j|1/2\eta^{(0)}, 1/2\delta^{(0)})_j \tag{16}$$

and

$$p(\mu_{j,}|\tau) = \prod_{j=1}^{K} N(\mu_{j,}|m^{(0)}, (\beta^{(0)}\tau_j) - 1) \tag{17}$$

So μ, τ can be decided by hyperparameters $\{m^{(0)}, \beta^{(0)}, \eta^{(0)}, \delta^{(0)}\}$

5.2 Application to a simulated example

We apply the algorithm to datasets simulated from a Gaussian HMM. We use the transition matrix and emission parameters (μ = 3, 2, 1) in McGory and Titterington's paper. We simulated 3 datasets comprising 150, 200 and 500 observations. For each dataset, we applied 20, 15, 10, 5 initial states. And we got the result as following.

Table 1 DIC observation based on different initial states

No.of initial states	DIC of 150 observation	DIC of 200 observation	DIC of 500 observation
20	-244.12	-396.81	-591.36
15	-231.74	-195.66	-601.68
10	-171.21	-208.31	-605.18
5	-134.37	-315.43	-571.73
estimated posterior	3.12,1.57,0.84	2.79,2.01,0.85	2.83,1.88,0.94

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0048

6. Conclusion

Overall, This paper mainly introduces the idea of variational inference and one of its most common algorithms, the Coordinate Ascent Mean-Field Variational Inference (CAVI). We also further explained the hidden markov model and its inference algorithm. Furthermore, in chapter 5 we briefly introduce Deviance Infor- mation Criterion (DIC) and Satanable-Akaike Information Criterion (WAIC) to evaluate the model we got in terms of the amount of information the models preserved. Lastly, we use Gaussian hidden Markov models to solve a simulated example.

References

[1] Vehtari Aki, Gelman Andrew, Gabry Jonah.: Practical Bayesian model evalu- ation using leave-one-out cross-validation and WAIC. Statistics and Computing, September 2017, Volume 27, Issue 5, pp 1413-1432

[2] Gelman Andrew, Hwang Jessica and Vehtari Aki.: Understanding predictive information criteria for Bayesian models.Statistics and Computing, November 2014, Volume 24, Issue 6, pp 997-1016.

[3] McGrory.C.A, Titterington D.M.: Variational Bayesian Analysis for Hidden Makrov Models. Aust. N.Z.J. Stat. 51(2), 2009, 227-244.

[4] Ormerod J.T., Wand M.P.: Explaining Variational Approximations. The Amer- ican Statistician Vol. 64, No.2, pp. 140-153.

[5] Blei David, Kucukelbir Alp, McAuliffe Jon.: Variational Inference: A Review for Statisticians. Cornell University.

[6] Opper Manfred, Winther Ole.: Mean field methods for classification with Gaussian processes.

[7] Bishop, Christopher.: Pattern Recognition and Machine Learning.