Variational Inference in Bayesian Models

Justin H. Le
Department of Electrical & Computer Engineering
University of Nevada, Las Vegas
lejustin.lv@gmail.com

August 14, 2016

1 Introduction

In Bayesian models, latent variables x influence the distribution of observations y, which we express probabilistically as $p(x, y) = p(y \mid x)p(x)$. To perform inference, we allow observations to influence our calculation of the posterior distribution of latent variables:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \tag{1}$$

The factor $p(y)$, known as the *evidence*, involves integration of the joint distribution $p(x, y)$ over all latent variables, rendering the above expression intractable for many of the models that interest us in practice. In these cases, approximate inference becomes necessary. Here, we discuss one popular approximate method known as variational inference.

2 Exact inference in graphical models

Let V be the set of nodes of a graph. On this graph, examples of inference include:

- Compute the marginal distribution $p(x_A)$ over a set of nodes $A \subset V$.
- Compute the conditional distribution $p(x_A \mid x_B)$ over two disjoint sets of nodes, A and B, where $A \cup B \subset V$.

In either situation above, we may consider A to be a set of *hidden* nodes and B to *observed* nodes (which we herein refer to as *evidence*).

A marginal distribution $p(x_m)$ at node x_m can be computed by summing/integrating the joint distribution $p(x_1, x_2, \ldots, x_m, \ldots)$ over all $x \neq x_m$. To perform exact inference of this marginal, we manipulate factors of the joint distribution to obtain a computable expression, e.g., a factor of the joint distribution can be moved out of a sum/integral that doesn't involve the variables of that factor. *Message-passing* can be used to improve the efficiency of

this calculation. Examples of message-passing for exact inference include the junction tree algorithm (for cyclic graphs) and the sum-product algorithm (for trees).

Sum-product message-passing can also be used to perform inference on graphs, in which case it produces an approximate result. Used in this manner, it is referred to as *loopy belief propagation*.

To be continued.

3 Variational approximation of distributions

Let x be random variable and $f(x)$ be its probability density. If $f(x)$ is concave, it can be expressed as a function of its *convex dual*, $f^*(x)$:

$$f(x) = \min_{\lambda}[\lambda^T x - f^*(\lambda)], \tag{2}$$

where λ is the *variational parameter*. The expression shows that $f(x)$ can be bounded above by a linear function of x, whose slope and bias are determined by λ. For a given x, the exact $f(x)$ can be recovered by choosing an optimal λ. In the setting of probabilistic inference, the above insights allow for the efficient computation of a joint distribution by approximating the associated local distributions variationally.

It is not straightforward to obtain this approximation for a general graph, whose nodes are each a random variable with, e.g., density $f(x)$. However, when variational bounds can be obtained for $f(x)$, they can effectively unlink x from its neighbors. Exact inference then becomes tractable on the remaining linked nodes.

Given the task of computing a posterior distribution:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{\int_x p(x, y)}, \tag{3}$$

where x is hidden, and y is observed (evidence), we often find the computation of the *partition function* in the denominator to be intractable. We instead seek a variational approximation to the posterior by formulating its calculation as an optimization problem:

$$\min_q[\mathbb{E}_q q(x) - \mathbb{E}_q p(x \mid y)] = \min_q[\mathrm{KL}(q(x) \parallel p(x \mid y))], \tag{4}$$

where q is a variational distribution, or a member of a *variational family* (the set of variational distributions over x). Eq. 4, which expresses the minimal Kullback-Leibler divergence $\mathrm{KL}(\cdot \parallel \cdot)$ between two distributions, can be solved efficiently for a variational family whose members facilitate the calculation of the expectations.

One such family is the *mean-field* family, in which each member is a distribution over independent hidden variables x, and where each x is associated with a separate variational parameter. Due to this assumption of independence and separation of parameters, each mean-field distribution can be expressed as a factored product of conditionals:

$$q(x) = q(z, \theta) = p(\theta \mid w) \prod_{n \in N} \prod_{j \in J} p(z_{nj} \mid \pi_{nj}), \tag{5}$$

where x has been represented by two sets of hidden variables, θ and z (see below). N and J are the number of contexts and the dimensionality of z, respectively, while w and π are the variational parameters of θ and z, respectively.

We refer to θ as the global variables of the model and z as the local variables. This distinction between global and local variables suits a wide variety of successful probabilistic models. As an example, for a Gaussian mixture model, the global variables correspond to the weights, means, and variances of components in the mixture, while the local variables correspond to the labels assigned to observations. For a hidden Markov model, the global variables are the transition and emission parameters, while the local variables are the states of the hidden Markov chain (which can likewise be interpreted as labels for observations). We refer to the index of an observation as the context n of z_n and y_n. In the HMM, a context is a point in time.

Assume that each local conditional $p(y_n, z_n \mid \theta)$ and the global prior $p(\theta)$ form a conjugate pair of an exponential family. Then a simple, closed-form expression can be obtained for updating the variational parameters iteratively until an optimal q is obtained.

To be continued.

4 Structured mean-field inference

Eq. 5 can be interpreted as a *full* mean-field distribution, in contrast with a *structured* one. For the HMM, The structured mean-field distribution does not factorize the local hidden variable z:

$$q(z, \theta) = p(\theta \mid w)p(z \mid \pi). \tag{6}$$

By retaining the sequential structure of z, this approach preserves information that may be encoded in its ordering, which becomes relevant when using an HMM to model time-dependent phenomena, in which the evolution of the state may be as informative as its instantaneous value, if not more. Other models may have other structures to be exploited and may lead to different factorizations than Eq. 6.

To see the effect of factorizing q, we first express the optimization problem 4 as its dual by maximizing a lower bound on the evidence:

$$\begin{aligned}
\log p(y) &= \log \int d\theta \sum_z p(\theta)p(y, z \mid \theta) \\
&= \log \int d\theta \sum_z p(\theta)p(y, z \mid \theta)\frac{q(z, \theta)}{q(z, \theta)} \\
&= \log \left\langle \frac{p(\theta)p(y, z \mid \theta)}{q(z, \theta)} \right\rangle_q \\
&\leq \left\langle \log \frac{p(\theta)p(y, z \mid \theta)}{q(z, \theta)} \right\rangle_q,
\end{aligned} \tag{7}$$

where the last line results from Jensen's inequality and the concavity of log. Then, we apply the structured mean-field approximation to obtain a cleaner bound:

$$\log p(y) \leq \left\langle \log \frac{p(\theta)p(y, z \mid \theta)}{q(z)q(\theta)} \right\rangle_q$$
$$= \left\langle \log \frac{p(\theta)}{q(\theta)} \right\rangle_q + \left\langle \log \frac{p(y, z \mid \theta)}{q(z)} \right\rangle_q . \tag{8}$$

We refer to 8 as the *evidence lower bound* or *ELBO*. Recall that in the mean-field approach, each variational distribution $q(\theta)$ and $q(z_n)$ is determined by its own parameter, w and π, respectively. Hence, to compute the optimal q, an update must be performed iteratively for each parameter. Updates on w depend on the gradient of the ELBO with respect to w, and updates on π are computed similarly.

Gradients can be obtained easily given the conjugacy described in Section 3. However, each iteration of these updates relies on the entire set of observations and thus fails to scale with the size of the data, which additionally prevents this method from being feasible for a streaming setting. To avoid this cost, the updates can be modified to each rely on only one instance (or small subset) of the data, sampled at random.

To be continued.

References

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational Inference: A Review for Statisticians". In: *arXiv preprint arXiv:1601.00670* (2016).

[2] Nicholas Foti et al. "Stochastic variational inference for hidden Markov models". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 3599–3607. URL: http://papers.nips.cc/paper/5560-stochastic-variational-inference-for-hidden-markov-models.pdf.

[3] Matthew D. Hoffman et al. "Stochastic Variational Inference". In: *Journal of Machine Learning Research* 14 (2013), pp. 1303–1347. URL: http://jmlr.org/papers/v14/hoffman13a.html.