CSC 2541: Bayesian Methods for Machine Learning

Radford M. Neal, University of Toronto, 2011

Lecture 10

Alternatives to Monte Carlo Computation

Since about 1990, Markov chain Monte Carlo has been the dominant method of computation for Bayesian models. Just before that, Monte Carlo using importance sampling was widely used. But what are the alternatives?

Some older methods:

- Analytical solutions. Usually confined to models with conjugate priors.
- Numerical quadrature (eg, Simpson's Rule). Applicable only to very low-dimensional problems.
- Gaussian approximation at the mode / Laplace's method of integration. Once a common method, still of some interest even in high dimensions.

Some newer methods:

- Quasi-Monte Carlo. Doesn't use random points, but rather points that look sort of random, but are designed to be more uniform.
- Variational methods. Look for an approximation in a tractable class of distributions.
- Methods related to "loopy belief propagation". Apply a method that works with acyclic graphs even when your graph has cycles.

Gaussian Approximation at the Mode

One can show that for many Bayesian models, the posterior distribution approaches a Gaussian form as the number of data points increases.

We might therefore approximate the posterior by a Gaussian distribution with the same mode, and the same curvature of the density at the mode.

Suppose that we wish to approximate $\pi(\theta|\text{data}) = (1/Z)\ \exp(h(\theta))$, where $h(\theta) = \log\text{prior} + \log\text{likelihood}$.

We find the location of the mode, θ^*, and the Hessian matrix of second derivatives at the mode, $h''(\theta^*)$.

The approximating Gaussian has mean θ^* and covariance matrix $[-h''(\theta^*)]^{-1}$.

We can see this by matching to the log of a multivariate Gaussian density:

$$\log N(\theta;\, \mu,\, \Sigma) \ \ = \ \ -(1/2)\log(2\pi) \ - \ (1/2)\log|\Sigma| \ - \ (1/2)(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)$$

for which the Hessian is $-\Sigma^{-1}$.

Properties of Gaussian Approximation

- We don't need to know the normalizing constant for the posterior, and (as we'll see) we can get an approximation to this normalizing constant (the marginal likelihood) from the Gaussian approximation.

- We need to be able to find the mode. This is usually easier than any Monte Carlo scheme, assuming that the distribution has a single mode (or at least a dominant mode).

- We need to compute second derivatives of the log posterior density at the mode. This is hard for some models, and becomes unattractive for high dimensional problems.

- Once we have the Gaussian approximation, we can easily find simple moments and quantiles, and we can find the expectation of more complex functions by simple Monte Carlo.

- The adequacy of the approximation may depend on the choice of parameterization — eg, whether to use $\theta \in (0, \infty)$ or instead reparameterize in terms of $\phi = \log(\theta)$.

The Laplace Approximation

The Laplace method for approximating an integral is related to Gaussian approximation, but can be a bit more elaborate.

Suppose we want to approximate a definite integral over R^D such as

$$I \;=\; \int \exp(h(\theta))\, d\theta$$

We can do a Taylor expansion of $h(\theta)$ around the location of its maximum, θ^*:

$$h(\theta) \;=\; h(\theta^*) \;+\; (\theta-\theta^*)^T h'(\theta^*) \;+\; (1/2)(\theta-\theta^*)^T h''(\theta^*)(\theta-\theta^*) \;+\; \cdots$$

Dropping terms past second order, and noting that $h'(\theta^*) = 0$, we get the approximation

$$\begin{aligned} I &\approx \int \exp\Big(h(\theta^*) + (1/2)(\theta-\theta^*)^T h''(\theta^*)(\theta-\theta^*)\Big) d\theta \\ &= \exp(h(\theta^*)) \int \exp\Big(-\,(1/2)(\theta-\theta^*)^T[-h''(\theta^*)](\theta-\theta^*)\Big) d\theta \\ &= \exp(h(\theta^*))\,(2\pi)^{D/2}\,|-h''(\theta^*)|^{-1/2} \end{aligned}$$

Laplace Approximation for Bayesian Inference

When $h(\theta) = \log \text{prior} + \log \text{likelihood}$, the normalizing constant for the posterior will be

$$\int \exp(h(\theta))\, d\theta$$

which we have just seen can be approximated as

$$\exp(h(\theta^*))\, (2\pi)^{D/2}\, |{-h''(\theta^*)}|^{-1/2}$$

This is also the marginal likelihood for the model, if no terms were dropped from the log likelihood.

To approximate the posterior expectation of a positive function, $a(\theta)$, we write

$$E[a(\theta)|\text{data}] \;=\; \frac{\int a(\theta) \exp(h(\theta))\, d\theta}{\int \exp(h(\theta))\, d\theta} \;=\; \frac{\int \exp(g(\theta))\, d\theta}{\int \exp(h(\theta))\, d\theta}$$

where $g(\theta) = h(\theta) + \log(a(\theta))$. We have seen how to approximate the denominator. We approximate the numerator in the same way, finding $\theta^\dagger$ that maximizes $g(\theta)$ and then computing $g''(\theta^\dagger)$. The result is

$$E[a(\theta)|\text{data}] \;\approx\; \frac{\exp(g(\theta^\dagger))\, (2\pi)^{D/2}\, |{-g''(\theta^\dagger)}|^{-1/2}}{\exp(h(\theta^*))\, (2\pi)^{D/2}\, |{-h''(\theta^*)}|^{-1/2}} \;=\; \frac{\exp(g(\theta^\dagger))\, |{-g''(\theta^\dagger)}|^{-1/2}}{\exp(h(\theta^*))\, |{-h''(\theta^*)}|^{-1/2}}$$

Variational Approximations

Another approach to approximating a posterior distribution is to define a class of tractable distributions, $Q_\phi(\theta)$, and then find the value of ϕ that makes Q_ϕ best match the posterior, $P(\theta|x)$, where x is the observed data.

Suppose we choose Kullback-Liebler divergence as our measure of how well Q matches P:

$$KL(Q||P) \;\;=\;\; -\int Q(\theta)\,\log\!\left(\frac{P(\theta|x)}{Q(\theta)}\right)d\theta$$

Then this approximation will also give us a lower bound on the marginal likelihood, since

$$\begin{aligned}\log P(x) \;\;&=\;\; \int Q(\theta)\,\log(P(x))\,d\theta \;-\; KL(Q||P) \;+\; KL(Q||P) \\ &=\;\; \int Q(\theta)\,\log\!\left(\frac{P(\theta,x)}{Q(\theta)}\right)d\theta \;+\; KL(Q||P)\end{aligned}$$

Since $KL(Q||P)$ is non-negative, the first term above is a lower bound on $\log P(x)$. It will typically be computable, when Q is tractable (eg, by simple Monte Carlo, if there's no better way).

Types of Variational Approximations

Note that $KL(Q||P)$ involves an expectation over Q, which we assume is tractable. This is one reason variational approximations are sometimes feasible. There are many possible choices for the class of approximating distributions.

For example:

- We might let Q_ϕ be Gaussian, with $\phi = (\mu, \Sigma)$. This Gaussian approximation to $P(\theta|x)$ will not necessarily be the same as the Gaussian approximation based on curvature at the mode.
- For multidimensional θ, we might require that $Q_\phi(\theta) = \prod_j Q_{\phi_j}(\theta_j)$. This assumption alone may be enough to force all the Q_{ϕ_j} to be tractable (eg, in "variational EM" for mixture models), or we might need to limit the form of these distributions.

Sometimes further approximations are needed, but if they are also lower bounds, we can preserve the property of lower bounding the marginal likelihood.

Properties of Variational Approximations

Variational approximation based on minimizing $KL(Q||P)$ will select a distribution Q that avoids putting putting significant probability in places where P does not. In particular, if P is multimodal, Q will fit just one mode, rather than span multiple modes (and the low probability region between).

This is desirable for mixture models, where multiple equivalent modes exist, but not so desirable if we want to see the true uncertainty in other posterior distributions with multiple modes.

As the posterior distribution becomes more complex (eg, for neural network models), it becomes harder and harder to imagine a tractable class Q_ϕ that will contain an adequate approximation.