STA 414/2104

Statistical Methods for Machine Learning and Data Mining

Radford M. Neal, University of Toronto, 2014

Week 6

Bayesian Inference by Monte Carlo

Monte Carlo Methods

A very general approach to Bayesian computation — applicable even to very high-dimensional problems — is to obtain a *sample* of points from the posterior distribution, and use it to make *Monte Carlo* estimates.

A single sample point will contain values for all the unknown parameters, hyperparameters, latent variables, missing data values, etc. — everything not known, except what we don't care about (and isn't linked to other things) or that we have integrated away analytically.

We use this sample to approximate expected values by averages over the sample points. For example, from K values, $\theta^{(1)}, \ldots \theta^{(K)}$, for a parameter, sampled from $P(\theta \,|\, \text{data})$, we can approximate the predictive probability that $Y = 1$ by

$$\begin{aligned} P(Y = 1 \mid \text{data}) &= \int P(Y = 1 \mid \theta)\, P(\theta \,|\, \text{data})\, d\theta \\ &\approx \frac{1}{K} \sum_{k=1}^{K} P(Y = 1 \mid \theta^{(k)}) \end{aligned}$$

If the $\theta^{(k)}$ values are independent, the approximation converges to the true value as $K \to \infty$, by the Law of Large Numbers.

Monte Carlo with Independent Points

Monte Carlo is simplest when we can directly sample K *independent* points from the distribution of interest.

Let's denote the probability/density function of interest as $\pi(x)$, and suppose that we are interested in the expectation of some function $a(x)$. Note that x is typically high dimensional.

The Monte Carlo estimate based on K sample points, $x^{(1)}, \ldots, x^{(K)}$, will be

$$\overline{a} \;=\; \frac{1}{K} \sum_{k=1}^{K} a(x^{(k)})$$

If the variance of $a(x)$ is finite, we can get an indication of the accuracy of this estimate from its standard error — an estimate of the standard deviation of $\overline{a}$ in imaginary repetitions of the estimation procedure. This standard error is

$$\text{S.E. } \bar{a} \;=\; \sqrt{s_a^2/K}$$

where s_a^2 is the sample variance of a:

$$s_a^2 \;=\; \frac{1}{K-1} \sum_{k=1}^{K} \Big(a(x^{(k)}) - \overline{a} \Big)^2$$

Application: General Expectations for Conjugate Models

Efficient direct sampling of independent points from the posterior is usually possible only for models with conjugate priors. Typically, posterior means of parameters can be found analytically for such models, so Monte Carlo isn't necessary.

However, even for a conjugate model, the expectation of some complicated function of the parameters may be an integral that isn't analytically tractable. But a Monte Carlo estimate based on independent points can be found as long as the posterior can be efficiently sampled.

This is what really makes conjugate models tractable, even when the dimension of the parameter space is high.

Importance Sampling

When there is no efficient way to sample independently from $\pi(x)$ we can instead sample independently from some "similar" distribution, $\pi^*(x)$, and estimate the expectation of $a(x)$ by

$$\widehat{a}_{IS} \;=\; \frac{\displaystyle\sum_{k=1}^{K} a(x^{(k)})\,\frac{\pi(x^{(k)})}{\pi^*(x^{(k)})}}{\displaystyle\sum_{k=1}^{K} \frac{\pi(x^{(k)})}{\pi^*(x^{(k)})}}$$

Note that we don't need the normalizing constants for π or π^*, since they will cancel in the ratio above.

As long as $\pi^*(x) > 0$ for all x where $\pi(x) > 0$, this converges to the expectation of $a(x)$ under π as $K \to \infty$. We can see this since

$$\frac{1}{K}\sum_{k=1}^{K} \frac{\pi(x^{(k)})}{\pi^*(x^{(k)})} \;\to\; E_{\pi^*}\left[\frac{\pi(x)}{\pi^*(x)}\right] \;=\; \int \left[\frac{\pi(x)}{\pi^*(x)}\right] \pi^*(x)\, dx \;=\; 1$$

$$\frac{1}{K}\sum_{k=1}^{K} a(x^{(k)})\frac{\pi(x^{(k)})}{\pi^*(x^{(k)})} \;\to\; E_{\pi^*}\left[a(x)\,\frac{\pi(x)}{\pi^*(x)}\right] \;=\; E_\pi[a(x)]$$

Accuracy of Importance Sampling

Here's the importance sampling estimate again:

$$\widehat{a}_{IS} \;=\; \frac{\displaystyle\sum_{k=1}^{K} a(x^{(k)})\, \frac{\pi(x^{(k)})}{\pi^*(x^{(k)})}}{\displaystyle\sum_{k=1}^{K} \frac{\pi(x^{(k)})}{\pi^*(x^{(k)})}}$$

If we know the normalizing constants for π or π^*, we could omit the denominator, since it converges to one. But that estimator is often less accurate, so we probably shouldn't.

We can get a standard error for $\widehat{a}_{IS}$ by taking the square root of an estimate of its variance:

$$\mathrm{Var}(\widehat{a}_{IS}) \;\approx\; \frac{\displaystyle\sum_{k=1}^{K} \left(\frac{\pi(x^{(k)})}{\pi^*(x^{(k)})}\, (a(x^{(k)}) - \widehat{a}_{IS}) \right)^2}{\left[\displaystyle\sum_{k=1}^{K} \frac{\pi(x^{(k)})}{\pi^*(x^{(k)})} \right]^2}$$

This is discussed in my paper on "Annealed Importance Sampling".

Usefulness of Importance Sampling

The usefulness of importance sampling depends crucially on whether a good π^* can be found, that can be efficiently sampled, *and* leads to $\widehat{a}_{IS}$ being accurate.

Accuracy will be poor if $\pi^*(x)$ is very small in a region with non-negligible probability under π — then few points will be sampled from a region that actually is important to estimating $E_\pi(a(x))$.

Worse, it's possible that **no** points will be sampled from this region — then the estimate will be inaccurate, but the standard error obtained may not indicate that it is inaccurate.

But you can't just make π^* be very broad — then most points sampled will be wasted, with $\pi(x)$ being very small.

Direct use of importance sampling for Bayesian inference is usually practical only in moderate dimensions (eg, 10), and then only after significant fiddling to get a good π^* (eg, some heavy-tailed distribution located at the posterior mode).

Importance Sampling Using the Prior

A particularly simple way to do importance sampling for Bayesian inference is to sample from prior distribution for the parameters, having density $P(\theta)$.

The importance sampling estimate of the expectation of some function of the parameters, $a(\theta)$, with respect to the posterior distribution is then

$$\widehat{a}_{IS} \;=\; \frac{\displaystyle\sum_{k=1}^{K} a(\theta^{(k)})\, \frac{P(\theta^{(k)})P(\text{data}|\theta^{(k)})}{P(\theta^{(k)})}}{\displaystyle\sum_{k=1}^{K} \frac{P(\theta^{(k)})P(\text{data}|\theta^{(k)})}{P(\theta^{(k)})}} \;=\; \frac{\displaystyle\sum_{k=1}^{K} a(\theta^{(k)})\, P(\text{data}|\theta^{(k)})}{\displaystyle\sum_{k=1}^{K} P(\text{data}|\theta^{(k)})}$$

When making predictions for a new data point, y_*, this becomes

$$E[y_*|\text{data}] \;=\; \frac{\displaystyle\sum_{k=1}^{K} E[y_*|\theta]\, P(\text{data}|\theta^{(k)})}{\displaystyle\sum_{k=1}^{K} P(\text{data}|\theta^{(k)})}$$

The denominator above is an estimate of the marginal likelihood for the model.

However, this works well only if the factor by which the posterior is more concentrated than the prior isn't too large (few parameters, not too much data).

Obtaining a Sample by Simulating a Markov Chain

When the posterior distribution is too complex to sample from directly, and we can't find a good importance sampling distribution, we can instead simulate a *Markov chain* that will converge (asymptotically) to the posterior distribution.

States from the latter portion of this Markov chain will come (asymptotically) from the posterior distribution, but they will be *dependent*.

We can still use these states to make Monte Carlo estimates, but we need to adjust the standard error to account for the dependence.

Finding such a Markov chain sounds hard, but fortunately there are general schemes that make this possible even for difficult problems. This *Markov chain Monte Carlo (MCMC)* approach is therefore very general. MCMC can also be very slow in some circumstances, but despite this, it is often the only viable approach to Bayesian inference using complex models.