Markov chain Monte Carlo, CS589

Justin Domke

Fall 2017

1 Motivation

Recall our basic strategy for Bayesian inference:

- Use the data to "calculate" the posterior $Pr(M|\text{Data}) \propto Pr(M) \prod_{i=1}^{N} Pr(Y_i|X_i, M)$. (Note here that we are only computing $Pr(M|\text{Data})$ up to a constant that depends on Data, but not on M).
- Given a new point X', make predictions for Y' via

$$Pr(Y'|X', \text{Data}) = \sum_M Pr(M|\text{Data}) Pr(Y'|X', M).$$

In general, there are many models (often infinite so we actually do $\int_M$ rather than $\sum_M$), so we can't sum over them by brute force. What we do instead is draw a set of "sample" models

$$M^1, M^2, ..., M^T \sim Pr(M|\text{Data}).$$

Then, we can approximate the distribution over Y' by

$$Pr(Y'|X', \text{Data}) \approx \frac{1}{T} \sum_{t=1}^{T} Pr(Y'|X', M^t).$$

The question is, how to we sample from $Pr(M|\text{Data})$?

2 Sampling from a circle

How to do it?

Idea: rejection sampling.

Generate a bunch of points inside a square, "throw away" the ones not inside a circle.

Lesson: Given some object (even a very simple one like a circle), you need an <u>algorithm</u> to draw samples from it.

3 The Metropolis Algorithm

Suppose that we want to sample from some distribution

$$p(M).$$

Now, we assume that we actually can't compute p, but only $\hat{p}$, which is the same up to some constant

$$\hat{p}(M) = Z \times p(M).$$

This is very convenient for Bayesian inference, but we'll see later on why we can get away with this.

3.1 Some confusion about notation

There were many questions in class about the exact meaning of $p(M)$ and $\hat{p}(M)$. The answer– for the purposes of the metroplis algorithm– is that they could be any distribution. But if you are using Metropolis to sample from a Bayesian posterior, we would want to sample from

$$p(M) = \frac{Pr(M)Pr(\text{Data}|M)}{Pr(\text{Data})}.$$

Here, the specific meaning is

- $Pr(M)$ - The **prior**.
- $Pr(Data|M) = \prod_{i=1}^{N} Pr(Y_i|X_i, M)$ - The **likelihood**.
- $Pr(Data)$ - is the **marginal evidence**.

The marginal evidence is very difficult to calculate. Fortunately, we don't need to calculate it. Instead, we will use

$$\hat{p}(M) = Pr(M)Pr(\text{Data}|M).$$

Note that this isn't technically a distribution, since it isn't correctly normalized. The constant in this case, of course, is

$$Z = Pr(\text{Data}).$$

That's hard to calculate, but we don't need to calculate it, fortunately!

3.2 The actual algoriothm

Metropolis

- Initialize m^0
- For $t = 0, ..., T - 1$

  - $m' \sim q(m'|m^i)$
  - If rand() $\leq \hat{p}(m')/\hat{p}(m)$
    * $m^{t+1} = m'$
  - Else
    * $m^{t+1} = m^t$
- Return m^1, ..., m^T

Here, we need a "proposal distribution" that has the property that $q(m'|m) = q(m|m')$.

Before trying to understand this any further, let's try coding a simple example.

$$\hat{p}(m) = \exp\left(-\frac{1}{2}(m-1.5)^2\right) + \frac{1}{2}\exp\left(-\frac{1}{2}(m+1.5)^2\right)$$

This example is shown in Python code. (This is available on Moodle, for you to play around with.)

4 Why Metropolis Works

4.1 Detailed Balance

Detailed balance

$$p(m)Pr(m \to n) = p(n)Pr(n \to m).$$

Here, $Pr(m \to n)$ means the probability that we are in state n at time $t+1$ if we are in state m at time t.

Suppose that the distribution over states at time t is r^t. Then, the distibution at time $t+1$ will be

$$r^{t+1}(m) = \sum_n r^t(n)Pr(n \to m).$$

Now, suppose that detailed balance holds. Furthermore, suppose that $q^t(n) = \pi(n)$. Then, we have that

$$\begin{aligned}
r^{t+1}(m) &= \sum_n r^t(n)Pr(n \to m) \\
&= \sum_n p(n)Pr(n \to m) \\
&= \sum_n p(m)Pr(m \to n) \\
&= p(m)\sum_n Pr(m \to n) \\
&= p(m).
\end{aligned}$$

This says that if we've converged to the stationary distribution, we <u>stay</u> at the stationary distribution.

4.2 Why Metropolis Satisfies Detailed Balance

What is $Pr(m \to n)$ for the Metropolis algorithm? Assume that $m \neq n$ (otherwise it's obvious). Then,

$$
\begin{aligned}
Pr(m \to n) &= q(n|m) \min\left(1, \frac{\hat{p}(n)}{\hat{p}(m)}\right) \\
Pr(n \to m) &= q(m|n) \min\left(1, \frac{\hat{p}(m)}{\hat{p}(n)}\right)
\end{aligned}
$$

So what is the ratio of these two things?

$$
\frac{Pr(m \to n)}{Pr(n \to m)} = \frac{q(n|m) \min\left(1, \frac{\hat{p}(n)}{\hat{p}(m)}\right)}{q(m|n) \min\left(1, \frac{\hat{p}(m)}{\hat{p}(n)}\right)}
$$

Now, notice that since $q(n|m) = q(m|n)$, these cancel. Furthermore, note that either $\hat{p}(n)/\hat{p}(m)$ is less than one or $\hat{p}(m)/\hat{p}(n)$ is. In either case, it evaluates to the same thing! Thus,

$$
\frac{Pr(m \to n)}{Pr(n \to m)} = \frac{\hat{p}(n)}{\hat{p}(m)}.
$$

This is equivalent to the detailed balance condition.