10-715: Adv. Intro to Machine Learning **Fall 2015**

Recitation 3: November 10

Eric Wong

3.1 Markov Chain

Given n states $x_1, \ldots, x_n$, let T define the transition probability matrix between the states. In other words, $T_{ij} = T(j|i)$ is the probability of going from state i to state j.

We say π is a stationary distribution if

$$\pi(x') = \sum_x \pi(x) T_{x,x'} \Leftrightarrow \pi^T = \pi^T T$$

There is a huge field of finding stationary distributions of various chains. For our purposes, we are interested in the following sufficient condition:

Theorem 3.1 *If the probability of going from $x' \to x$ is equal to the probability of going from $x \to x'$, we say we have detailed balance. Equivalently:*

$$p(x')T(x|x') = p(x)T(x'|x)$$

If we have detailed balance, then a stationary distribution exists.

Why is this useful? You can think of a stationary distribution as a time-average number of steps spent in various states. This is because under certain conditions (irreducible positive recurrent Markov chains), there is a unique distribution given by these proportions.

Alternatively, it can be proved that the limiting distribution $\pi_j = \lim_{n\to\infty} T_{ij}^{(n)}$, when it exists, is the stationary distribution (note that $T^{(n)}$ gives the transition probability matrix after n steps). Under certain conditions, the limiting distribution can be proven to exist, however in general it does not always.

There is a whole field of study about Markov chains and these distributions. If you find this interesting, consider taking 15-857 Analytical Performance Modeling with Mor Harchol-Balter or see her textbook on queueing theory.

3.1.1 Markov Chain Monte Carlo (MCMC)

Suppose we wish to sample from some distribution. If the distribution is extremely complicated, this can be a non-trivial matter to do directly! Instead, we will use a technique known as MCMC to get around this by sampling from easier distributions.

Let π be the target distribution. The strategy of the algorithm is as follwos: We will construct a transition matrix T from an easier distribution that satisfies detailed balance, such that the resulting stationary distribution is our target distribution, π.

Then, by performing random walks on this Markov chain, we can "sample" from the stationary distribution (which will be our target distribution).

3.1.2 Metropolis-Hastings

One well known class of MCMC methods is the Metroplis Hastings algorithm. The main process of MH is quite simple:

1. Let $Q(x|x')$ be a function that proposes a move to state x from x'.
2. Accept the move with probability $A(x'|x)$. Otherwise stay.

The induced T is now $T(x \to x') = Q(x'|x)A(x'|x)$. As of this point, we have not determined Q or A yet. We will need to choose Q, A to satisfy detailed balance and achieve the target distribution.

To satisfy detailed balance, it must be that

$$\pi(x)Q(x'|x)A(x'|x) = \pi(x')Q(x|x')A(x|x')$$

We can satisfy this by setting $A(x'|x) = \min\left(1, \frac{\pi(x')Q(x|x')}{\pi(x)Q(x'|x)}\right)$. You can verify this quickly.

Note that $Q(x|x')$ has been chosen completely arbitrarily. The only requirements of the outlined process is that we are able to sample from Q, which should be an "easier" distribution than the original desired one.

- A reasonable choice for $Q(x|x')$ is to pick a Gaussian centered around x', so that closer states are more likely to be visited. The sequence of samples is then a random walk.
- In general, you will want to space out your samples far enough to avoid correlation between subsequent samples.
- You can start the algorithm from a random initial point. However you usually need a "burn-in" period to remove bias from your initial starting point and reach the equilibrium distribution. It is generally accepted to burn 1000 samples, but your results may vary. A chain that appears to have converged could actually just be lingering in some state space!

3.1.3 Gibbs Sampling

Gibbs sampling is yet another MCMC algorithm. Actually, it is a special case of the Metropolis Hastings algorithm that is particular suitable for graphical models. Let

$$Q(x_i'|x_i) = p(x_i'|\mathbf{x}_{-i})$$

where $\mathbf{x}_{-i}$ is the set of all variables not include x_i. Note that our transitions are going to only update one random variable at a time. Why do we do this? Efficiency!

1. Samples are ALWAYS accepted:

$$A(x_i'|x_i) = \min\left(1, \frac{\pi(\mathbf{x}')p(x_i|\mathbf{x}_{-i})}{\pi(\mathbf{x})p(x_i'|\mathbf{x}_{-i})}\right) = \min\left(1, \frac{p(x_i'|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i|\mathbf{x}_{-i})}{p(x_i|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i'|\mathbf{x}_{-i})}\right) = 1$$

2. The graphical model structure implies that $p(x_i|\mathbf{x}_{-i})$ only depends on the values in the Markov blanket of x_i.

You can either sweep across all samples and repeat for each iteration, or randomly pick random variables to update at each time step. You might also be able to sample in blocks, depending on the component structure of your graphical model.