EE698R : Advanced Topics in Machine Learning

Lecture 10.3

Gibbs Sampling and Metropolis Hastings Algorithm

Arunim Joarder (180141) and Ayush Jain (180173)

April 3, 2022

1 Convergence of Markov Chain

A Markov Chain will converge to a stationary distribution π_x if:

$$\pi_x(x') = \sum_x p_{(X^{t+1}|X^t)}(x'|x)\pi_x(x)$$
$$\pi_x(x') = \sum_x T(x \to x')\pi_x(x)$$

If the above equation is satisfied, this means that if the markov chain or the random walk is allowed to run for a sufficient amount of time, it will simulate the distribution π_x.

2 Gibbs Sampling

The Gibbs sampling algorithm is used when the target distribution $\pi(x)$ is difficult to sample from, but we sample using the transitional distribution $T(x \to x')$ such that the samples converge to the ones from $\pi(x)$.
The condition for convergence to a stationary distribution π_x is:
$\pi_x(x') = \sum_x p_{x_{t+1}|x_t}(x'|x)\pi_x(x)$
$\pi_x(x') = \sum_x T(x \to x')\pi_x(x)$
There are multiple transitional distributions that converge to the stationary distribution. In Gibbs Sampling, we update individual variables using their conditional distributions.

2.1 Algorithm

Algorithm 1 Gibbs Sampling

Start with any $x^0 = (x_1^0, x_2^0, x_3^0)$
for $k = 1, 2, 3, ...$ **do**
  Sample $x^{(k)}$ using the conditional distributions
  $x_1^k \sim p_{X_1|X_2,X_3}(x_1|x_2 = x_2^{k-1}, x_3 = x_3^{k-1})$
  $x_2^k \sim p_{X_2|X_1,X_3}(x_2|x_1 = x_1^k, x_3 = x_3^{k-1})$
  $x_3^k \sim p_{X_3|X_1,X_2}(x_3|x_1 = x_1^k, x_2 = x_2^k)$
  We obtain $x_k = (x_1^k, x_2^k, x_3^k)$
end for

The order in which the individual variables are updated does not matter.

2.2 Samples obtained from Gibbs Sampling follow the Markov chain stationary distribution equation

To Prove: $p(x', y', z') = \sum_{x,y,z} T(x, y, z \to x', y', z')p(x, y, z)$
Proof: In Gibbs Sampling, each individual variable gets updated using the univariate conditional

distribution. The transitional probabilities correspond to the product of conditional distributions of individual variables.
$T(x, y, z \to x', y', z') = p_{X|Y,Z}(x'|y, z)p_{Y|X,Z}(y'|x', z)p_{Z|X,Y}(z'|x', y')$
Substituting this,

$$\begin{aligned}
\text{RHS} &= \sum_{x,y,z} T(x, y, z \to x', y', z')p(x, y, z) \\
&= \sum_{x,y,z} p_{X|Y,Z}(x'|y, z)p_{Y|X,Z}(y'|x', z)p_{Z|X,Y}(z'|x', y')p(x, y, z) \\
&= \sum_{y,z} p(x'|y, z)p(y'|x', z)p(z'|x', y') \sum_{X} p(x, y, z) \\
&= \sum_{y,z} p(x'|y, z)p(y, z)p(y'|x', z)p(z'|x', y') \qquad (\text{as } \sum_{X} p(x, y, z) = p(y, z)) \\
&= \sum_{y,z} \Big(p(x', y, z)p(y'|x', z)\Big)p(z'|x', y') \qquad (\text{as } p(x'|y, z)p(y, z) = p(x', y, z)) \\
&= \sum_{z} \Big(p(x', z)p(y'|x', z)\Big)p(z'|x', y') \qquad (\text{as } \sum_{y} p(x', y, z) = p(x', z)) \\
&= \sum_{z} \Big(p(y', x', z)\Big)p(z'|x', y') \qquad (\text{as } \sum_{y} p(x', y, z) = p(x', z)) \\
&= p(y', x')p(z'|x', y') \qquad (\text{as } \sum_{z} p(y', x', z) = p(y', x')) \\
&= p(x', y', z')
\end{aligned}$$

LHS = RHS
Hence Proved.
So, By Gibbs Sampling we have found a sampling procedure that constructs a Markov Chain having stationary distribution π_x.

2.3 Example: 2-Dimensional Gaussian

We wish to sample from a 2-D Gaussian having a distribution

$$\mathbf{x} \sim \mathcal{N}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} |\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}) \tag{1}$$

We use the conditional distributions of x_1 and x_2, and perform Gibbs sampling to obtain the sample.

$$\begin{aligned}
p(x_1|x_2 = k) &= \mathcal{N}(\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(k - \mu_2), \bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \\
p(x_2|x_1 = k) &= \mathcal{N}(\bar{\mu} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(k - \mu_1), \bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})
\end{aligned}$$

How to verify that samples $\mathbf{x_s}$ follow the Normal Distribution $\mathcal{N}(x_1, x_2|\mu, \mathbf{\Sigma})$?

1. Estimate sample mean and sample variance and compare to the ones from the original distribution.
2. Construct a histogram to perform visual inspection.
3. Metrics like Maximum Likelihood, KL Divergence can be used.

2.4 Drawbacks of Gibbs Sampling

1. **Slow Convergence**.

   - Mixing: Roughly speaking, for any MCMC algorithms it's performance can be measured by how quick the sample statistics (mean) converge to the original distribution. It is also called mixing rate. Gibbs Sampling traverses the space very slowly and has a mixing issue.
   - It follows sequential sampling. There must be a order in which variables are updated. During update of variable x_k, it is necessary that $x_1, x_2, ..., x_{k-1}$ are updated. Hence, this cannot be parallelized and the algorithms is slow. A trick to boost the speed is to initiate several Markov chains running in parallel.

2. **High Correlation**
   In Gibbs sampling, updates in samples correspond to slow jumps or traversals in the sample space. Large jumps is generally not possible.
   Consider two consecutive samples: $\mathbf{x_s^t}$ and $\mathbf{x_s^{t+1}}$. These samples are very correlated.

3. **Deriving Conditional Distributions may not be possible.** It might be difficult to get analytical expressions for conditional distributions. Implementation of Gibbs Sampling will not be possible in this case.

As the dimension of $\mathbf{x}$ increases, the algorithms becomes slower and slower. The successive samples are also highly correlated.

3 Metropolis Hastings Algorithm

The Metropolis Hastings (MH) algorithm is used when a target distribution $\pi(x)$ is provided, which is easy to compute but difficult to sample from. The MH algorithm tries to estimate a transitional distribution $T(x \to x')$ such that its equilibrium spatial distribution becomes equivalent to $\pi(x)$.
The algorithm works by assuming a proposal transitional distribution $Q(x \to x')$ for changing states then and applying rejection sampling with an acceptance probability $A(x \to x')$ (or critic). So the following relations hold,

$$T(x \to x') = Q(x \to x')\, A(x \to x') \qquad \forall x' \neq x \tag{2}$$

$$T(x \to x) = Q(x \to x)\, A(x \to x) + \sum_{x' \neq x} \left(Q(x \to x')\, [1 - A(x \to x')]\right) \tag{3}$$

3.1 Algorithm

Algorithm 2 Metropolis Hastings Algorithm

for $k = 1, 2, 3, ...$ **do**
  Sample x' from proposal distribution $Q(x^k \to x')$
  if accepted **then** ▷ Accept with probability $A(x^k \to x')$
    $x^{k+1} \leftarrow x'$
  else
    $x^{k+1} \leftarrow x^k$
  end if
end for

3.2 Choosing A and Q

3.2.1 To find suitable $A(x \to x')$

We need to find a suitable form of $A(x \to x')$ following the constraint given by Eq.4, as proved earlier.

$$\pi(x') = \sum_x (\pi(x)T(x \to x')) \tag{4}$$

We claim that Eq.4 will automatically be satisfied if Eq.5 (also known as the Detailed Balance Principle) is satisfied.

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x) \tag{5}$$

Proof

$$\begin{aligned}
\sum_x (\pi(x)T(x \to x')) &= \sum_x (\pi(x')T(x' \to x)) && \text{(from Eq.5)} \\
&= \pi(x') \sum_x (T(x' \to x)) \\
&= \pi(x') && \text{(sum of probability over all possible } x \text{ gives 1)}
\end{aligned}$$

Thus we prove that Eq.4 is satisfied if Eq.5 is satisfied.
Now sustituting Eq.2 into Eq.5 we get,

$$\pi(x)Q(x \to x')A(x \to x') = \pi(x')Q(x' \to x)A(x' \to x) \tag{6}$$

$$\frac{A(x \to x')}{A(x' \to x)} = \frac{\pi(x')Q(x' \to x)}{\pi(x)Q(x \to x')} \equiv \rho \tag{7}$$

The value of ρ can easily be calculated once $Q(x \to x')$ has been chosen, and based on this ρ we can define $A(x \to x')$.

if $\rho < 1$ **then**
Let,
$A(x \to x') = \rho$
$A(x' \to x) = 1$

else if $\rho > 1$ **then**
Let,
$A(x \to x') = 1$
$A(x' \to x) = \frac{1}{\rho}$

As we only care about the value of $A(x \to x')$, we can write it as,

$$A(x \to x') = \min(1, \rho) \tag{8}$$

$$A(x \to x') = \min(1, \frac{\pi(x')Q(x' \to x)}{\pi(x)Q(x \to x')}) \tag{9}$$

This definition of the critic distribution in Eq.9 works in general cases and can be used anywhere.

3.2.2 Choosing $Q(x \to x')$

We can choose any form of $Q(x \to x')$ as long as it follows a few conditions.

1. $Q(x \to x') > 0 \forall x'$: This condition is necessary as it ensures reachability to any point x' given that we start from x. If $Q(x \to x')$ were zero for some pair (x, x'), it would be very hard to get to x' starting from x, leading to slower convergence.

2. Q should not be too narrow: A very narrow Q in the spatial dimensions would take a very long time to cover the entire region of interest as every x' generated from x will be very *close* to x, as Q is narrow.

3. Q should not be too wide: A very wide Q would higher rejection rate.

Further information regarding the Gibbs Sampling and Metropolis Hastings Algorithm can be found here [Bis06].

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.