CAP 6617 ADVANCED MACHINE LEARNING

NAME: SANKET KUMAR UF ID: 4513-3352

UNDERSTANDING THE METROPOLIS-HASTINGS ALGORITHM

AIM: A detailed, introductory exposition of the Metropolis-Hastings algorithm and a powerful Markov chain method to simulate multivariate distributions are given.

- **INTRODUCTION:** Metropolis-Hastings algorithm is a Markov Chain Monte Carlo(MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. These algorithms are generally used for sampling from multi-dimensional distributions, especially when the number of dimensions are high. The algorithm can be implemented using “Acceptance-Rejection” sampling and one “block-at-a-time”.

- **ACCEPTANCE-REJECTION SAMPLING:** The objective is to generate samples from the absolutely continuous target density. To obtain samples $X \sim f(x)$ where $f(x)$ is the probability density function for X, we sample a proposal function $X^* \sim h(x)$ and $U \sim Uniform(0,1)$. After sampling, we accept $X = X^*$ *if* $U <= f(X^*)/ch(X^*)$, otherwise return to sampling again. For a given proposal distribution $h(x)$, because the probability of acceptance is $1/c$, an optimal c is $sup_x f(x)/g(x)$ where c satisfies $cg(x) >= f(x)$ for all x. $f(x)$ only needs to be known up to a normalizing constant.

- **MARKOV CHAIN MONTE CARLO SIMULATION:** In MCMC methods, to generate samples from a target distribution $f(.)$, the methods find and utilize a transition kernel $P(x, dy)$ whose nth iterate converges to $f(.)$ for large n. The process is started at an arbitrary x and iterated a large number of times. After this large number, the distribution of observations generated from the simulation is approximately the target distribution.
  The basic idea is we perform a “random walk” through the probability distribution, favoring values with higher probabilities. The probability of a certain state being reached, depends only on the previous state of the chain. If we repeat this enough under the right conditions, we will hit every point in the space with a frequency proportional to its probability. After certain iterations, the system will converge to a point irrespective of the starting point.
  The problem is to find an appropriate transition kernel. Let’s say $P(x, dy) = p(x, y)dy + r(x)\delta_x(dy)$, where $p(x, x) = 0$ and δ_x is the membership of x in dy. If $p(x, y)$ satisfies the reversibility condition, $f(x)p(x, y) = f(y)p(y, x)$, then $f(.)$ is the invariant density of $P(X, .)$. This reversibility condition must be satisfied by $p(x, y)$. To find this $p(x, y)$ with this property, we use Metropolis-Hastings algorithm.

- **METROPOLIS-HASTINGS ALGORITHM:** We choose a proposal distribution $q(x, y)$, which states that when a process is at state x, the density generates a value y from $q(x, y)$. For most of the cases: $f(x)q(x, y) > f(y)q(y, x)$. This means the process moves from x to y too often and from y to x too rarely. Probability of move $a(x, y)<1$ is introduced to reduce the number of moves from x to y. Now the function becomes $p_{MH}(x, y) \equiv q(x, y)a(x, y)$ for $x \neq y$. $a(x, y)$ is determined by requiring that $p_{MH}(x, y)$ satisfies the reversibility condition, which gives $a(x, y) = min[f(y)q(y, x)/f(x)q(x, y), 1]$. If a candidate value is rejected, the current value is taken as the next item in the sequence. For calculating $a(x, y)$, there is no need of knowing the normalizing constant for $f(x)$ because it appears both in numerator and denominator. If the proposal function is symmetric, $a(x, y)$ reduces to $f(y)/f(x)$ because $q(x, y) = q(y, x)$. If the jump goes “uphill”, the change is accepted, otherwise it is accepted with a non-zero probability. The regularity conditions for convergence to occur are and the process to be ergodic are: 1) **Irreducible** – for every state, there is a non-zero probability of moving to any other state. 2) **Aperiodic** – the process must not get trapped in cycles.

- To implement M-H algorithm, it is necessary that a suitable proposal function be specified. Typically, this function is selected from a family of distributions that requires the specification of such tuning parameters as the location and scale.