An introduction to MCMC in Machine Learning

Peadar Coyle

October 14, 2012

Contents

1 Motivation **3**

2 Monte Carlo principle **4**

3 Association learning **4**

1 Motivation

MCMC techniques have been applied to solve integration and optimisation problems in large dimensional spaces. These two types of problem play a fundamental role in machine learning, physics, statistics, econometrics and decision analysis. The following are just some examples.

1. *Bayesian inference and learning.* Given some unknown variables $x \in \mathcal{X}$ and data $y \in \mathcal{Y}$, the following typical intractable problems are central to Bayesian statistics

   - Normalisation. To obtain the posterior $p(x|y)$ given the prior $p(x)$ and likelihood $p(y|x)$, the normalising factor in Bayes' theorem needs to be computed

   $$p(x|y) = \frac{p(y|x)p(x)}{\int_{\mathcal{X}} p(y|x')p(x')dx'}$$

   - Marginalisation. Given the joint posterior of (x,z) $\in \mathcal{X} \times \mathcal{Z}$, we may often be interested in the marginal posterior

   $$p(x|y) = \int_{\mathcal{Z}} p(x, z|y)dz.$$

   - Expectation. The objective of the analysis is often to obtain summary statistics of the form

   $$\mathbb{E}_{p(x|y)}(f(x)) = \int_{\mathcal{X}} f(x)p(x|y)dx \qquad (1)$$

   form some function of interest $f : \mathcal{X} \to \mathbb{R}^{n_f}$ integrable with respect to $p(x|y)$. Examples of appropriate functions include the conditional mean, in which case f(x) $=$ x, or the conditional covariance of x where f(x) $= xx' - \mathbb{E}_{p(x|y)}(x)\mathbb{E}'_{p(x|y)}(x)$

2. Statistical mechanics. Here, one needs to compute the partition function Z of a system with states s and Hamiltonian E(s)

   $$Z = \sum_s \exp\left[-\frac{E(s)}{kT}\right] \qquad (2)$$

   where k is the Boltzmann's constant and T denotes the temperature of the system. Summing over the large number of possible configurations is prohibitively expensive. Note that problems of computing the partition function and the normalising constant in statistical inference are analogous.

2 Monte Carlo principle

The idea of Monte Carlo simulation is to draw an i.i.d. set of samples $\{x^{(i)}\}_{i=1}^{N}$ from a target density p(x) defined on a high-dimensional space $\mathcal{X}$ (e.g. the set of possible configurations of a system, the space on which the posterior is defined, or the combinatorial set of feasible solutions). These N samples can be used to approximate the target density with the following empirical point-mass function

$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x^{(i)}}(x).$$

where $\delta_{x^{(i)}}(x)$ denotes the Dirac delta mass at $x^{(i)}$. Consequently, one can approximate the integrals (or very large sums) I(f) with tractable sums $I_N(f)$ that converge as follows

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \xrightarrow[N\to\infty]{a.s.} I(f) = \int_{\mathcal{X}} f(x)p(x)dx$$

That is, the eestimate $I_N(f)$ is unbiased and by the strong law of large numbers, it will almost surely (a.s.) converge to I(f). If the variance (in the univariate case for simplicity) of f(x) satisfies $\sigma_f^2 \overset{\triangle}{=} \mathbb{E}_{p(x)}(f^2(x)) - I^2(f) < \infty$, the the variance of the estimator $I_N(f)$ is equal to $\text{var}(I_N(f)) = \frac{\delta_f^2}{N}$ and a central limit theorem yields convergence in distribution of the error

$$\sqrt{N}(I_N(f) - I(f))\ N \overset{\Longrightarrow}{\to} \infty\ \mathcal{N}(0, \sigma_f^2)$$

where $\Longrightarrow$ denotes convergence in distribution. The avantage of Monte Carlo integration over deterministic integration arises in the fact that former positions the integration grind

3 Association learning

Definition 3.1. (Support) Support is defined on itemset Z $\subset$I as the proportion of transactions in which all items in Z are found together in the database:

$$supp(Z) = \frac{freq(Z)}{|D|}$$

where freq(Z) denotes the frequency of itemset Z (number of transactions in which Z occurs) in database D, and |D| is the number of transactions in the database.