Monte_Carlo

June 1, 2023

1 STAT 207: Monte Carlo Methods

Monte Carlo methods are computational techniques used to solve problems and make probabilistic predictions by generating random samples. These methods rely on the principle of statistical sampling to approximate solutions or estimate quantities that are difficult or intractable to compute analytically.

The name "Monte Carlo" is derived from the famous Monte Carlo Casino in Monaco, known for its games of chance and randomness.

- Integration
- Simulation
- Optimization
- MCMC

1.1 Monte Carlo Integration

- Monte Carlo integration is a rough and ready technique for calculating highdimensional integrals and dealing with nonsmooth integrands.
- Monte Carlo techniques ignore smoothness and substitute random points for fixed quadrature points.
- To approximate the integral

$$E[f(X)] = \int f(x)\, d\mu(x),$$

  we take i.i.d. samples $X_1, ..., X_n$ from μ and estimate the integral by the sample average $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$.
- Such Monte Carlo estimates converge to $E[f(X)]$ as $n \to \infty$.
- If $f(x)$ is square integrable, by CLT, the estimator is approximately normally distributed around $E[f(X)]$ with standard deviation $\sqrt{\mathrm{Var}[f(X)]/n}$, which can be estimated as $\sqrt{v/n}$, with

$$v = \frac{1}{n-1}\sum_{i=1}^{n}\left[f(X_i) - \frac{1}{n}\sum_{j=1}^{n} f(X_j)\right]^2.$$

- CLT tells us:

  - the Monte Carlo error estimate does not depend directly on the dimensionality of the underlying space;
  - the error declines at the slow rate of $n^{-1/2}$.

1.2 Importance Sampling

- Importance sampling is one technique for variance reduction.
- Let $d\mu(x) = g(x)d\nu(x)$
- Find $h(x)$ (importance distribution) relative to ν with $h(x) > 0$ when $f(x)g(x) \neq 0$. Then we have

$$\int f(x)\,d\mu(x) = \int f(x)g(x)\,d\nu(x) = \int \frac{f(x)g(x)}{h(x)} h(x)\,d\nu(x).$$

- Sample $Y_1, \dots, Y_n$ i.i.d. from $h(x)$, then the sample average

$$\frac{1}{n}\sum_{i=1}^{n} \frac{f(Y_i)g(Y_i)}{h(Y_i)} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)w(Y_i)$$

  offers an alternative unbiased estimator of $\int f(x)\,d\mu(x)$.

- The ratios $w(Y_i) = \frac{g(Y_i)}{h(Y_i)}$ are called **importance weights**.
- The weighted estimator has smaller variance than the naive estimator $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ if and only if the second moments of the two sampling distributions satisfy

$$\int \left[\frac{f(x)g(x)}{h(x)}\right]^2 h(x)\,d\nu(x) \leq \int f(x)^2 g(x)\,d\nu(x).$$

- If we choose $h(x) = \frac{|f(x)|g(x)}{\int |f(z)|g(z)\,d\mu(z)}$, then the Cauchy-Schwarz inequality implies

$$\begin{aligned}\int \left[\frac{f(x)g(x)}{h(x)}\right]^2 h(x)\,d\nu(x) &= \left[\int |f(x)|g(x)\,d\nu(x)\right]^2 \\ &\leq \int f(x)^2 g(x)\,d\nu(x) \int g(x)\,d\nu(x) \\ &= \int f(x)^2 g(x)\,d\nu(x).\end{aligned}$$

- One problem is that $\int |f(x)|g(x)\,d\nu(x)$ is unknown. How to choose $h(x)$?
- In practice, one or both of the densities $g(x)$ and $h(x)$ may only be known up to an unspecified constant.
- The fix:

$$\frac{\sum_{i=1}^{n} f(Y_i)w(Y_i)}{\sum_{i=1}^{n} w(Y_i)} = \frac{\frac{1}{n}\sum_{i=1}^{n} f(X_i)cw(Y_i)}{\frac{1}{n}\sum_{i=1}^{n} cw(Y_i)}.$$

1.2.1 Example: Binomial Tail Probabilities

- $X \sim Binom(m, p)$, $P(X \geq x)$ is very small if x is much larger than mp.
- Consider importance sampling with random binomial deviates with the same number of trials m but a higher success probability q.
- Consider the following functions:

$$\begin{aligned} f(y) =& 1_{\{y=x\}} \\ g(y) =& \binom{m}{y} p^y (1-p)^{m-y} \\ h(y) =& \binom{m}{y} q^y (1-q)^{m-y} \\ w(y) =& \left(\frac{p}{q}\right)^y \left(\frac{1-p}{1-q}\right)^{m-y}. \end{aligned}$$

- Choose q so that $mq = x$.

1.2.2 Example: Expected Returns to the Origin

- In a three-dimensional, symmetric random walk on the integer lattice, the expected number of returns to the origin equals

$$\frac{1}{2^3} \int_{-1}^{1} \int_{-1}^{1} \int_{-1}^{1} \frac{3}{3 - \cos(\pi x_1) - \cos(\pi x_2) - \cos(\pi x_3)} \, dx_1 \, dx_2 \, dx_3.$$

- A crude Monte Carlo estimate based on 10,000 uniform deviates from the cube $[-1, 1]^3$ is 1.478 ± 0.036.
- The singularity of the integrand at the origin 0 explains the inaccuracy and implies that the estimator has infinite variance.
- Thus, the standard error 0.036 attached to the estimate 1.478 is bogus.
- Let $S_3 = \{(x_1, x_2, x_3) : r = \sqrt{x_1^2 + x_2^2 + x_3^2} \leq 1\}$ be the unit sphere in R^3.

1.3 Finite-State Markov Chains

- Markov chains are one of the richest sources of good models for capturing dynamical behavior with a large stochastic component.
  - discrete-time and continuous-time chains
  - hidden Markov chains

1.4 Discrete-Time Markov Chains

- Transition probability matrix $P = (p_i j)$, with $p_i j = P(Z_n = j | Z_{n-1} = i)$ and $\sum_j p_i j = 1$.
- Markov property:

- P^n provides the n-step transition probabilities $p_ij^{(n)}$.
- Does P^n converge?
- Solve $\pi = \pi P$, the equilibrium (or stationary) distribution of the chain.
- Two ergodic conditions for uniqueness of the problem:
  - Aperiodicity: for any state in the chain, the chain can return to that state after a certain number of steps without following any fixed pattern or cycle.
  - Irreducibility: every state is reachable from every other state. All states communicate.
- Merge into a single assumption: for some n, P^n has all positive entries.
- Under the ergodic condtions, we have the **ergodic theorem**. Let $f(z)$ be some function defined on the states of an ergodic chain. Then $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(Z_i)$ exists and equals the theoretical mean $E_\pi[f(Z)] = \sum_z \pi_z f(z)$ of $f(Z)$ under the equilibrium distribution π.
- The equilibrium condition $\pi = \pi P$,

  $$\pi_j = \sum_i \pi_i p_{ij},$$

  has a stronger version

  $$\pi_j p_{ji} = \pi_i p_{ij}, \quad \forall (i, j).$$

  This is called the **detailed balance**.

1.4.1 Example: Random Walk on a Graph

- Consider a connected graph with vertex set V and edge set E.
- The number of edges $d(v)$ incident on a given vertex v is called the degree of v.
- Define the transition probability matrix $P = (p_{uv})$ by

  $$p_{uv} = \begin{cases} \frac{1}{d(u)} & \text{for } \{u, v\} \in E \\ 0 & \text{for } \{u, v\} \notin E. \end{cases}$$

- If V has m edges, then the equilibrium distribution of the chain has components $\pi_v = d(v)/(2m)$.

1.5 Hidden Markov Chains

- Hidden Markov chains incorporate both observed data and missing data.
- The missing data are the sequence of states visited by a Markov chain, $Z_1, ..., Z_n$.
- The observed data provide partial information about this sequence of states, $Y_i = y_i$.
- The likelihood of the observed data:

  $$P = Pr(Y_1 = y_1, ..., Y_n = y_n)$$

- It can be constructed from 3 components:

- (a) the initial distribution π at the first epoch of the chain,
- (b) the epoch-dependent transition probabilities $p_{ijk} = Pr(Z_{i+1} = k|Z_i = j)$, and
- (c) the conditional densities $\phi_i(y_i|j) = Pr(Y_i = y_i|Z_i = j)$.

- Baum's forward and backward algorithms to evaluate P and its partial derivatives.
- The Viterbi algorithm solves the problem of finding a most probable sequence of states of the hidden Markov chain given the observed data by dynamic programming.

1.5.1 Connections to the EM Algorithm

- Baum's algorithms can be integrated in the E step of the EM algorithm to solve the MLE of HMM.
- Denote X_{ij} the missing indicator of the chain taking state j at epoch i.
- The complete data loglikelihood is

$$L(\theta) = \sum_j X_{1j} \ln \pi_j + \sum_{i=1}^n \sum_j X_{ij} \ln \phi_i(Y_i|j) + \sum_{i=1}^n \sum_j \sum_k X_{ij} X_{i+1,k} \ln p_{ijk}.$$

...