Sampling methods and MCMC

Course of Machine Learning
Master Degree in Computer Science

University of Rome "Tor Vergata"

a.a. 2020-2021

Giorgio Gambosi

General issue

How can we sample from any distribution, especially if we do not have an analytical representation of it?

Sampling: easy case

Assume we know $p(x)$: we wish to find a function $f(z)$ such that if $z \sim U(0,1)$, then $f(z) \sim p(z)$.

- This is equivalent to saying that for each $z \in [0,1]$ the cumulative probability of z wrt to the uniform distribution, which is $P_U(\zeta \leq z) = z$ itself, should be equal to the cumulative probability of $f(z)$ wrt $p(z)$, that is $P_p(\zeta \leq f(z))$
- that is,

$$z = \int_0^z d\zeta = \int_{-\infty}^{f(z)} p(\zeta) d\zeta = P_p(f(z))$$

  as a corollary, since it results $z = P_p(f(z))$, we have that $f(z) = P_p^{-1}(z)$

Sampling in the easy case: an example

Example

Given $\mathcal{R}$, we use it to produce exponentially distributed values, that is values distributed according to

$$p(x|\lambda) = \lambda e^{-\lambda x}, 0 \leq x < \infty$$

The exponential cumulative distribution is

$$P(x) = \int_0^x \lambda e^{-\lambda \xi} d\xi = 1 - e^{-\lambda x}$$

by setting $z = P(f(z)) = 1 - e^{-\lambda f(z)}$ we get

$$\begin{aligned} e^{-\lambda f(z)} &= 1 - z \\ -\lambda f(z) &= \ln(1 - z) \\ f(z) &= -\frac{1}{\lambda} \ln(1 - z) \end{aligned}$$

Sampling on general distributions

Approaches

Applying the above method is possible only for simple distributions. In most cases, you cannot immediately derive values distributed according to $p(x)$

Many sampling methods have been introduced

- rejection sampling
- importance sampling
- adaptive rejection sampling
- sampling-importance-sampling
- ...

Markov chains

Definition

Given a (possibly infinite) sequence of random variables $\mathbf{X} = (X_0, X_1, \ldots)$ and a state space $\mathcal{X}$ of possible values for all $X_i \in \mathbf{X}$, a Markov chain on $\mathbf{X}$ is a stochastic process which defines for each ordered pair $< x_i, x_j > \in \mathcal{X}^2$, a probability p_{ij} of transition from x_i to x_j such that $p(X_t = x_i | X_{t-1} = x_j) = p_{ij}$, for all $t > 0$.

State probability

Given an initial state, that is a vaue assigned to X_0, the distribution $p(X_t = x_j | X_0 = x_k)$ of each random variable X_t on the set of state can be easily obtained (by matrix multiplication).

Markov chains

Stationary distribution

Under suitable conditions on its structure, a Markov chain is ergodic, that is the probability $p(X_t = x_j | X_0 = x_k)$, as $n \to \infty$,

- is independent from the initial state

$$p(X_t = x_j | X_0 = x_k) = p(X_t = x_j)$$

- is stationary

$$p(X_t = x_j) = p(X_{t+1} = x_j)$$

Markov chain Montecarlo (MCMC)

Idea

Given a hard to sample distribution $p(x)$, derive an ergodic Markov chain such that:

- its transition probability $q(x_i|x_{i-1})$ is easy to sample
- its stationary distribution is $p(x)$

Markov chain Montecarlo (MCMC)

Ho to use it?

Given the Markov chain,

- a sequence of random transitions is performed, starting from any initial state (value of x).
- apply a certain number of initial transitions (burn-in time)
- after that, at each step the value $\overline{x}$ reached by the MC is tested wrt a predefined criterion: if the test is positive, the value is returned

The returned values are (approximately) distributed as $p(x)$: hence their sequence can be used as a sequence of samplings from $p(x)$

MCMC methods

Several MCMC methods have been defined, differing each other by the structure of the chain and the acceptance criterion applied.

Metropolis algorithm

Idea

After the burning time, let $x^{(i-1)}$ be the current state and let $\overline{x}$ be the value produced by a random transition from $x^{(i-1)}$, obtained by sampling $q(x|x^{(i-1)})$

$\overline{x}$ is accepted, and returned as a sample, with probability

$$A(\overline{x}, x^{(i-1)}) = \min\left(1, \frac{p(\overline{x})}{p(x^{(i-1)})}\right)$$

Notice that if $\overline{x}$ has higher probability than $x^{(i-1)}$ with respect to the target distribution $p(x)$, it is accepted, while if its probability is smaller, it is accepted with probability equal to the ratio between them.

If $\overline{x}$ is accepted, then $x^i = \overline{x}$ becomes the current state, otherwise the current state is not modified, that is $x^{(i)} = x^{(i-1)}$)

Note

Observe that the same holds if $\pi(x) = Kp(x)$ is applied in the definition

Metropolis-Hastings algorithm

Idea

- Applied for non symmetric $q(x|x')$
- In this case,

$$A'(x, x') = \min\left(1, \frac{p(x)q(x'|x)}{p(x')q(x|x')}\right)$$

Gibbs sampling

Use

Gibbs sampling is a MCMC applied in cases when:

- x has dimensionality at least 2, $\mathbf{x} = (x_1, \ldots, x_m)$, with $m > 1$
- for all $i = 1, \ldots, m$, the conditional distribution $p(x_i|\mathbf{x}_{-i})$ is easy to sample, where $\mathbf{x}_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m\}$

Idea

Instead of sampling the next state in a single step from $q(\mathbf{x}|\mathbf{x}')$, a sequence of m transitions is sampled, each wrt a component x_i of $\mathbf{x}$ and to distribution $p(x_i|\mathbf{x}_{-i})$.

The basic idea in Gibbs sampling is that rather than probabilistically picking the next state of all at once, a separate probabilistic choice is performed for each of the m dimensions, with each choice depending on the other $k - 1$ dimensions.

MCMC and bayesian models

- MCMC can be applied (as it frequently happens) in bayesian inference by observing that the posterior distribution is defined as

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\theta)p(\theta)}{Z}$$

  where Z is usually hard to compute
- Let us remind that MCMC is able to sample a distribution $p(\mathbf{x})$ assuming that a proportional function $\pi(\mathbf{x}) = Kp(\mathbf{x})$ can be evaluated, for any unknown K
- Thus, samples of the posterior distribution of parameters can be obtained if both the prior $p(\theta)$ and the likelihood $p(\mathbf{X}|\theta) = \prod_i p(\mathbf{x}_i|\theta)$ can be evaluated for any value θ

Sampling the evidence

- Actually, the evidence

$$p(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta$$

could be explicitly evaluated, if necessary, as the average of a set of m values

$$p(\mathbf{X}|\theta_i) \qquad i = 1, \ldots, m$$

computed from the set of samples $\theta_1, \ldots, \theta_m$ of $p(\theta)$

Sampling the predictive distribution

- For what regards the predictive distribution

$$p(\mathbf{x}|\mathbf{X}) = \int_{\theta} p(\mathbf{x}|\theta)p(\theta|\mathbf{X})d\theta$$

the same considerations apply, averaging the set of values

$$p(\mathbf{x}|\theta_i) \qquad i = 1, \ldots, m$$

computed from the set of samples $\theta_1, \ldots, \theta_m$ of the posterior distribution $p(\theta|\mathbf{X})$