10-708: Probabilistic Graphical Models, Spring 2015

16 : Monte Carlo Methods

Lecturer: Eric P. Xing *Scribes: Aaron Q Li, Jay-Yoon Lee*

1 Representation

Many distributions can be represented in closed form, example:

$$f(x) = \frac{1}{(2\pi|\Sigma|)^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

And the expectation $E_p[f(x)]$ is easy to derive.

But there are even more distributions that cannot be expressed in closed form.

2 Monte Carlo methods

Concept: Draw random samples from a distribution and compute marginals and expectations using sample-based average. e.g:

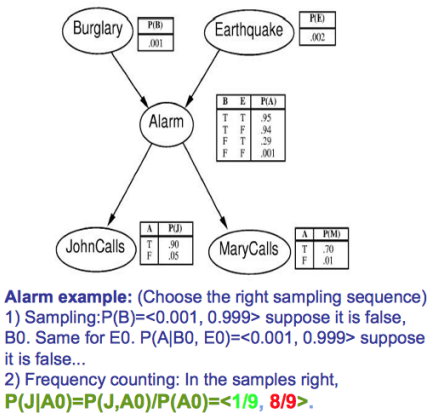
$$E[f(x)] = \frac{1}{N}\sum_i f(x)^{(i)}$$

This works with arbitrary model and is asymptotically exact. However, there are three challenges:

- How to draw samples from a given dist (not all distributions can be trivially sampled).
- How to make better use of the samples (not all sample are useful, or eqally useful).
- How to know when we have enough sample for estimation.

Example:

Naive Sampling: In the lecture slide, there is an example is provided using samples generated from a Bayesian network where variable values are discrete. We can constrcut the samples according to probabilities given in a BN by traversing through the tree while using the truth table on each node. The probability estimation can be done using simple frequency counting. However, problems arise when some value of some variable is not present in the pool of samples, thus giving unknown (or zero, according to the frequency counting



E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

Figure 1: Example of Monte Carlo method: Naive Sampling

paradigm) probabilities of some marginal and conditional distribution. This implies we need more samples to accurately estimate a probability. However, in many situations the number of samples required is exponential in numbers. Furthermore, the naive sampling method would be too costly for high dimensional cases, and therefore we introduce other alternatives from rejection sampling to weighted resampling.

3 Rejection Sampling

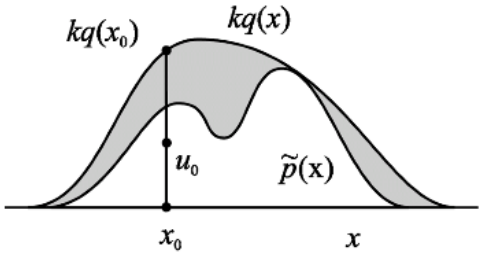


Figure 2: Illustration of the Rejecton Sampling

Suppose we want to sample from a distribution $\Pi(x) = \frac{1}{Z}\Pi'(x)$, where Z is an unknown normaliser. In many situations, $\Pi(x)$ is difficult to sample but $\Pi'(x)$ is easy to evaluate. Rejection sampling provides a simple procedure to make use of this property and sample from the original distribution by utilising a simpler distribution, $Q(x)$. The procedure is as the following:

$$\begin{aligned} \text{Draw} \quad & x' \sim Q(x) \\ \text{Accept with proportion to} \quad & \frac{\Pi(x')}{kQ(x')} \end{aligned}$$

Where k is a constant chosen in a way so to ensure $kQ(x) \geq \pi(x)$. The correctness of this procedure can be easily shown by using Bayesian inference as described below.

$$\begin{aligned} p(x) &= \frac{[\Pi'(x)/kQ(x)]Q(x)}{\int [\Pi'(x)/kQ(x)]Q(x)dx} \\ &= \frac{\Pi'(x)}{\int \Pi'(x)dx} = \Pi(x). \end{aligned}$$

However, as one can visualise, if the proposal distribution does not match the shape of original distribution closely, this procedure can yield a high number of rejected samples, thus making it inefficient.

Moreover, the value of k has to be determined before the procedure can start. This is not possible in many situations.

4 Importance Sampling

Instead of formulating a distribution with a density function everywhere greater or equal to original distribution, and subsequently rejecting samples, importance sampling uses a weighting scheme on the generated samples. Suppose we generate M samples from the proposal distribution q, and assume $\pi(x^{(t)})$ can be evaluated (**note**: this is non-trivial assumption). Then the mean of arbitrary function $f(x)$ can be computed by

$$E[f(x)] = \frac{1}{M}\sum_{i=1}^{M} f(x^{(i)})w^{(i)}$$

Where $w^{(t)} = \frac{\pi(x^{(t)})}{q(x^{(t)})}$ and $x^{(t)}$ are the samples from q.

In the case which $\pi(x)$ cannot be evaluated directly. Assume $p'(x) = \alpha p(x)$, where α is an unknown normaliser. Let $r(x) = \frac{\pi'(x)}{q(x)}$, it follows that

$$E_q[r(x)] = \int \frac{\pi'(x)}{q(x)} q(x)dx = \alpha$$

Now it can be shown that (see slides)

$$E_p[f(x)] = \frac{1}{M}\sum_{i=1}^{M} f(x^{(i)})r(x^{(t)}) / \sum_i r(x^{(t)})$$

In other words, the weights $\tilde{w}^{(t)}$ is now determined by $r(x^{(t)})/\sum_i r(x^{(t)})$, and $E[f(x)] = \frac{1}{M}\sum_{i=1}^{M} f(x^{(i)})\tilde{w}^{(i)}$.

5 Weighted Resampling

Note even though importance sampling resolves one issue in rejection sample, that the value of k has to be determined a priori, the efficiency of the procudure still relies on how well the proposal distribution q matches the original distribution π. To see this, suppose the original distribution has a substantial mass concenrtrated in a small region, and the proposal distribution does not. Following the procedure of importance sampling, it is not hard to see that samples generated in this region is going to have a much higher weight than samples in other regions. Thus, the estimated result is likely to contain large error.

There are two solutions to this issue. One is to use a heavy tail proposal distribution, the other is to resample from the samples according to the weights. However, this means a high number of samples is required to give reasonable estimate. The textbook (Mackay 29.2) shows an example which the weight difference could result requiring exponential number of samples in order to give any reasonable estimate of $E_p[f(x)]$.

6 Particle Filter

Particle Filter is a tool to estimate the hidden state x_t (at time t) of a hidden markov model with known transition probability $\pi(x_{t+1}|x_t)$, and the expectation value over a function at state x_t, approximated by

$$\int f(x_t)p(x_t|y_{0:t})dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_t^{(i)})$$

Where $y_{0:t}$ are the observations from time 0 to t, and $x_t^{(i)}$ with $i = 1, ..., N$ are the samples. Depending on the proposal distribution, there are many ways to sample each x_t and update the weights (see Wikipedia, for example). The one introduced in class uses the following mechanism: following the same derivation in importance sampling, it can be shown that the quantity $p(x_t|y_{1:t})$ can be represented by

$$\begin{aligned} x_t^{(m)} &\sim p(x_t|y_{1:t-1}) \\ w_t^{(m)} &= \frac{p(y_t|x_t^{(m)})}{\sum_m p(y_t|x_t^{(m)})} \end{aligned}$$

Where $m = 1, ..., M$ are the samples and $w_t^{(m)}$ are the weights. We draw x_{t+1} according to the following distribution:

$$\begin{aligned} p(x_{t+1}|y_{1:t}) &= \int p(x_{t+1}|x_t)p(x_t|y_{1:t})dx_t \\ &= \sum_{m=1}^{M} w_t^{(i)} p(x_{t+1}|x_t^{(m)}) \end{aligned}$$

And repeat the previous process of drawing M samples for timestamp $t+1$. i.e

$$\begin{aligned} x_{t+1}^{(m)} &\sim p(x_{t+1}|y_{1:t}) \\ w_{t+1}^{(m)} &= \frac{p(y_{t+1}|x_{t+1}^{(m)})}{\sum_m p(y_{t+1}|x_{t+1}^{(m)})} \end{aligned}$$

(To be continued)