Algorithms for Graphical Models (AGM)

Gibbs sampling

$Date: 2008/10/21 09:52:02 $

AGM-13

In this lecture

- Markov chain Monte Carlo
- Gibbs sampling

Limitations of importance sampling

- If evidence is improbable we end up with very low weighted samples.

- The bottom line is that importance sampling becomes less useful the further the proposal distribution is from the target distribution.

- We want the instantiated variables to have a more direct effect on what gets sampled (for both their descendant and non-descendant nodes).

Markov chain Monte Carlo

- *Markov chain Monte Carlo*: Instead of sampling from the target distribution, sample from a *sequence of distributions* which gets progressively closer to the target distribution.

- If we do this for long enough we will end up sampling from a distribution very close to the target distribution.

- We now have two dimensions of approximation: the distributions from which we sample only approximate the target distribution and, (as with all sampling) any sample only provides an approximate picture of the distribution from which it is sampled.

Why 'Monte Carlo'?

- Any simulation-based and thus approximate computational method is a *Monte Carlo* method.

- They have become popular in statistics with the advent of cheap, powerful computers.

- Named after the casino

(Finite) Markov chains

A *finite (homogenous) Markov chain* is a linear infinite BN:

$$X_0 \to X_1 \to X_2 \to \ldots$$

where

1. each X_i has the same *finite* set of values; and

2. each CPT $P(X_{i+1}|X_i)$ is the same.

Markov chain intuitions

- The values of the X_i are best thought of as *states* of some dynamic system.

- Sampling from a Markov chain thus corresponds to one possible evolution of such a system.

- For example, if the states are possible locations of an object, a 'run' of the Markov chain corresponds to the object moving probabilistically—where the next move only depends on the current position.

Markov chain terminology

- X_0 is the *initial distribution*.

- The probabilities $P(X_{i+1} = s|X_i = s')$ are *transition probabilities*.

- The CPT of transition probabilities—remember there's only one—is the *transition matrix*.

- We can number the states $s_1, s_2, \ldots s_m$ and let entry p_{jk} in the transition matrix be $P(X_{i+1} = s_k|X_i = s_j)$, the probability of moving from state s_j to state s_k

Using transition matrices

Matrix multiplication gives the distribution after one iteration of the chain:

```
> rosenthal.mat
     [,1] [,2] [,3] [,4]
[1,]  0.4  0.2  0.3  0.1
[2,]  0.4  0.4  0.2  0.0
[3,]  0.6  0.2  0.1  0.1
[4,]  0.7  0.1  0.0  0.2
```

```
> start
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
```

```
> start %*% rosenthal.mat
     [,1] [,2] [,3] [,4]
[1,]  0.4  0.2  0.3  0.1
```

Variable elimination by matrix multiplication

```
> start %*% rosenthal.mat %*% rosenthal.mat %*% rosenthal.mat
      [,1]  [,2]  [,3]  [,4]
[1,] 0.465 0.237 0.212 0.086

> start2
     [,1] [,2] [,3] [,4]
[1,]    0    0    1    0

> start2 %*% rosenthal.mat %*% rosenthal.mat %*% rosenthal.mat
      [,1]  [,2]  [,3]  [,4]
[1,] 0.473 0.237 0.204 0.086
```

Stationary distribution

- Notice that the two different Markov chains in the previous slide appear to be 'forgetting' their initial distributions and both be converging to a common distribution.

- Let P be the transition matrix for a Markov chain. If π is a distribution such that $\pi P = \pi$ (matrix multiplication) then π is said to be a *stationary distribution*.

- If P_i is a stationary distribution, then $P_i = P_{i+1} = P_{i+2} = \ldots$.

What's stationary

- In a run of the chain, we (generally) continue to move between states once we hit a stationary distribution, but the *probability* of being in any given state will then be constant.

MCMC for approximate sampling

- The aim of Markov chain Monte Carlo (MCMC) methods is to design a Markov chain whose stationary distribution is the target distribution . . .

- . . . such that P_i quickly converges to the stationary distribution *irrespective of the initial distribution*.

- We can then run the chain to produce a sample; throwing away an initial 'burn-in' sample which is too influenced by the initial distribution.

MCMC for joint distributions

- To do approximate probabilistic inference for a joint distribution (e.g. a Bayesian network) we design a Markov chain each state of which is a full joint instantiation of the distribution.

- So a transition is a move from one joint instantiation to another.

- One popular option is to make this transition one variable at a time: *Gibbs sampling*

Gibbs sampling

- Order the variables in the BN somehow: $V_1, V_2, \ldots V_n$

- Suppose the current state is $V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n$. Sample a new value for V_1 from $P(V_1|V_2 = v_2, \ldots, V_n = v_n)$. Let v_1' be the new value.

- Next sample a new value for V_2 from $P(V_2|V_1 = v_1', \ldots, V_n = v_n)$. Then V_3 from $P(V_3|V_1 = v_1', V_2 = v_2', \ldots, V_n = v_n)$

- Continue similarly for $V_4, V_5, \ldots V_n$ until we have a new state $V_1 = v_1', V_2 = v_2', \ldots, V_n = v_n'$

Already instantiated variables

- If a variable is instantiated (i.e. the distribution we want to sample from is a *conditional* distribution),
- Then we don't need to sample a value for it,
- We just 'clamp' it to whatever value it is instantiated to.
- So the more evidence we have, the easier Gibbs sampling is.

It doesn't always work

- Consider a BN $A \rightarrow B$ where both variables are binary, A has a uniform distribution and B has the same value as A with probability one.

- If we start at $A = 0, B = 0$ (with probability one) then a realisation of the Markov chain will remain stuck there,

- even though $P(A = 0.5)$

- This chain is not *ergodic* and so does not converge to the desired distribution.

Continuous variables

- In this module we restrict attention almost entirely to *discrete* random variables—those that have only finitely many values.

- This is artificial: most real problems require continuous random variables: think of temperature, weight, time etc.

- Since it does not depend on manipulating factors, Gibbs sampling can be used for continuous distributions (as well as those with a mixture of discrete and continuous).

- Continuous distributions are defined by a *density function*: probabilities are computed by integrating this function.

The Normal model

- Numerical data is very often modelled by a Normal distribution (hence the name).

- Also known as the Gaussian distribution.

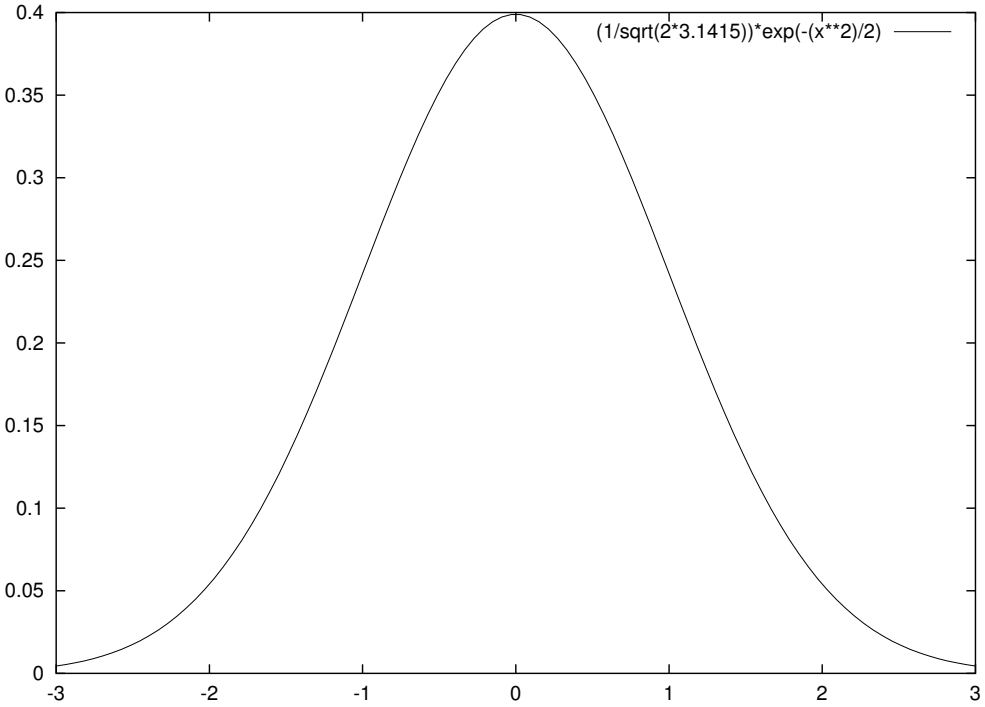
$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$P(z \leq x \leq z') = \int_z^{z'} f(x|\mu,\sigma^2)\,dx$$

The standard Normal distribution

Here is

$$x \sim \text{Norm}(\mu = 0, \sigma^2 = 1)$$



AGM-13

A multivariate continuous distribution

Suppose:

1. $X \sim \text{Norm}(\mu = 0, \sigma^2 = 1)$

2. $Y \sim \text{Norm}(\mu = X, \sigma^2 = 3)$

Then we have a BN $X \rightarrow Y$ to which we can apply Gibbs sampling.

Cue demo (data produced by the BUGS system).

Exploiting conditional independence in Gibbs sampling

- When we sample from $P(V_i|V_1 = v_1, V_2 = v_2, \ldots, V_{i-1} = v_{i-1}, V_{i+1} = v_{i+1}, V_n = v_n)$ we can take advantage of conditional independence.

- Conditional on its neighbours in the interaction graph, V_i is independent of all other variables, so their current values are irrelevant and constructing the right sampling distribution is much simpler.

- This allows rapid Gibbs sampling in very big distributions.

Markov blankets

- Neighbours in the interaction graph form the *Markov blanket* for a variable.

- A Markov blanket shields a variable from the influence of other variables.

- For a BN, the Markov blanket is the variable's children, parents and co-parents in the DAG.