Particle methods and MCMC

CS B553
Spring 2013

Announcements

- A3 posted
  - Due Friday March 8, 11:59PM
- Plan for rest of course
  - Fourth and final assignment after spring break
  - Final project
  - A few more quizzes

Final project

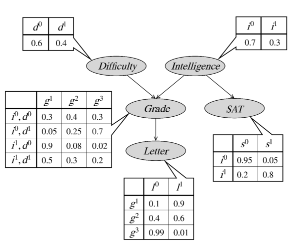
- Choose a topic of interest to you, related to probabilistic approaches (graphical models)
  - Option 1: Choose a research paper that applies probabilistic approaches to a problem. Re-implement, improve, and/or validate their results.
  - Option 2: Apply probabilistic approaches to some new problem of interest to you.
- Work alone or in partnerships
- Deliverables
  - Project proposal: Friday March 22
  - Interim report: Monday April 8
  - Project presentation: Week of April 22
  - Project report and code: Wednesday May 1

Particle-based techniques

- A *particle* is an assignment of values to (some) variables of a graphical model
  - Full particles: assignments of values to all variables
  - Collapsed particles: assignments to some variables
- Basic idea: Sets of particles can be used to approximate a distribution
  - E.g. Many samples from a distribution can be a good representation of original distribution

Forward sampling

- For a Bayes net, we can sample particles using the simple *Forward sampling* algorithm
  - Sample values from priors at root nodes
  - For a node X for which values have been sampled for all parents, sample from P(X | Parents(X))



Computing marginals

- Forward sampling gives a very simple technique for computing marginals over set of variables X
  - Collect many particles using Forward sampling
  - For each possible value of X, count the percentage of sampled particles that have that value:

$$\hat{P}(\mathbf{X} = \mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x})$$

  - where I is an indicator function that is 1 if the equality is true, and 0 otherwise

Handling evidence

- In general, we're interested in computing marginals conditioned on some evidence, i.e. P(X | Y=y)
- One easy way to do this with forward sampling:
  - Sample many particles from the Bayes net
  - If a particle has Y=y, then keep it, else discard it
  - Compute marginals as before, using only the remaining particles
- But this wastes a lot of effort – most samples need to be discarded immediately!

Importance sampling

- Likelihood weighting is a specific case of a more general algorithm, *importance sampling*
- Useful when:
  - We'd like to sample from some distribution P, but doing this is difficult
  - We have some other distribution Q that's close to P and that is easier to sample from
  - Q is called a *proposal distribution*

Importance sampling

- Generate samples $\xi[1], \xi[2], ..., \xi[N]$ from a proposal distribution Q
  - If P=Q, i.e. the proposal distribution is exact,

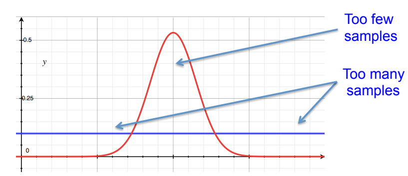
$$\hat{P}(\mathbf{X} = \mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x})$$

  - If P is only an approximation of Q, then we need to apply a correction term to each sample,

$$\hat{P}(\mathbf{X} = \mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x})\frac{P(\xi[i])}{Q(\xi[i])}$$

A simple, trivial example

- Suppose that we want to sample from a Gaussian distribution (P)
  - It's much easier to sample from a uniform distribution (Q)



$$\hat{P}(\mathbf{X} = \mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x})\frac{P(\xi[i])}{Q(\xi[i])}$$

Importance sampling

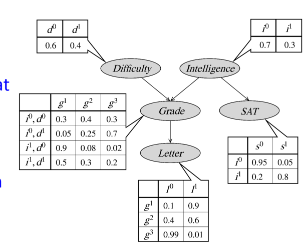
- This formulation assumes we can compute P() exactly
  - But sometimes we can compute P() only up to a normalization constant, i.e. can compute $\tilde{P}()$
- *Normalized importance sampling* avoids computing P() exactly:

$$\hat{P}(\mathbf{X} = \mathbf{x}) = \frac{\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x})\frac{\tilde{P}(\xi[i])}{Q(\xi[i])}}{\sum_{i=1}^{N} \frac{\tilde{P}(\xi[i])}{Q(\xi[i])}}$$

$$= \frac{\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x}) w[i]}{\sum_{i=1}^{N} w[i]}$$

Recall: Likelihood weighting

- Compute a likelihood weight for each particle
  - Initial weight=1
  - Sample values from priors at root nodes
  - For unobserved X for which values have been sampled for all parents, sample from P(X | Parents(X))
  - For observed Y=y, set Y=y but then update weight: w=w * P(Y=y | Parents(Y))



Computing marginals with LW

- LW produces a weighted set of particles
  - To compute P(X=x | Y=y), take sum of weights of particles with X=x, over sum of weights of all sampled particles
- Given samples N samples $(\xi_1, w_1), (\xi_2, w_2), \ldots, (\xi_N, x_N)$,

$$\hat{P}(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = \frac{\sum_{i=1}^{N} w[i] \;\; I(\mathbf{x}[i] = \mathbf{x})}{\sum_{i=1}^{N} w[i]}$$

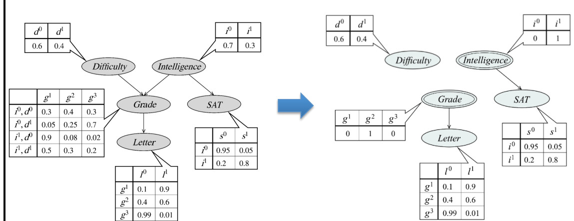
  - where $\mathbf{x}[i]$ refers to the x variables of sample $\xi[i]$, and I is an indicator function that is 1 if the two sets of values are equal

Likelihood weighting

- We can view likelihood weighting as a special case of importance sampling
  - We want to estimate P(X | Y).
  - We don't know how to sample from P(X | Y) directly
  - But we can sample from a different distribution, in which we hard code all of the variables in Y to their observed values, and sample from the rest
  - This proposal distribution Q is sampled from the *mutilated* version of the original Bayes network

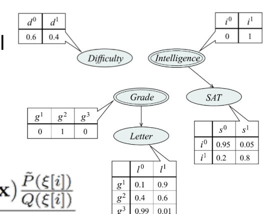
Mutilated networks

- Remove incoming links to observed variables, and set their CPDs to be deterministic
  - Say we observe G and I



Likelihood weighting

- LW samples from the mutilated BN as a proposal distribution
  - Then computes marginals using importance sampling

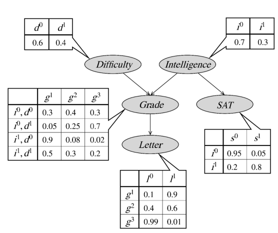


$$\hat{P}(\mathbf{X} = \mathbf{x}) = \frac{\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x}) \frac{\tilde{P}(\xi[i])}{Q(\xi[i])}}{\sum_{i=1}^{N} \frac{\tilde{P}(\xi[i])}{Q(\xi[i])}}$$

$$= \frac{\sum_{i=1}^{N} I(\mathbf{x}[i] = \mathbf{x}) w[i]}{\sum_{i=1}^{N} w[i]}$$

Guarantees

- With high probability, importance sampling will find the correct marginal distribution, ***eventually***.
  - Rate of convergence depends on the quality of Q; i.e. how similar it is to P
  - E.g. Suppose we either observe D and I, or we observe L and S. Which is likely to converge faster?



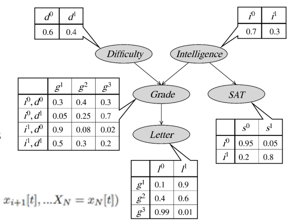
Importance Sampling

- Disadvantages
  - Not clear how to do this on a Markov net
  - If P and Q are not similar, could take a very long time to converge to correct marginal

Markov Chain Monte Carlo (MCMC)

- General class of techniques that produce a *sequence* of samples
- Main idea: Save effort by using information from *past samples* in producing *future samples*
  - Initial samples are from a proposal distribution Q
  - Subsequent sampling is biased towards P
  - Eventually the samples are drawn from a distribution that is closer and closer to P

Example: Gibbs sampling



- Generate initial sample *x[0]*
- For each sample *t=1…T*
  - Let *x[t] = x[t-1]*
  - For each unobserved variable X_i,
    - Sample a value for X_i given values for all other variables in *x[t];* i.e. sample from:

$$P(X_i|X_1 = x_1[t], ...X_{i-1} = x_{i-1}[t], X_{i+1} = x_{i+1}[t], ...X_N = x_N[t])$$
$$= P(X_i|\mathbf{X_{-i}} = \mathbf{x_{-i}}[t])$$

where $\mathbf{X_{-i}} = \mathbf{X} - \{X_i\}$

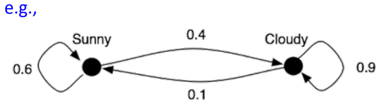
    - Put this sampled value in $x_i[t]$

Properties of Gibbs sampling

- Gibbs can be applied to Markov or Bayes networks
  - Unlike forward sampling and importance sampling, which can in general only be applied to Bayes nets
- Gibbs sampling will converge to sampling from the correct distribution, ***eventually***
  - Under weak assumptions (that the clique potential functions are positive)
  - But may require a long time to converge
  - Why does this happen?

Markov chains

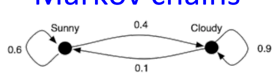
- Stochastic process model
  - Due to Andrey Markov (1906)
  - e.g.,



- The Markov assumption:
  - The probability of transitioning to each new state depends *only* on the current state (and not on the prior states)
  - More formally,

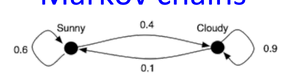
$$P(Q_{t+1} = q_{t+1}|Q_t = q_t, Q_{t-1} = q_{t-1}, ..., Q_0 = q_0) = P(Q_{t+1} = q_{t+1}|Q_t = q_t)$$

Markov chains



- Suppose there's a 80% chance of sun on day 0. What is the probability of sun on day 3?

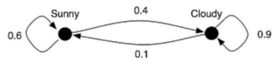
Markov chains



- Suppose there's a 80% chance of sun on day 0. What is the probability of sun on day 3?

$$P(Q_3 = \text{☼}) = P(Q_3 = \text{☼}|Q_2 = \text{☼})P(Q_2 = \text{☼}) + P(Q_3 = \text{☼}|Q_2 = \text{🌧})P(Q_2 = \text{🌧})$$

Markov chains

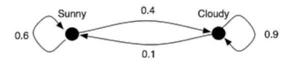


- Suppose there's a 80% chance of sun on day 0. What is the probability of sun on day 3?

$$\begin{aligned}
P(Q_3 = \text{☼}) &= P(Q_3 = \text{☼}|Q_2 = \text{☼})P(Q_2 = \text{☼}) + P(Q_3 = \text{☼}|Q_2 = \text{☁})P(Q_2 = \text{☁}) \\
&= 0.6P(Q_2 = \text{☼}) + 0.1P(Q_2 = \text{☁}) \\
&= 0.6(0.6P(Q_1 = \text{☼}) + 0.1P(Q_1 = \text{☁})) + 0.1(0.4P(Q_1 = \text{☼}) + 0.9P(Q_1 = \text{☁}))
\end{aligned}$$

25

Markov chains

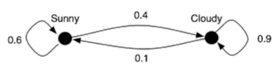


- Suppose there's a 80% chance of sun on day 0. What is the probability of sun on day 3?

$$\begin{aligned}
P(Q_3 = \text{☼}) &= P(Q_3 = \text{☼}|Q_2 = \text{☼})P(Q_2 = \text{☼}) + P(Q_3 = \text{☼}|Q_2 = \text{☁})P(Q_2 = \text{☁}) \\
&= 0.6P(Q_2 = \text{☼}) + 0.1P(Q_2 = \text{☁}) \\
&= 0.6(0.6P(Q_1 = \text{☼}) + 0.1P(Q_1 = \text{☁})) + 0.1(0.4P(Q_1 = \text{☼}) + 0.9P(Q_1 = \text{☁})) \\
&= 0.6(0.6(0.6P(Q_0 = \text{☼}) + 0.1P(Q_0 = \text{☁})) + 0.1(0.4P(Q_0 = \text{☼}) + 0.9P(Q_0 = \text{☁}))) \\
&+ 0.1(0.4(0.4P(Q_0 = \text{☼}) + 0.1P(Q_0 = \text{☁})) + 0.9(0.4P(Q_0 = \text{☼}) + 0.9P(Q_0 = \text{☁})))
\end{aligned}$$

26

Markov chains

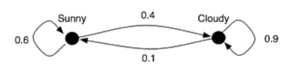


- Suppose there's a 80% chance of sun on day 0. What is the probability of sun on day 3?

$$\begin{aligned}
P(Q_3 = \text{☼}) &= P(Q_3 = \text{☼}|Q_2 = \text{☼})P(Q_2 = \text{☼}) + P(Q_3 = \text{☼}|Q_2 = \text{☁})P(Q_2 = \text{☁}) \\
&= 0.6P(Q_2 = \text{☼}) + 0.1P(Q_2 = \text{☁}) \\
&= 0.6(0.6P(Q_1 = \text{☼}) + 0.1P(Q_1 = \text{☁})) + 0.1(0.4P(Q_1 = \text{☼}) + 0.9P(Q_1 = \text{☁})) \\
&= 0.6(0.6(0.6P(Q_0 = \text{☼}) + 0.1P(Q_0 = \text{☁})) + 0.1(0.4P(Q_0 = \text{☼}) + 0.9P(Q_0 = \text{☁}))) \\
&+ 0.1(0.4(0.4P(Q_0 = \text{☼}) + 0.1P(Q_0 = \text{☁})) + 0.9(0.4P(Q_0 = \text{☼}) + 0.9P(Q_0 = \text{☁}))) \\
&= 0.6(0.6(0.6(0.8) + 0.1(0.2)) + 0.1(0.4(0.8) + 0.9(0.2))) \\
&+ 0.1(0.4(0.6(0.8) + 0.1(0.2)) + 0.9(0.4(0.8) + 0.9(0.2)))
\end{aligned}$$

27

Markov chains



- Suppose there's an 80% chance of sun on day 0. What is the probability of sun on day 3?

$$\begin{aligned}
P(Q_3 = \text{☼}) &= P(Q_3 = \text{☼}|Q_2 = \text{☼})P(Q_2 = \text{☼}) + P(Q_3 = \text{☼}|Q_2 = \text{☁})P(Q_2 = \text{☁}) \\
&= 0.6P(Q_2 = \text{☼}) + 0.1P(Q_2 = \text{☁}) \\
&= 0.6(0.6P(Q_1 = \text{☼}) + 0.1P(Q_1 = \text{☁})) + 0.1(0.4P(Q_1 = \text{☼}) + 0.9P(Q_1 = \text{☁})) \\
&= 0.6(0.6(0.6P(Q_0 = \text{☼}) + 0.1P(Q_0 = \text{☁})) + 0.1(0.4P(Q_0 = \text{☼}) + 0.9P(Q_0 = \text{☁}))) \\
&+ 0.1(0.4(0.4P(Q_0 = \text{☼}) + 0.1P(Q_0 = \text{☁})) + 0.9(0.4P(Q_0 = \text{☼}) + 0.9P(Q_0 = \text{☁}))) \\
&= 0.6(0.6(0.6(0.8) + 0.1(0.2)) + 0.1(0.4(0.8) + 0.9(0.2))) \\
&+ 0.1(0.4(0.6(0.8) + 0.1(0.2)) + 0.9(0.4(0.8) + 0.9(0.2))) \\
&= 0.275
\end{aligned}$$

28