Sistemas Inteligentes

Approximate Inference (Pieter Abbeel UC Berkeley)

José Eduardo Ochoa Luna
Dr. Ciencias - Universidade de São Paulo

Maestría C.C. Universidad Católica San Pablo
Sistemas Inteligentes

29 de noviembre 2018

Approximate Inference

- Simple algorithms like variable elimination may be too slow
- Many interesting classes of models may not admit exact polynomial-time solutions at all
- Much research effort is spent on developing algorithms that yield approximate solutions to the inference problem

Approximate algorithms

- Variational methods: formulate inference as an optimization problem

Approximate algorithms

- Variational methods: formulate inference as an optimization problem
- Sampling methods: produce answers by repeatedly generating random numbers from a distribution of interest

Approximate algorithms

- Variational methods: formulate inference as an optimization problem
- Sampling methods: produce answers by repeatedly generating random numbers from a distribution of interest
- Sampling methods have historically been the main way of performing approximate inference, although over the past 15 years variational methods have emerged as viable (and often superior) alternatives.

Sampling

- Sampling is a lot like repeated simulation: predicting the weather, basketball games,

Sampling

- Sampling is a lot like repeated simulation: predicting the weather, basketball games,
- Basic Idea
  - Draw N samples from a sampling distribution
  - Compute an approximate posterior probability
  - Show this convergences to the true probability P

Sampling

- Sampling is a lot like repeated simulation: predicting the weather, basketball games,
- Basic Idea
  - Draw N samples from a sampling distribution
  - Compute an approximate posterior probability
  - Show this convergences to the true probability P
- Why sample?
  - Learning: get samples from a distribution you don't know
  - Inference: getting a sample is faster than computing the right answer

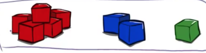
Sampling

1. Get sample u from uniform distribution over $[0, 1)$
2. Convert this sample u into an outcome for the given distribution by having each outcome associated with a sub-interval of $[0, 1)$, with sub-interval size equal to probability of the outcome

C	P(C)
red	0.6
green	0.1
blue	0.3

$$0 \le u < 0.6, \to C = red$$
$$0.6 \le u < 0.7, \to C = green$$
$$0.7 \le u < 1, \to C = blue$$

- If random() returns u = 0.83, then our sample is C = blue
- E.g, after sampling 8 times:



P(blue) $= 2/8$

Sampling in Bayes Nets

- Prior Sampling
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling

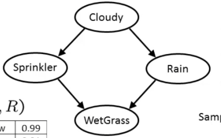
Prior Sampling

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

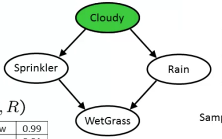
Prior Sampling

Prior Sampling

C → S → R → W
C → R → S → W

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

Prior Sampling

Prior Sampling

C → S → R → W
C → R → S → W

$P(C)$

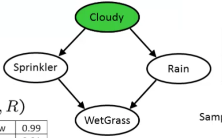
+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8



$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

Prior Sampling

Prior Sampling

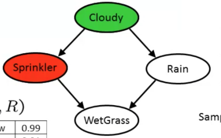
C → S → R → W
C → R → S → W

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

Prior Sampling

Prior Sampling

$C \to S \to R \to W$

$C \to R \to S \to W$

$P(C)$

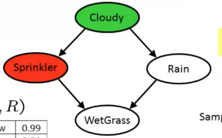
+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8



$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

Prior Sampling

Prior Sampling

C → S → R → W
C → R → S → W

$P(C)$

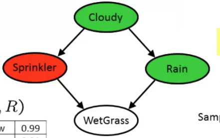
+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8



$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

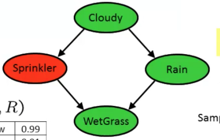
Prior Sampling

Prior Sampling

C → S → R → W
C → R → S → W

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

Prior Sampling

Prior Sampling

$C \to S \to R \to W$
$C \to R \to S \to W$

$P(C)$

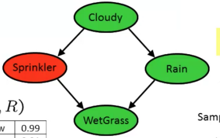
+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8



$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, -s, +r, +w
-c, +s, -r, +w
...

Prior Sampling

- For $i = 1, 2, \ldots, n$
  - Sample x_i from $P(X_i | Parents(X_i))$

Prior Sampling

- For $i = 1, 2, \ldots, n$
  - Sample x_i from $P(X_i | Parents(X_i))$
- Return $(x_1, x_2, \ldots, x_n)$

Prior Sampling

- For $i = 1, 2, \ldots, n$
  - Sample x_i from $P(X_i|Parents(X_i))$
- Return $(x_1, x_2, \ldots, x_n)$
- Repeat this process for more samples

Prior Sampling

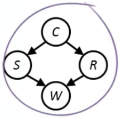
- This process generates samples with probability: $S_{PS}(x_1, \ldots, x_n) = \prod_{i=1}^{n}(P(x_i|Parents(X_i)) = P(x_1, \ldots, x_n)$
- Let the number of samples of an event be $N_{PS}(x_1, \ldots, x_n)$
- Then

$$\begin{aligned} \lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N \\ &= S_{PS}(x_1, \ldots, x_n) \\ &= P(x_1, \ldots, x_n) \end{aligned}$$

- the sampling procedure is consistent

Example

- We'll get a bunch of samples from the BN:
  - +c, -s, +r, +w
  - +c, +s, +r, +w
  - -c, +s, +r, -w
  - +c, -s, +r, +w
  - -c, -s, -r, +w



- If we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too

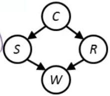
Example

- We' ll get a bunch of samples from the BN:
  - +c, -s, +r, +w
  - +c, +s, +r, +w
  - -c, +s, +r, -w
  - +c, -s, +r, +w
  - -c, -s, -r, +w

P(C|+w)

+c 3/4

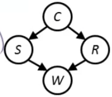
-c 1/4



- If we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too
  - What about P(C| +w)? P(C| +r, +w)? P(C| -r, -w)?

Example

- We'll get a bunch of samples from the BN:
  - +c, -s, +r, +w
  - +c, +s, +r, +w
  - -c, +s, +r, -w
  - +c, -s, +r, +w
  - -c, -s, -r, +w



$P(C|+w)$

+c $3/4$

-c $1/4$

- If we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too
  - What about P(C| +w)? P(C| +r, +w)? P(C| -r, -w)?

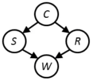
$P(+c|+r,+w)=1$

$-c \quad =0$

Rejection Sampling

Let’s say we want $P(C)$

- No point keeping all samples around
- Just tally counts of C as we go

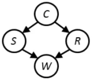


+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w

Rejection Sampling

Let's say we want $P(C| + s)$

- Same things: tally C outcomes but ignore (reject) samples which don't have $S = +s$
- This is called rejection sampling
- It is also consistent for conditional probabilities (i. e., correct in the limit)



+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w

Rejection Sampling

- IN: evidence instantiation
- For $i = 1, 2, \ldots, n$
  - Sample x_i from $P(X_i|Parents(X_i))$
  - if x_i not consistent with evidence then reject: Return and no sample is generated in this cycle
- Return $(x_1, x_2, \ldots, x_n)$

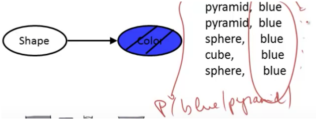
Problem with Rejection Sampling

- If evidence is unlikely, rejects lots of samples
- Evidence not exploited as you sample

Likelihood Weighting

Idea: fix evidence variables and sample the rest

- Problem: sample distribution not consistent
- Solution: weight by probability of evidence given parents



Likelihood Weighting

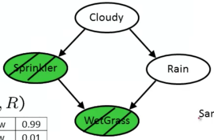
Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S, R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

...

$w = 1.0$

Likelihood Weighting

Likelihood Weighting

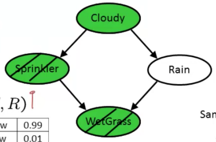
+s, +w

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

...

$$w = 1.0 \times 0.1$$

Likelihood Weighting

Likelihood Weighting

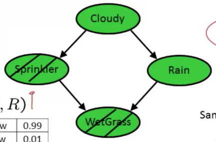
+s, +w

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

...

$$w = 1.0 \times 0.1$$

Likelihood Weighting

Likelihood Weighting

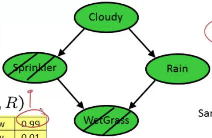
+s, +w

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

...

$$w = 1.0 \times 0.1 \times 0.99$$

Likelihood Weighting

Likelihood Weighting

+s, +w

$P(C)$

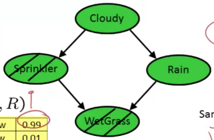
+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8



$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

...

$w = 1.0 \times 0.1 \times 0.99$ = 0.099

Likelihood Weighting

- IN: evidence instantiation
- $w = 1.0$
- For $i = 1, 2, \ldots, n$
  - if X_i is an evidence variable
    - $X_i =$ observation x_i for X_i
    - Set $w = w * P(x_i|Parents(X_i))$
  - else Sample x_i from $P(X_i|Parents(X_i))$
- Return $(x_1, x_2, \ldots, x_n), w$

Likelihood Weighting Problem

- Evidence influences the choice of downstream variables, but not upstream one (C isn't more likely to get a value matching the evidence)
- We would like to consider evidence when we sample every variable

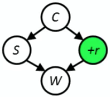
Gibbs Sampling

- Procedure: keep track of a full instantiation $x_1, x_2, \ldots, x_n$. Start with an arbitrary instantiation consistent with evidence. Sample one variable at a time, conditioned on all the rest, but keep evidence fixed. Keep repeating this for a long time
- Property: in the limit of repeating this infinitely many times the resulting sample is coming from the correct distribution $P(-|evidence)$
- Rationale: both upstream and downstream variable condition on evidence

Gibbs Sampling

Gibbs Sampling Example: P(S | +r)

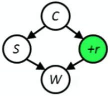
- Step 1: Fix evidence
  - R = +r



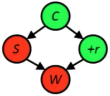
Gibbs Sampling

Gibbs Sampling Example: P(S | +r)

- Step 1: Fix evidence
  - R = +r



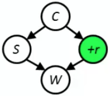
- Step 2: Initialize other variables
  - Randomly



Gibbs Sampling

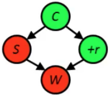
Gibbs Sampling Example: P(S | +r)

- Step 1: Fix evidence
  - R = +r



- Steps 3: Repeat
  - Choose a non-evidence variable X

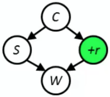
- Step 2: Initialize other variables
  - Randomly



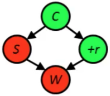
Gibbs Sampling

Gibbs Sampling Example: P(S | +r)

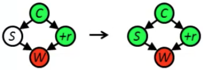
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly

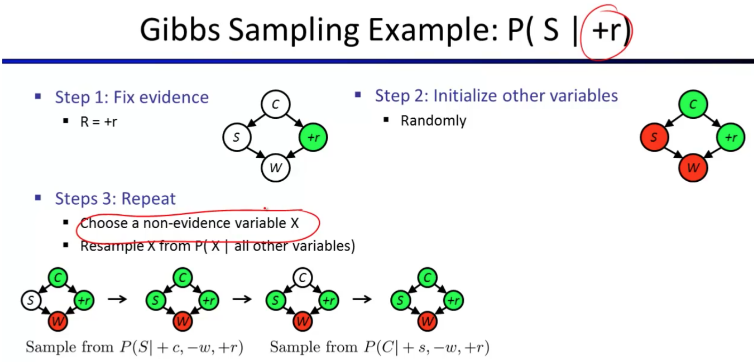


- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)



Sample from $P(S| + c, -w, +r)$

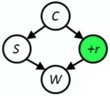
Gibbs Sampling



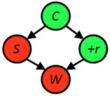
Gibbs Sampling

Gibbs Sampling Example: P(S | +r)

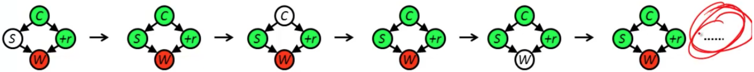
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly



- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

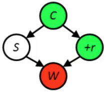


Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$ Sample from $P(W|+s,+c,+r)$

Efficient Resampling of One Variable

Sample from $P(S|+c,+r,-w)$

$$
\begin{aligned}
P(S|+c,+r,-w) &= \frac{P(S,+c,+r,-w)}{P(+c,+r,-w)} \\
&= \frac{P(S,+c,+r,-w)}{\sum_s P(s,+c,+r,-w)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s,+r)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{P(+c)P(+r|+c)\sum_s P(s|+c)P(-w|s,+r)} \\
&= \frac{P(S|+c)P(-w|S,+r)}{\sum_s P(s|+c)P(-w|s,+r)}
\end{aligned}
$$



Many things cancel out, only CPTS with S remain

Gibbs Sampling

- Gibbs sampling produce sample from the query distribution $P(Q|e)$ in limit of re-sampling infinitely often
- Gibbs sampling is a special case of more general methods called Markov Chain Monte Carlo Methods (MCMC), Metropolis-Hastings (MH) is one of the more famous MCMC methods and Gibbs sampling is a special case of MH
- Monte Carlo Methods are just sampling

Tarea

- Modelar la Student Bayes Network, usar por ejemplo SamIam
- Crear un Jupyter Notebook e implementar Variable Elimination para esta RB
- Verificar si las inferencias obtenidas son las mismas que en SamIam