CS 583: Probabilistic Graphical Models

Topic: Sampling

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

SAMPLING MOTIVATION

- Exact inference is NP-hard
- Exact inference with an arbitrary structured network containing thousands of variables is impractical
- Various approximate inference techniques
  - Variational inference
  - Sampling

SAMPLING

- The basic idea
  - Generate data using the network and the parameters
  - Use the data to answer the queries
- For this to work
  - Sampling needs to be more efficient than running inference
  - Enough data need to be sampled for precision

WE'LL COVER

- **Forward sampling**
  - Bayesian networks, no evidence
- **Rejection sampling**
  - Bayesian networks, with evidence
- **Likelihood weighting**
  - Bayesian networks, with evidence
- **Gibbs sampling**
  - Bayesian networks and Markov networks, with or without evidence

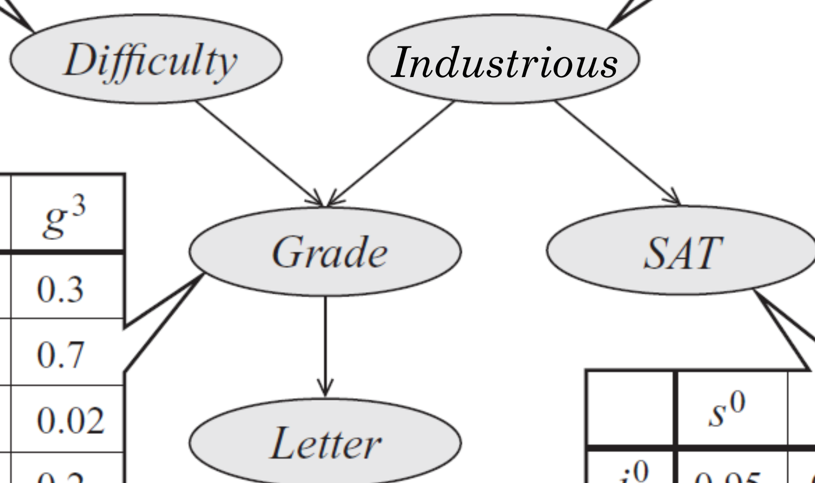
Preliminaries: how to sample from a distribution

- Discrete
  - Binary $[p, 1\text{-}p]$
    - Sample a random number r from [0,1]. If $r<p$, then it is the first value, otherwise it is the second.
  - Multinomial $[p^1, p^2, \ldots, p^n]$
    - Create $[0, p^1, p^1+p^2, p^1+p^2+p^3, \ldots, p^1+p^2+\ldots+p^n]$
    - Sample a random number r from [0, 1]. Find i where $p^1+p^2+\ldots+p^{i-1} < r < p^1+p^2+\ldots+p^i$
- Continuous
  - Depends on the distribution
  - For e.g., many different methods to sample from Gaussian distribution

Forward sampling

- Use for Bayesian networks and marginal probabilities
- For each variable V_i that is ready
  - Sample a value v_i for V_i using $P(V_i \mid Pa(V_i))$
- Repeat this process M times to generate M instances
- A variable is ready if
  - It has no parents, or
  - You have sampled its all parents
- To compute marginal
  - Use maximum likelihood estimate
    - Count and normalize

Forward sampling on the student network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

1. D and I are ready.
2. Sample D from P(D): d^0.
3. Only I is ready.
4. Sample I from P(I): i^1.
5. G and S are now ready.
6. Sample G from $P(G|i^1,d^0)$: g^1.
7. S and L are ready.
8. Sample S from $P(S|i^1)$: s^0.
9. L is ready.
10. Sample L from $P(L|g^1)$: l^0
11. The instance is $\langle d^0,i^1,s^0,g^1,l^0\rangle$
12. Repeat the process from step 1 M times.

FORWARD SAMPLING ON THE STUDENT NETWORK

- The sampled data is

Iteration	D	I	S	G	L
1	d^0	i^1	s^0	g^1	l^0
…	…	…	…	…	…
M	…	…	…	…	…

$P(D=d^0)$ = # of rows with $D=d^0$ / (# of rows with $D=d^0$ + # of rows with $D=d^1$)

Bounds

- Absolute error ε bound with probability at least 1-δ
  - $M \geq \ln(2/\delta) / 2\varepsilon^2$
  - E.g.
    - δ=0.1, ε=0.1 $\Rightarrow$ M ≥ 150
    - δ=0.01, ε=0.1 $\Rightarrow$ M ≥ 265
    - δ=0.01, ε=0.01 $\Rightarrow$ M ≥ 26,491
- Relative error ε bound with probability at least 1-δ
  - $M \geq 3\ln(2/\delta) / P(\boldsymbol{y})\varepsilon^2$
  - A big problem with using this bound is that we do not know P($\boldsymbol{y}$)

Conditional probability queries

- P($\boldsymbol{y}$, $\boldsymbol{e}$) and P($\boldsymbol{e}$) can be separately estimated
- Then P($\boldsymbol{y}$ | $\boldsymbol{e}$) = P($\boldsymbol{y}$, $\boldsymbol{e}$)/ P($\boldsymbol{e}$)
- For this to work, both P($\boldsymbol{y}$, $\boldsymbol{e}$) and P($\boldsymbol{e}$) need to be estimated with *relative* low error
- If we estimate P($\boldsymbol{y}$, $\boldsymbol{e}$) and P($\boldsymbol{e}$) with small *absolute* error, then P($\boldsymbol{y}$, $\boldsymbol{e}$)/ P($\boldsymbol{e}$) can be arbitrarily off.

Where is the evidence?

- If evidence is only at the root variables, it is easy; don't sample those variables; just set them to their respective values
  - E.g., if $\mathbf{E} = \{d^1, i^0\}$ in the student network, then don't sample D and I. Just set D=d^1 and I=i^0
- If the evidence is at the intermediate or leaf nodes (e.g., if any of G, S, L is in the evidence)
  - Rejection sampling
  - Likelihood sampling

REJECTION SAMPLING

- Given evidence $\boldsymbol{e}$
- Sample an instance $\boldsymbol{x}^{(i)}$ using forward sampling
- If $\boldsymbol{x}^{(i)}$ and and $\boldsymbol{e}$ disagree, then reject the instance
- To compute the conditional, use MLE
  - Count and normalize
- If we generate M instances, how many of them will be rejected/kept?

Likelihood weighting

- Sample like forward sampling, except
  - When a variable is in the evidence set,
    - Set its value to evidence value
- Each instance has a weight
  - $w = \Pi_{\boldsymbol{v} \in \boldsymbol{e}} \mathrm{P}(\boldsymbol{v} \mid \mathrm{Pa}(\boldsymbol{v}))$
- Counts are now weighted by each instance's weight

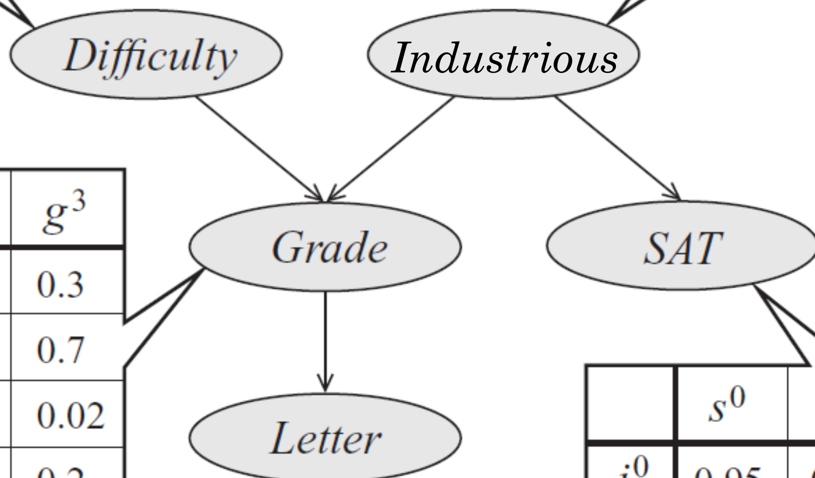
Likelihood weighting on a chain

- Network
  - $A \rightarrow B$
- Parameters
  - $\mathrm{P}(A) = [p;\ 1\text{-}p]$
  - $\mathrm{P}(B \mid A{=}t) = [q;\ 1\text{-}q]$
  - $\mathrm{P}(B \mid A{=}f) = [r;\ 1\text{-}r]$
- $\mathrm{P}(A \mid B{=}t) = ?$

P(A | B=T)

- Exact inference
  - P(A=t | B=t) =
    - P(A=t, B=t) / P(B=t)
    - P(A=t)P(B=t | A=t) / Σ_A P(A)P(B=t | A)
    - $p^*q / (p^*q + (1-p)^*r)$
- Likelihood weighting
  - Sample M instances
    - Sample A randomly from [p, 1-p]
    - Set B=t
    - The weight of the instance i is
      - If A=t, w_i=P(B=t | A=t)=q, else w_i=P(B=t | A=f)=r
  - Out of M instances
    - Approximately p^*M have A=t and each has weight q
    - Approximately $(1-p)^*M$ have A=f and each has weight r
    - P(A=t | B=t) = $p^*M^*q / (p^*M^*q + (1-p)^*M^*r) = p^*q / (p^*q + (1-p)^*r)$

Likelihood weighting on the Student network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Assume $S=s^1$

1. **w=1**
2. D and I are ready.
3. Sample D from P(D): d^0.
4. I is ready.
5. Sample I from P(I): i^1.
6. G and S are now ready.
7. Sample G from $P(G|i^1,d^0)$: g^1.
8. S and L are ready.
9. **Set S=s1**
10. **$w=w*P(s^1|i^1)$**
11. L is ready.
12. Sample L from $P(L|g^1)$: l^0
13. The instance is $<d^0,i^1,s1,g^1,l^0>$ **and its weight is w**
14. Repeat the process from step 1 M times.

Gibbs sampling

- Works for both
  - Bayesian and Markov networks
  - With and without evidence
- Huge body of work on it
- I will cover the simplest version
- More details can be found at Chapter 12 Section 3

Gibbs sampling

- All variables: $\mathcal{X}$, evidence variables: $\boldsymbol{E}$, variables of interest: $\boldsymbol{Y} \subseteq \mathcal{X} \setminus \boldsymbol{E}$

1. Set evidence variables $\boldsymbol{E}$ to their values $\boldsymbol{e}$
2. Initialize the remaining variables $\mathcal{X} \setminus \boldsymbol{E}$ somehow (random is (probably) OK)
3. For each variable $X_i \in \mathcal{X} \setminus \boldsymbol{E}$
   - Sample X_i using $P(X_i \mid \mathcal{X} \setminus X_i)$
4. Discard the first N instances
5. Use the last M instances to compute $P(\boldsymbol{Y} \mid \boldsymbol{e})$

P(X_i | $\mathcal{X} \setminus X_i$)

- P(I|D=d^0, G=g^2, L=l^1, S=s^1) = ?

$$P(I=i^0 \mid D=d^0, G=g^2, L=l^1, S=s^1)$$

$$=\frac{P(i^0,d^0,g^2,l^1,s^1)}{P(d^0,g^2,l^1,s^1)}$$

$$=\frac{P(i^0,d^0,g^2,l^1,s^1)}{P(i^0,d^0,g^2,l^1,s^1)+P(i^1,d^0,g^2,l^1,s^1)}$$

$$=\frac{P(i^0)P(d^0)P(g^2 \mid i^0,d^0)P(l^1 \mid g^2)P(s^1 \mid i^0)}{P(i^0)P(d^0)P(g^2 \mid i^0,d^0)P(l^1 \mid g^2)P(s^1 \mid i^0)+P(i^1)P(d^0)P(g^2 \mid i^1,d^0)P(l^1 \mid g^2)P(s^1 \mid i^1)}$$

$$=\frac{P(i^0)P(d^0)P(g^2 \mid i^0,d^0)P(l^1 \mid g^2)P(s^1 \mid i^0)}{P(d^0)P(l^1 \mid g^2)\left(P(i^0)P(g^2 \mid i^0,d^0)P(s^1 \mid i^0)+P(i^1)P(g^2 \mid i^1,d^0)P(s^1 \mid i^1)\right)}$$

$$=\frac{P(i^0)P(g^2 \mid i^0,d^0)P(s^1 \mid i^0)}{P(i^0)P(g^2 \mid i^0,d^0)P(s^1 \mid i^0)+P(i^1)P(g^2 \mid i^1,d^0)P(s^1 \mid i^1)}$$

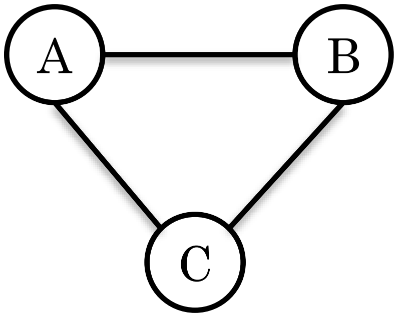
$P(X_i \mid \mathcal{X} \setminus X_i)$

- Multiply all the factors that include X_i using the most recently sampled (or evidence) values for the remaining variables
- Normalize it
- The approach works for both Bayesian and Markov networks

Markov network example

A	B	φ(A,B)
T	T	5
T	F	1
F	T	1
F	F	5

A	φ(A)
T	2
F	1



B	φ(B)
T	1
F	4

A	C	φ(A,C)
T	T	6
T	F	1
F	T	1
F	F	6

C	φ(C)
T	1
F	8

B	C	φ(B,C)
T	T	1
T	F	10
F	T	10
F	F	1

Start with random values: A=F, B=T, C=T.
Sample A. Which distribution do we sample A from?