Bayes Networks

CS540 Bryan R Gibson University of Wisconsin-Madison

Slides adapted from those used by Prof. Jerry Zhu, CS540-1

Outline

- Joint Probability:
  great for inference, terrible to obtain and store
- Bayes Nets: build joint distributions in manageable chunks
  - using Independence and Conditional Independence
- Inference in Bayes Nets
  - naive algorithms can be terribly inefficient
  - more efficient algorithms can be found
- Parameter Learning in Bayes Nets

Creating a Joint Distribution

- Making a joint distribution of N variables
  1. List all combinations of values
     (if each variable has k values, k^N combinations)
  2. Assign each combination a probability
  3. Check that they sum to 1

Weather	Temp	Prob.
sunny	hot	150/365
sunny	cold	50/365
cloudy	hot	40/365
cloudy	cold	60/365
rainy	hot	5/365
rainy	cold	60/365
		365/365

Using a Joint Distribution

- Once you have the joint distribution, you can do everything e.g. marginalization:

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

- Example: $P(\text{sunny or hot}) = (150 + 50 + 40 + 5)/365$

  convince yourself this is the same as $P(\text{sunny}) + P(\text{hot}) - P(\text{sunny and hot})$

Weather	Temp	Prob.
sunny	hot	150 365
sunny	cold	50 365
cloudy	hot	40 365
cloudy	cold	60/365
rainy	hot	5 365
rainy	cold	60/365

Using a Joint Distribution (cont.)

- You can also do inference:

$$P(Q \mid E) = \frac{\sum_{\text{rows matching } Q \text{ AND } E} P(\text{row})}{\sum_{\text{rows matching } E} P(\text{row})}$$

- Example: $P(\text{hot} \mid \text{rainy}) = 5/65$

Weather	Temp	Prob.
sunny	hot	150/365
sunny	cold	50/365
cloudy	hot	40/365
cloudy	cold	60/365
rainy	hot	5 365
rainy	cold	60 365

The Bad News

- Joint distribution can take up a huge amount of space
- Remember: for N variables each taking k values,
  the joint distribution table has k^N numbers
- It would be good to be able to use fewer numbers ...

Using fewer numbers

- Example: Suppose there are 2 events
  - B: there's a burglary in your house
  - E: there's an earthquake
- The joint distribution has 4 entries
  $P(B, E),\ P(B, \neg E),\ P(\neg B, E), P(\neg B, \neg E)$
- Do we have to come up with these 4 numbers?
- Can we 'derive' them just using P(B) and P(E) instead?
- What assumption do we need?

Independence

- Assume: "Whether there's a burglary doesn't depend on whether there's an earthquake"
- This is encoded as

$$P(B \mid E) = P(B)$$

- This is a strong statement!
- Equivalently:

$$P(E \mid B) = P(E)$$
$$P(B, E) = P(B)P(E)$$

- It requires domain knowledge outside of probability.
- It needed an understanding of causation

Independence (cont.)

- With independence, we have

$$P(B, \neg E) = P(B)P(\neg E)$$
$$P(\neg B, E) = P(\neg B)P(E)$$
$$P(\neg B, \neg E) = P(\neg B)P(\neg E)$$

- Say $P(B) = 0.001$, $P(E) = 0.002$, $P(B \mid E) = P(B)$
- The joint probability table is:

Burglary	Earthquake	Prob
B	E	
B	$\neg E$	
$\neg B$	E	
$\neg B$	$\neg E$	

- Now we can do anything, since we have the joint.

A More Interesting Example ...

- Let:
  - B: there's a burglary in your house
  - E: there's an earthquake
  - A: your alarm goes off
- Your alarm is supposed to go off when there's a burglary ...
- but sometimes it doesn't ...
- and sometimes it is triggered by an earthquake.
- The knowledge we have so far:
  - $P(B) = 0.001$, $P(E) = 0.002$, $P(B \mid E) = P(B)$
  - Alarm is NOT independent of whether there's a burglary, nor is it independent of earthquake
- We already know the joint of B, E. All we need is:

$$P(A \mid \text{Burglary} = b, \text{Earthquake} = e)$$

for the 4 cases of $b = \{B, \neg B\}$, $e = \{E, \neg E\}$ to get full joint.

A More Interesting Example (cont.)

- B: there's a burglary in your house
- E: there's an earthquake
- A: your alarm goes off
- Your alarm is supposed to go off when there's a burglary
  but sometimes it doesn't
  and sometimes it is triggered by an earthquake.

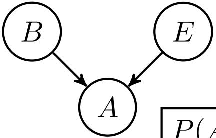
$P(B)=0.001$	$P(A \mid B, E)=0.95$
$P(E)=0.002$	$P(A \mid B, \neg E)=0.94$
$P(B \mid E)=P(B)$	$P(A \mid \neg B, E)=0.29$
	$P(A \mid \neg B, \neg E)=0.001$

- These 6 numbers specify the joint, instead of 7
- Savings are larger with more variables!

Introducing Bayes Nets

$$
\boxed{\begin{array}{rl}
P(B)=0.001 & P(A \mid B, E)=0.95 \\
P(E)=0.002 & P(A \mid B, \neg E)=0.94 \\
P(B \mid E)=P(B) & P(A \mid \neg B, E)=0.29 \\
 & P(A \mid \neg B, \neg E)=0.001
\end{array}}
$$

$$\boxed{P(B) = 0.001}$$



$$\boxed{P(E) = 0.002}$$

$$
\boxed{\begin{array}{l}
P(A \mid B, E) = 0.95 \\
P(A \mid B, \neg E) = 0.94 \\
P(A \mid \neg B, E) = 0.29 \\
P(A \mid \neg B, \neg E) = 0.001
\end{array}}
$$

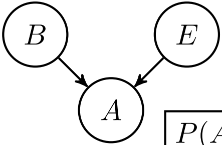
Joint Probability with Bayes Nets

$$P(x_1, \ldots, x_n) = \prod_i P(x_i \mid \text{parents}(x_i))$$

$$\text{Example: } P(\neg B, E, \neg A) = P(\neg B) P(E \mid \neg B) P(\neg A \mid \neg B, E)$$

$$= P(\neg B) P(E) P(\neg A \mid \neg B, E)$$

$$\boxed{P(B) = 0.001}$$



$$\boxed{P(E) = 0.002}$$

$$\boxed{\begin{aligned} P(A \mid B, E) &= 0.95 \\ P(A \mid B, \neg E) &= 0.94 \\ P(A \mid \neg B, E) &= 0.29 \\ P(A \mid \neg B, \neg E) &= 0.001 \end{aligned}}$$

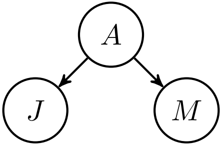
More to the story . . .

- A: your alarm sounds
- J: your neighbor John calls
- M: your neighbor Mary calls
- John and Mary don't communicate but
  they will both call if they hear the alarm

- What kind of independence do we have?

  Conditional Independence: $P(J, M \mid A) = P(J \mid A)P(M \mid A)$

- What does the Bayes Net look like?



Now An Example with 5 Variables

- B: there's a burglary in your house
- J: there's an earthquake
- A: your alarm sounds
- J: your neighbor John calls
- M: your neighbor Mary calls

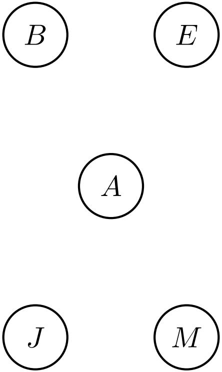
- B, E are independent

- J is only directly influenced by A
  J is conditionally independent of B, E, M given A

- M is only directly influenced by A
  M is conditionally independent of B, E, J given A

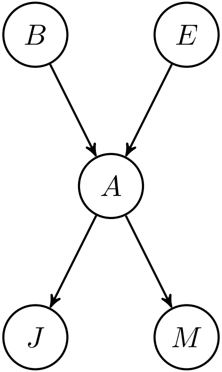
Creating a Bayes Net

- Step 1: add variables (one variable per node)



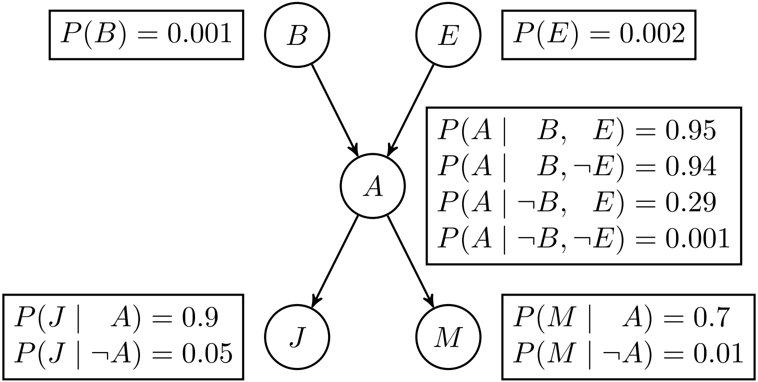
Creating a Bayes Net (cont.)

- Step 2: add directed edges
  - graph must be acyclic
  - if node X has parents $Q_1, \ldots Q_m$, you are promising that any variable that's not a descent of X is conditionally indpendent of X given $Q_1, \ldots, Q_m$



Creating a Bayes Net (cont.)

- Step 3: add CPTs
  - each CPT lists $P(X \mid \text{Parents})$ for all comb. of parent values
  - e.g. you must specify $P(J \mid A)$ AND $P(J \mid \neg A)$,
  - they don't need to sum to 1!

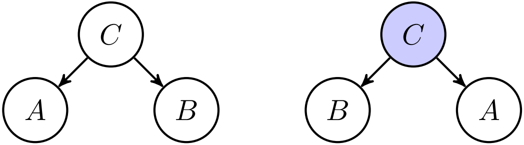


Creating a Bayes Net: Summary

1. Choose a set of relevant variables
2. Choose an ordering of them, say $x_1, \ldots, x_n$
3. for $i = 1$ to n
    3.1 Add node x_i to the graph
    3.2 Set parents(x_i) to be the minimal subset of $\{x_1, \ldots, x_{i-1}\}$ s.t. x_i is cond. indep. of all other members of $\{x_1, \ldots, x_{i-1}\}$ given parents(x_i)
    3.3 Define the CPT's for $P(x_i \mid$ assignment of parents$(x_i))$

Representing Conditional Independence

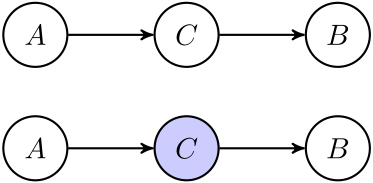
- Case 1: Tail-to-Tail



- A,B in general not independent
- But A,B conditionally independent given C
- C is “tail-to-tail” node: if C is observed, it blocks path

Representing Conditional Independence (cont.)

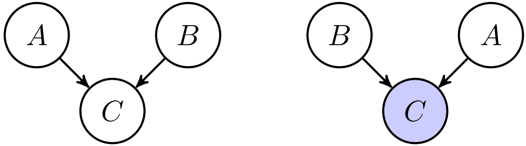
- Case 2: Head-to-Tail



- A,B in general not independent
- But A,B conditionally independent given C
- C is "head-to-tail" node: if C is observed, it blocks path

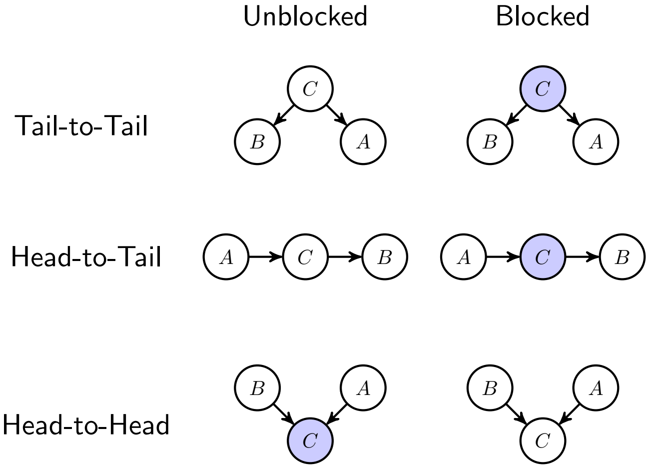
Representing Conditional Independence (cont.)

- Case 3: Head-to-Head



- A,B in general independent
- But A,B NOT conditionally independent given C
- C is "head-to-head" node: if C is observed, it unblocks path, or, importantly, if any of C's decendents are observed

Representing Conditional Independence (cont.)

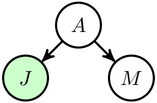


Representing Conditional Independence: Example

	Unblocked	Blocked

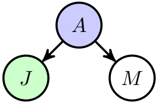
Tail-to-Tail

Unblocked:



$P(M \mid J) \neq P(M)$

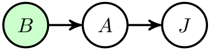
Blocked:



$P(M \mid J, A) = P(M \mid A)$

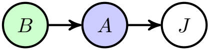
Head-to-Tail

Unblocked:



$P(J \mid B) \neq P(J)$

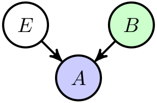
Blocked:



$P(J \mid B, A) = P(J \mid A)$

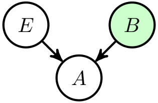
Head-to-Head

Unblocked:



$P(E \mid B, A) \neq P(E \mid A)$

Blocked:

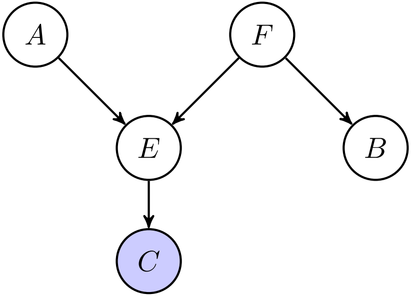


$P(E \mid B) = P(E)$

D-Separation

- For any groups of nodes A,B and C:
- A and B are independent given C if:
  - all (undirected) paths from any node in A to any node in B are blocked
- A path is blocked if it includes a node s.t. either:
  - The arrows on the path meet head-to-tail or tail-to-tail at the node, and the node is in C, or
  - The arrows meet head-to-head at the node, and neither the node, nor any of its descendents, is in C

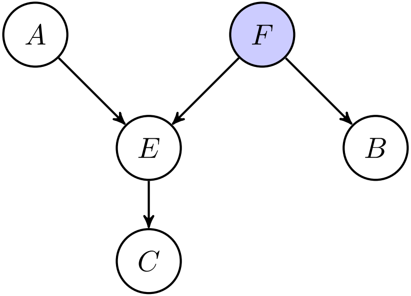
D-Separation: Examples



- The path from A to B is not blocked by either E or F
- But A,B conditionally dependent given C:

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

D-Separation: Examples (cont.)

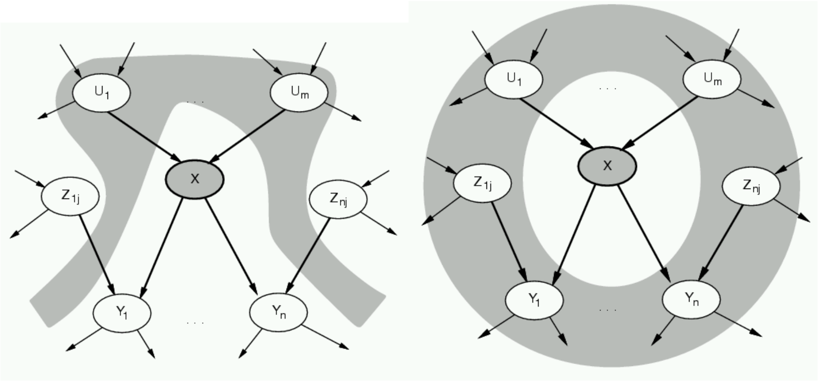


- The path from A to B is blocked both at E and F
- But A,B conditionally independent given F:

$$P(A, B \mid F) = P(A \mid F)P(B \mid F)$$

Conditional Independence in Bayes Nets

- a node is cond. indep. of its non-descendents given its parents
- a node is cond. indep. of all other nodes given its
  Markov Blanket (parents, children, spouses)



Compactness of a Bayes Net

- Bayes Nets encode joint dists., often with far fewer parameters
- Recall, a full joint table needs k^N parameters
  - N variables, k values per variable
  - grows exponentially with N
- If the Bayes Net is sparse, e.g. each node has at most M parents ($M << N$), it only needs $O(Nk^M)$ parameters
  - grows linearly with N
  - can't have too many parents though

Summary so far . . .

- We can define a Bayes Net, using a small number of parameters, to describe a joint probability.
- Any joint probability can be computed as:

$$P(x_1, \ldots, x_n) = \prod_i P(x_i \mid \text{parents}(x_i))$$

- The above joint probability can be computed in time linear with number of nodes N
- With this distribution, we can compute any conditional probability $P(Q \mid E)$, thus we can perform inference.
- How?

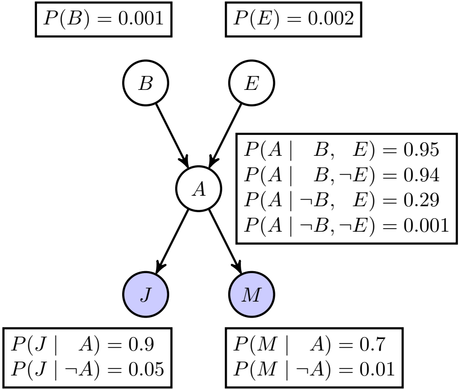
Inference by Enumeration

$$P(Q \mid E) = \frac{\sum_{\text{joint matching } Q \text{ AND } E} P(\text{joint})}{\sum_{\text{joint matching } E} P(\text{joint})}$$

For Example: $P(B \mid J, \neg M)$

1. Compute $P(B, J, \neg M)$
2. Compute $P(J, \neg M)$
3. Return

$$\frac{P(B, J, \neg M)}{P(J, \neg M)}$$



Inference by Enumeration

Sum up:

$P(B, J, \neg M, A, E)$

$P(B, J, \neg M, A, \neg E)$

$P(B, J, \neg M, \neg A, E)$

$P(B, J, \neg M, \neg A, \neg E)$

Each one O(N) for sparse graph

Q AND E P(joint)

tching E P(joint)

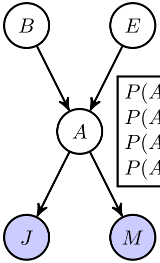
For Example: $P(B \mid J, \neg M)$

1. Compute $P(B, J, \neg M)$
2. Compute $P(J, \neg M)$
3. Return

$$\frac{P(B, J, \neg M)}{P(J, \neg M)}$$

$B) = 0.001$

$P(E) = 0.002$



$P(A \mid B, E) = 0.95$
$P(A \mid B, \neg E) = 0.94$
$P(A \mid \neg B, E) = 0.29$
$P(A \mid \neg B, \neg E) = 0.001$

$P(J \mid A) = 0.9$
$P(J \mid \neg A) = 0.05$

$P(M \mid A) = 0.7$
$P(M \mid \neg A) = 0.01$

Inference by Enumeration

$$P(Q \mid E) = \frac{\sum_{\text{joint matching } Q \text{ AND } E} P(\text{joint})}{\sum_{\text{joint matching } E} P(\text{joint})}$$

For Example: $P(B \mid J, \neg M)$

1. Compute $P(B, J, \neg M)$
2. Compute $P(J, \neg M)$
3. Return

$$\frac{P(B, J, \neg M)}{P(J, \neg M)}$$

Sum up:

$P(J, \neg M, B, A, E)$

$P(J, \neg M, B, A, \neg E)$

$P(J, \neg M, B, \neg A, E)$

$P(J, \neg M, B, \neg A, \neg E)$

$P(J, \neg M, \neg B, A, E)$

$P(J, \neg M, \neg B, A, \neg E)$

$P(J, \neg M, \neg B, \neg A, E)$

$P(J, \neg M, \neg B, \neg A, \neg E)$

Each one O(N) for sparse graph

$= 0.002$

$\mid B, E) = 0.95$
$\mid B, \neg E) = 0.94$
$\mid \neg B, E) = 0.29$
$\mid \neg B, \neg E) = 0.001$

$P(J \mid A) = 0.9$
$P(J \mid \neg A) = 0.05$

$P(M \mid A) = 0.7$
$P(M \mid \neg A) = 0.01$

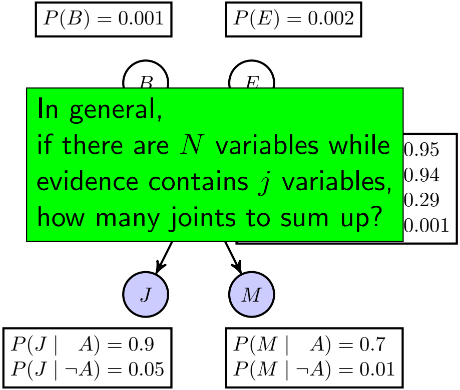
Inference by Enumeration

$$P(Q \mid E) = \frac{\sum_{\text{joint matching } Q \text{ AND } E} P(\text{joint})}{\sum_{\text{joint matching } E} P(\text{joint})}$$

For Example: $P(B \mid J, \neg M)$

1. Compute $P(B, J, \neg M)$
2. Compute $P(J, \neg M)$
3. Return

$$\frac{P(B, J, \neg M)}{P(J, \neg M)}$$



In general,
if there are N variables while evidence contains j variables, how many joints to sum up?

Inference by Enumeration (cont.)

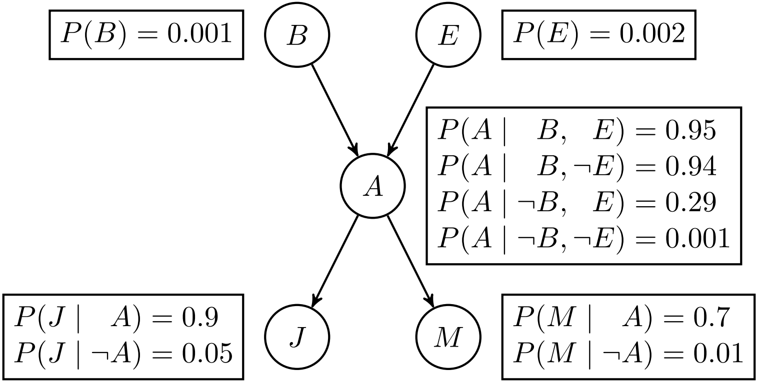
- In general, if there are N variables while evidence contains j variables, how many joints do we need to sum up? $k^{(N-j)}$
- It is this summation that makes inference by enumeration inefficient
- Some computation can be saved by carefully ordering the terms and re-using intermediate results (variable elimination)
- A more complex algorithm called join tree or junction tree can save even more computation
- **The bad news:**
  Exact inference with an arbitrary Bayes Net is intractable

Approximate Inference by Sampling

- Inference can be done approximately by sampling
- General sampling approach:
  1. Generate many, many samples
     (each sample a complete assignment of all variables)
  2. Count the fraction of samples matching query and evidence
  3. As the number of samples approaches ∞, the fraction converges to the posterior $P(Q \mid E)$
- We'll see 3 sampling algorithms (there are more . . .)
  1. Simple sampling
  2. Likelihood weighting
  3. Gibbs sampler

Alg.1: Simple Sampling

- This Bayes Net defines a joint distribution
- Can we generate a set of samples that have the same underlying joint distribution?

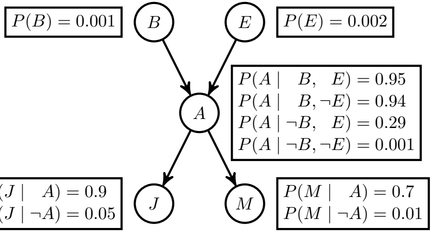


Alg.1: Simple Sampling (cont.)

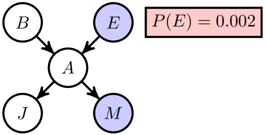
To generate one sample:

1. Sample B: $x =$ rand(0,1). If $(x < 0.001)$ $B = true$ else $B = false$
2. Sample E: $x =$ rand(0,1). If $(x < 0.002)$ $E = true$ else $E = false$
3. If $(B == true$ AND $E == true)$ sample $A \sim \{0.95, 0.05\}$
   elseif $(B == true$ AND $E == false)$ sample $A \sim \{0.94, 0.06\}$
   elseif $(B == false$ AND $E == true)$ sample $A \sim \{0.29, 0.71\}$
   else sample $A \sim \{0.001, 0.999\}$
4. Similarly sample J
5. Similarly sample M

Repeat for more samples



Alg.1: Inference with Simple Sampling

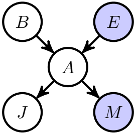


Ex: infer B given E, M i.e. $P(B \mid E, M)$

- First we generate lots of samples
- Keep samples with $E = true$ and $M = true$, toss out the others
- In the N of them that we keep, count the N_1 ones with $B = true$, i.e. those that fit our query
- Return the estimate: $P(B \mid E, M) \approx N_1/N$
- The more samples, the better the estimate
- You should be able to generalize this method to arbitrary BN
- Can you see a problem with simple sampling?

Alg.1: Inference with Simple Sampling (cont.)

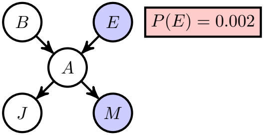
- Since $P(E) = 0.002$,
  we expect only 1 sample out of every 500 to have $E = true$
- We'll throw away 499 samples, a huge waste
- This observation leads to ...



$P(E) = 0.002$

Alg.2: Likelihood Weighting

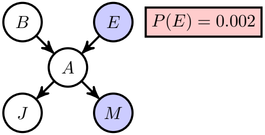
- Say we've generated B and we're about to generate E
- E is an evidence node, known to be true
- Using simple sampling, we will generate
  - $E = true$ 0.2% of the time
  - $E = false$ 99.8% of the time
- Instead, let's always generate $E = true$, but weight the sample down by $P(E) = 0.002$
- Initially the sample has weight $w = 1$, now it $w = w * 0.002$
- We're "virtually throwning away" samples



Alg.2: Likelihood Weighting (cont.)

- Continue and generate A, J as before
- When it's time to generate evidence M from $P(M \mid A)$,
  again always generate $M = true$,
  but weight the sample by $w = w * P(M \mid A)$
- If $A = true$ and $P(M \mid A) = 0.7$, the final weight for this sample is $w = 0.002 * 0.7$
- Repeat and keep all samples, each with a weight: $w_1, \ldots, w_n$
- Return the estimate:

$$P(B \mid E, M) \approx \frac{\sum_{B=true} w_i}{\sum_{all} w_i}$$



Apply this weighting trick every time we generate a value for an evidence node

Alg.3: Gibbs Sampler

- the simplest method in the family of Markov Chain Monte Carlo (MCMC) methods

1. Start from an arbitrary sample, with evidence nodes fixed to their true values,

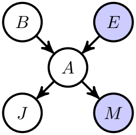
   e.g. $(B = true, E = true, A = false, J = false, M = true)$

2. For each hidden node X, fixing all other nodes, resample its value from $P(X = x \mid \text{Markov-blanket}(X))$,

   e.g.: $B \sim P(B \mid E = true, A = false)$

   Update B with its new sampled value, move on to A, J

3. We now have one new sample. Repeat ...



Alg.3: Gibbs Sampler (cont.)

- Keep all samples: $P(B \mid E, M)$ is the fraction with $B = true$
- In general:

Where $X = x$

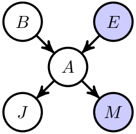
$$P(X = x \mid \text{Markov-blanket}(X)) \propto$$
$$P(X = x \mid \text{parents}(X)) * \prod_{Y_j \in \text{children}(X)} P(y_j \mid \text{parents}(Y_j))$$

- Compute the above for $X = x_1, \ldots, x_k$, then normalize

- More tricks:

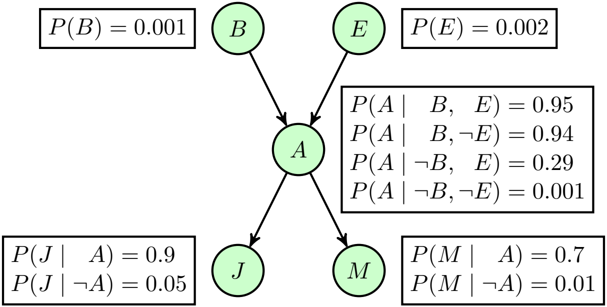
  · 'burn-in': don't use the first n_b samples (e.g. $n_b = 1000$)

  · after burn-in, only use one in every n_s samples (e.g. $n_s = 50$)



Parameter (CPT) Learning for BNs

- Where do you get these CPT values?
  - Ask domain experts, or
  - Learn from data

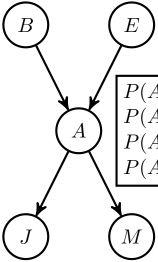


Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

← Given this data,
How do you learn this CPT?

$\boxed{P(B) = 0.001}$ $\boxed{P(E) = 0.002}$



$$P(A \mid B, E) = 0.95$$
$$P(A \mid B, \neg E) = 0.94$$
$$P(A \mid \neg B, E) = 0.29$$
$$P(A \mid \neg B, \neg E) = 0.001$$

$$P(J \mid A) = 0.9$$
$$P(J \mid \neg A) = 0.05$$

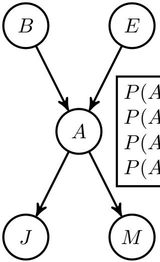
$$P(M \mid A) = 0.7$$
$$P(M \mid \neg A) = 0.01$$

Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

Count $|B|$ and $|\neg B|$ in dataset,
$P(B) = |B| \;/\; (|B| + |\neg B|)$

$\boxed{P(B) = 0.001}$ $\quad$ $\boxed{P(E) = 0.002}$



$$
\begin{aligned}
P(A \mid B, E) &= 0.95 \\
P(A \mid B, \neg E) &= 0.94 \\
P(A \mid \neg B, E) &= 0.29 \\
P(A \mid \neg B, \neg E) &= 0.001
\end{aligned}
$$

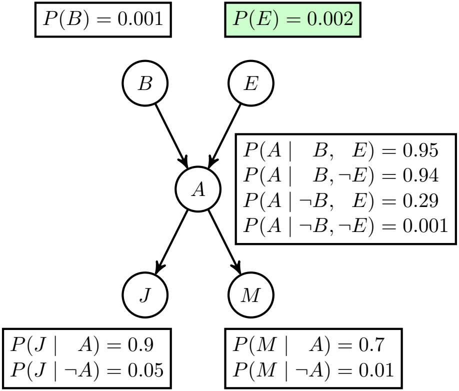
$$
\begin{aligned}
P(J \mid A) &= 0.9 \\
P(J \mid \neg A) &= 0.05
\end{aligned}
$$

$$
\begin{aligned}
P(M \mid A) &= 0.7 \\
P(M \mid \neg A) &= 0.01
\end{aligned}
$$

Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

Count $|E|$ and $|\neg E|$ in dataset,
$P(E) = |E| \;/\; (|E| + |\neg E|)$



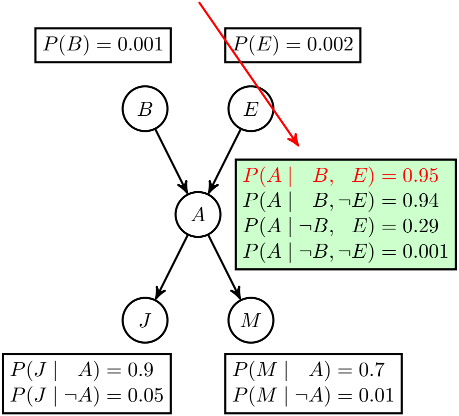
Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

Count $|A|$ and $|\neg A|$ in dataset

where $B = true$, $E = true$,

$P(A \mid B, E) = |A| \ / \ (|A| + |\neg A|)$



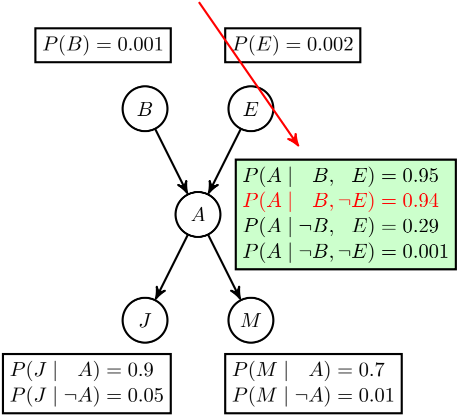
Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

Count $|A|$ and $|\neg A|$ in dataset

where $B = true$, $E = false$,

$P(A \mid B, \neg E) = |A| \ / \ (|A| + |\neg A|)$



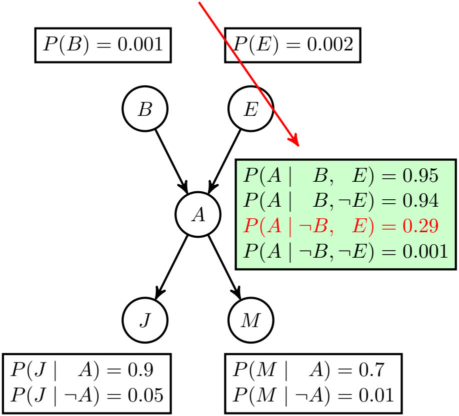
Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

Count $|A|$ and $|\neg A|$ in dataset

where $B = false$, $E = true$,

$P(A \mid \neg B, E) = |A| \ / \ (|A| + |\neg A|)$



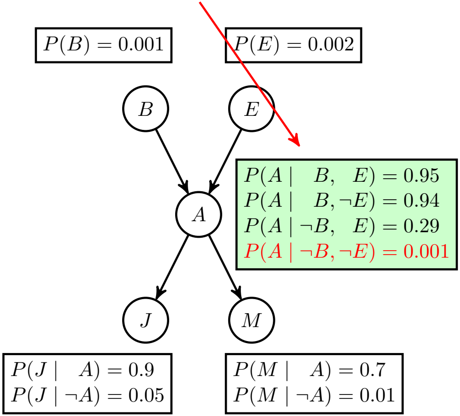
Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

Count $|A|$ and $|\neg A|$ in dataset

where $B = false$, $E = false$,

$P(A \mid \neg B, \neg E) = |A| \;/\; (|A| + |\neg A|)$



Parameter (CPT) Learning for BNs (cont.)

```
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B, ~E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A,  J, ~M)
(~B,  E,  A,  J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
(~B, ~E, ~A, ~J, ~M)
( B,  E,  A, ~J,  M)
(~B, ~E, ~A, ~J, ~M)
...
```

- 'Unseen event' problem
- Going back to:

  Count $|A|$ and $|\neg A|$ in dataset
  where $B = true$, $E = true$,
  $P(A \mid B, E) = |A| \ / \ (|A| + |\neg A|)$
- What if there are no rows with $(B, E, \neg A, *, *)$ in the dataset?
- Do we want to set:
  $P(A \mid B, E) = 1$,
  $P(\neg A \mid B, E) = 0$?
- Why or why not?

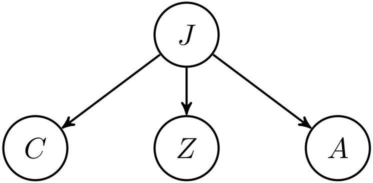
Parameter (CPT) Learning for BNs: Smoothing

- $P(X = x \mid \text{parents}(X)) = (\text{frequency of } x \text{ given parents})$ is called the Maximum Likelihood Estimate (MLE)
- The MLE is vulnerable to the 'unseen event' problem when the dataset is small:
  e.g. flip coin 3 times: all heads $\rightarrow$ one-sided coin?
- 'Add one' smoothing: the simplest solution

Parameter (CPT) Learning for BNs: Smoothing (cont.)

- 'Add one' smoothing: add 1 to all counts
- e.g. Count $|A|$, $|\neg A|$ in dataset where $B = true$, $E = true$
  - $P(A \mid B, E) = (|A|+1) \;/\; (|A|+1 + |\neg A|+1)$
  - If $|A| = 1$, $|\neg A| = 0$:
    - without smoothing: $P(A \mid B, E) = 1$, $P(\neg \mid B, E) = 0$
    - with smoothing: $P(A \mid B, E) = 0.67$, $P(\neg \mid B, E) = 0.33$
  - If $|A| = 100$, $|\neg A| = 0$:
    - without smoothing: $P(A \mid B, E) = 1$, $P(\neg \mid B, E) = 0$
    - with smoothing: $P(A \mid B, E) = 0.99$, $P(\neg \mid B, E) = 0.01$
- Smoothing saves you when you don't have enough data, and hides away when you do
- It's a form of Maximum a posteriori (MAP) estimate

A Special Bayes Net: Naïve Bayes Classifier



J: Person is a junior

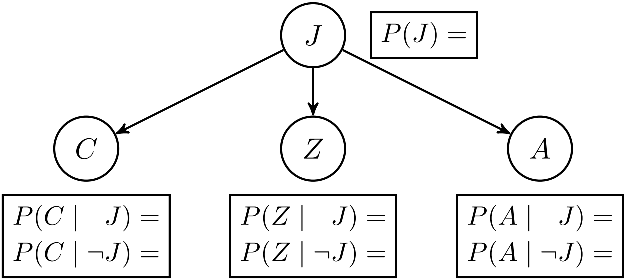
C: Brought coat to class

Z: Lives in 53706

A: Saw Avatar more than once

- What do the CPTs look like?

A Special Bayes Net: Naïve Bayes Classifier (cont.)



J: junior

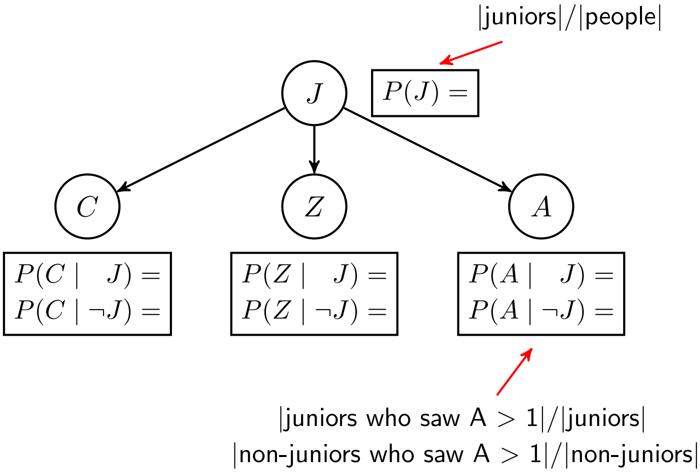
C: coat

Z: 53706

A: Avatar

- Suppose we have dataset of 30 people who attend a lecture.
- How can we use this to estimate the values in the CPTs?

A Special Bayes Net: Naïve Bayes Classifier (cont.)



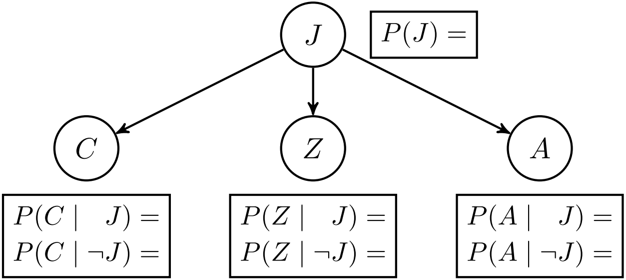
J: junior

C: coat

Z: 53706

A: Avatar

A Special Bayes Net: Naïve Bayes Classifier (cont.)



J: junior

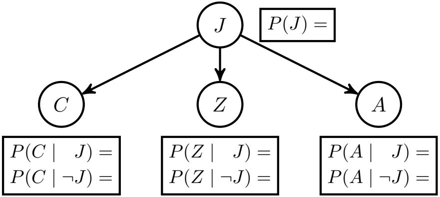
C: coat

Z: 53706

A: Avatar

- A new person shows up wearing a "I live right beside the Union Theater where I saw Avatar every night" jacket
- What's the probability that the person is a Junior?

A Special Bayes Net: Naïve Bayes Classifier (cont.)



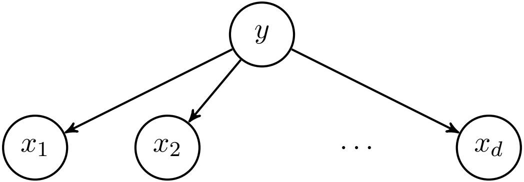
J: junior
C: coat
Z: 53706
A: Avatar

- Input (Evidence, $\mathbf{x}$) : C, Z, A
- Output (Query, y) : J?

$$P(J \mid C, Z, A) = P(J, C, Z, A)/P(C, Z, A)$$
$$= \frac{P(J, C, Z, A)}{[P(J, C, Z, A) + P(\neg J, C, Z, A)]}$$

$$P(J, C, Z, A) = P(J)P(C \mid J)P(Z \mid J)P(A \mid J)$$
$$P(\neg J, C, Z, A) = P(\neg J)P(C \mid \neg J)P(Z \mid \neg J)P(A \mid \neg J)$$

A Special Bayes Net: Naïve Bayes Classifier (cont.)



- Naïve Bayes Classifiers have a special structure:
  - a "class" node y at the root
  - evidence nodes $\mathbf{x}$ (observed features) as leaves
  - conditional independence between all evidence given class
    (strong assumption, usually wrong, but usually empirically ok)

And that's it for now: What you should know ...

- Inference using joint distribution
- Problems with joint distribution
- Bayes Net: representation (nodes,edges,CPTs) and meaning
- How to compute joint probabilities from Bayes Net
- Inference by enumeration
- Inference by sampling
  - simple sampling, likelihood weighting, Gibbs
- CPT parameter learning from data
- Naïve Bayes