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

<table>
<thead>
<tr>
<th>Weather</th>
<th>Temp</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>150/365</td>
</tr>
<tr>
<td>sunny</td>
<td>cold</td>
<td>50/365</td>
</tr>
<tr>
<td>cloudy</td>
<td>hot</td>
<td>40/365</td>
</tr>
<tr>
<td>cloudy</td>
<td>cold</td>
<td>60/365</td>
</tr>
<tr>
<td>rainy</td>
<td>hot</td>
<td>5/365</td>
</tr>
<tr>
<td>rainy</td>
<td>cold</td>
<td>60/365</td>
</tr>
</tbody>
</table>

\[ \frac{365}{365} \]
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: \( P(\text{sunny or hot}) = \frac{150 + 50 + 40 + 5}{365} \)

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

<table>
<thead>
<tr>
<th>Weather</th>
<th>Temp</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>150 365</td>
</tr>
<tr>
<td>sunny</td>
<td>cold</td>
<td>50 365</td>
</tr>
<tr>
<td>cloudy</td>
<td>hot</td>
<td>40 365</td>
</tr>
<tr>
<td>cloudy</td>
<td>cold</td>
<td>60/365</td>
</tr>
<tr>
<td>rainy</td>
<td>hot</td>
<td>5 365</td>
</tr>
<tr>
<td>rainy</td>
<td>cold</td>
<td>60/365</td>
</tr>
</tbody>
</table>
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: \( P(\text{hot} \mid \text{rainy}) = \frac{5}{65} \)

<table>
<thead>
<tr>
<th>Weather</th>
<th>Temp</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>150/365</td>
</tr>
<tr>
<td>sunny</td>
<td>cold</td>
<td>50/365</td>
</tr>
<tr>
<td>cloudy</td>
<td>hot</td>
<td>40/365</td>
</tr>
<tr>
<td>cloudy</td>
<td>cold</td>
<td>60/365</td>
</tr>
<tr>
<td>rainy</td>
<td>hot</td>
<td>5/365</td>
</tr>
<tr>
<td>rainy</td>
<td>cold</td>
<td>60/365</td>
</tr>
</tbody>
</table>
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
With independence, we have

\[
\begin{align*}
P(B, \neg E) &= P(B)P(\neg E) \\
P(\neg B, E) &= P(\neg B)P(E) \\
P(\neg B, \neg E) &= P(\neg B)P(\neg E)
\end{align*}
\]

Say \( P(B) = 0.001, \ P(E) = 0.002, \ P(B \mid E) = P(B) \)

The joint probability table is:

<table>
<thead>
<tr>
<th>Burglary</th>
<th>Earthquake</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B )</td>
<td>( E )</td>
<td></td>
</tr>
<tr>
<td>( B )</td>
<td>( \neg E )</td>
<td></td>
</tr>
<tr>
<td>( \neg B )</td>
<td>( E )</td>
<td></td>
</tr>
<tr>
<td>( \neg B )</td>
<td>( \neg E )</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>B: burglary</td>
<td>0.001</td>
</tr>
<tr>
<td>E: earthquake</td>
<td>0.002</td>
</tr>
<tr>
<td>A: alarm off</td>
<td>0.95</td>
</tr>
<tr>
<td>B, E</td>
<td>0.94</td>
</tr>
<tr>
<td>B, ¬E</td>
<td>0.29</td>
</tr>
<tr>
<td>¬B, E</td>
<td>0.001</td>
</tr>
</tbody>
</table>

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

\[
\begin{align*}
P(B) &= 0.001 & P(A | B, E) &= 0.95 \\
P(E) &= 0.002 & P(A | B, \neg E) &= 0.94 \\
P(B | E) &= P(B) & P(A | \neg B, E) &= 0.29 \\
& & P(A | \neg B, \neg E) &= 0.001
\end{align*}
\]

\[
\begin{align*}
P(A | B, E) &= 0.95 & P(A | B, \neg E) &= 0.94 \\
P(A | \neg B, E) &= 0.29 & P(A | \neg B, \neg E) &= 0.001
\end{align*}
\]
Joint Probability with Bayes Nets

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

Example: \( P(\neg B, E, \neg A) = P(\neg B)P(E \mid \neg B)P(\neg A \mid \neg B, E) \)
\[ = P(\neg B)P(E)P(\neg A \mid \neg B, E) \]
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?

![Bayes Net Diagram](attachment:Bayes_Net_14.png)
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)

  $B$  $E$
  
  $A$
  
  $J$  $M$
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 independent 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!

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

\[
P(J \mid A) = 0.9 \\
P(J \mid \neg A) = 0.05
\]

\[
P(M \mid A) = 0.7 \\
P(M \mid \neg A) = 0.01
\]
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 $\text{parents}(x_i)$ 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 $\text{parents}(x_i)$
   
   3.3 Define the CPT’s for $P(x_i \mid \text{assignment of parents}(x_i))$
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
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
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 descendents are observed
Representing Conditional Independence (cont.)

Tail-to-Tail

Unblocked

Tail-to-Tail

Blocked

Head-to-Tail

Unblocked

Head-to-Tail

Blocked

Head-to-Head

Unblocked

Head-to-Head

Blocked
Representing Conditional Independence: Example

Unblocked

Tail-to-Tail

\[ P(M \mid J) \neq P(M) \]

Head-to-Tail

\[ P(J \mid B) \neq P(J) \]

Head-to-Head

\[ P(E \mid B, A) \neq P(E \mid A) \]

Blocked

\[ P(M \mid J, A) = P(M \mid A) \]

\[ P(J \mid B, A) = P(J \mid A) \]

\[ P(E \mid B, A) \neq P(E \mid A) \]

\[ P(E \mid B) = P(E) \]
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 $C$:

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$
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 conditionally independent of its non-descendents given its parents
- a node is conditionally independent of all other nodes given its Markov Blanket (parents, children, spouses)
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 ($M \ll N$), 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

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

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

For Example:

\[
P(B | J, \neg M)
\]

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

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

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

### Evidence:

- \( P(J \mid A) = 0.9 \)
- \( P(J \mid \neg A) = 0.05 \)
- \( P(M \mid A) = 0.7 \)
- \( P(M \mid \neg A) = 0.01 \)

- \( B, E = 0.002 \)
- \( B, \neg E = 0.29 \)
- \( \neg B, E = 0.95 \)
- \( \neg B, \neg E = 0.94 \)
Inference by Enumeration

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

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

1. Compute \( P(B, J, \neg M) \)
2. Compute \( P(J, \neg M) \)
3. Return \[ \frac{P(B, J, \neg M)}{P(J, \neg M)} \]

\( P(B) = 0.001 \)
\( P(E) = 0.002 \)

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

\[ P(J \mid A) = 0.9 \]
\[ P(J \mid \neg A) = 0.05 \]
\[ P(M \mid A) = 0.7 \]
\[ P(M \mid \neg A) = 0.01 \]
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 $\infty$, 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?

\[
\begin{align*}
P(B) &= 0.001 \\
P(E) &= 0.002 \\
P(A | B, E) &= 0.95 \\
P(A | B, \neg E) &= 0.94 \\
P(A | \neg B, E) &= 0.29 \\
P(A | \neg B, \neg E) &= 0.001 \\
P(J | A) &= 0.9 \\
P(J | \neg A) &= 0.05 \\
P(M | A) &= 0.7 \\
P(M | \neg A) &= 0.01
\end{align*}
\]
Alg. 1: Simple Sampling (cont.)

To generate one sample:

1. Sample $B$: $x = \text{rand}(0,1)$. If $(x < 0.001)$ $B = \text{true}$ else $B = \text{false}$
2. Sample $E$: $x = \text{rand}(0,1)$. If $(x < 0.002)$ $E = \text{true}$ else $E = \text{false}$
3. If $(B == \text{true} \ \text{AND} \ E == \text{true})$ sample $A \sim \{0.95, 0.05\}$
   elseif $(B == \text{true} \ \text{AND} \ E == \text{false})$ sample $A \sim \{0.94, 0.06\}$
   elseif $(B == \text{false} \ \text{AND} \ E == \text{true})$ 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?
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 . . .
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 $w = w \times 0.002$
- We’re “virtually throwing 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 $M = \text{true}$, but weight the sample by $w = w \cdot P(M \mid A)$
- If $A = \text{true}$ and $P(M \mid A) = 0.7$, the final weight for this sample is $w = 0.002 \cdot 0.7$
- Repeat and keep all samples, each with a weight: $w_1, \ldots, w_n$
- Return the estimate:

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

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 …

Diagram:
- \(B\)
- \(E\)
- \(A\)
- \(J\)
- \(M\)
Alg. 3: Gibbs Sampler (cont.)

- Keep all samples: \( P(B \mid E, M) \) is the fraction with \( B = \text{true} \)
- In general:
  \[
P(X = x \mid \text{Markov-blanket}(X)) \propto P(X = x \mid \text{parents}(X)) \times \prod_{Y_j \in \text{children}(X)} P(y_j \mid \text{parents}(Y_j))
\]
- 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

\[
\begin{align*}
P(B) &= 0.001 \\
P(E) &= 0.002 \\
P(A | B, E) &= 0.95 \\
P(A | B, \neg E) &= 0.94 \\
P(A | \neg B, E) &= 0.29 \\
P(A | \neg B, \neg E) &= 0.001 \\
P(J | A) &= 0.9 \\
P(J | \neg A) &= 0.05 \\
P(M | A) &= 0.7 \\
P(M | \neg A) &= 0.01
\end{align*}
\]
Parameter (CPT) Learning for BNs (cont.)

Given this data, How do you learn this CPT?

\[ P(B) = 0.001 \]
\[ P(E) = 0.002 \]

\[ P(A | B, E) = 0.95 \]
\[ P(A | B, \neg E) = 0.94 \]
\[ P(A | \neg B, E) = 0.29 \]
\[ P(A | \neg B, \neg E) = 0.001 \]

\[ P(J | A) = 0.9 \]
\[ P(J | \neg A) = 0.05 \]

\[ P(M | A) = 0.7 \]
\[ P(M | \neg A) = 0.01 \]

\[ (\neg B, \neg E, \neg A, J, \neg M) \]
\[ (\neg B, \neg E, \neg A, \neg J, \neg M) \]
\[ (\neg B, \neg E, \neg A, J, \neg M) \]
\[ (\neg B, \neg E, \neg A, \neg J, \neg M) \]
\[ (\neg B, \neg E, \neg A, J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]
\[ (\neg B, \neg E, \neg A, \neg J, M) \]

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

\[
\begin{align*}
&\neg B, \neg E, \neg A, J, \neg M \\
&\neg B, \neg E, \neg A, \neg J, \neg M \\
&\neg B, \neg E, \neg A, J, \neg M \\
&\neg B, \neg E, \neg A, \neg J, \neg M \\
&\neg B, \neg E, \neg A, J, \neg M \\
&B, \neg E, A, J, M \\
&\neg B, \neg E, \neg A, \neg J, \neg M \\
&\neg B, \neg E, \neg A, J, \neg M \\
&\neg B, \neg E, \neg A, \neg J, \neg M \\
&\neg B, \neg E, \neg A, J, \neg M \\
&B, E, A, J, M \\
\end{align*}
\]

... 

Count \(|B|\) and \(|\neg B|\) in dataset,

\[
P(B) = \frac{|B|}{|B| + |\neg B|}
\]

\[
P(B) = 0.001 \quad P(E) = 0.002
\]

\[
\begin{align*}
P(A | B, E) &= 0.95 \\
P(A | B, \neg E) &= 0.94 \\
P(A | \neg B, E) &= 0.29 \\
P(A | \neg B, \neg E) &= 0.001
\end{align*}
\]

\[
\begin{align*}
P(J | A) &= 0.9 \\
P(J | \neg A) &= 0.05 \\
P(M | A) &= 0.7 \\
P(M | \neg A) &= 0.01
\end{align*}
\]
Parameter (CPT) Learning for BNs (cont.)

\[ \sim B, \sim E, \sim A, J, \sim M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]
\[ \sim B, \sim E, \sim A, J, \sim M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]
\[ B, \sim E, A, J, M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]
\[ \sim B, \sim E, \sim A, J, M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]
\[ \sim B, \sim E, \sim A, J, \sim M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]
\[ B, E, A, J, M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]
\[ \sim B, \sim E, \sim A, J, \sim M \]
\[ B, E, A, \sim J, M \]
\[ \sim B, \sim E, \sim A, \sim J, \sim M \]

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

\[
\begin{align*}
&\sim B, \sim E, \sim A, J, \sim M \\
&\sim B, \sim E, \sim A, \sim J, \sim M \\
&\sim B, \sim E, \sim A, J, \sim M \\
&\sim B, \sim E, \sim A, \sim J, \sim M \\
&( B, \sim E, A, J, M) \\
&\sim B, \sim E, \sim A, \sim J, \sim M \\
&\sim B, \sim E, \sim A, J, \sim M \\
&( B, \sim E, A, J, M) \\
&\sim B, \sim E, \sim A, \sim J, \sim M \\
&( B, \sim E, \sim A, J, M) \\
&( B, \sim E, \sim A, J, \sim M) \\
&( B, \sim E, \sim A, J, M) \\
&( B, \sim E, \sim A, \sim J, \sim M) \\
&( B, \sim E, \sim A, \sim J, \sim M) \\
&( B, \sim E, \sim A, \sim J, \sim M) \\
&( B, \sim E, \sim A, \sim J, M) \\
&( B, \sim E, \sim A, \sim J, M) \\
&\sim B, \sim E, \sim A, \sim J, \sim M \\
&( B, E, A, J, M) \\
&\sim B, \sim E, \sim A, \sim J, \sim M
\end{align*}
\]

Count \(|A|\) and \(|\neg A|\) in dataset where \(B = true, E = true\),
\[
P(A \mid B, E) = \frac{|A|}{|A| + |\neg A|}
\]

\[
\begin{align*}
P(B) &= 0.001 \\
P(E) &= 0.002 \\
P(A \mid B, E) &= 0.95 \\
P(A \mid B, \neg E) &= 0.94 \\
P(A \mid \neg B, E) &= 0.29 \\
P(A \mid \neg B, \neg E) &= 0.001 \\
P(J \mid A) &= 0.9 \\
P(J \mid \neg A) &= 0.05 \\
P(M \mid A) &= 0.7 \\
P(M \mid \neg A) &= 0.01
\end{align*}
\]
Parameter (CPT) Learning for BNs (cont.)

\[(\sim B, \sim E, \sim A, J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(B, \sim E, A, J, M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[(\sim B, \sim E, \sim A, \sim J, \sim M)\]
\[
\text{Count } |A| \text{ and } |\neg A| \text{ in dataset }
\]
\[\text{where } B = true, \ E = false, \]
P\((A \mid B, \neg E) = \frac{|A|}{(|A| + |\neg A|)}\)

\[
\begin{align*}
P(B) &= 0.001 \\
P(E) &= 0.002 \\
P(A \mid B, E) &= 0.95 \\
P(A \mid B, \neg E) &= 0.94 \\
P(A \mid \neg B, E) &= 0.29 \\
P(A \mid \neg B, \neg E) &= 0.001 \\
P(J \mid A) &= 0.9 \\
P(J \mid \neg A) &= 0.05 \\
P(M \mid A) &= 0.7 \\
P(M \mid \neg A) &= 0.01
\end{align*}
\]
Parameter (CPT) Learning for BNs (cont.)

\[(\sim B, \sim E, \sim A, J, \sim M) \quad (\sim B, \sim E, \sim A, \sim J, \sim M) \quad (\sim B, \sim E, \sim A, J, \sim M) \quad (\sim B, \sim E, \sim A, \sim J, \sim M) \quad (B, \sim E, A, J, M) \quad (\sim B, \sim E, \sim A, J, \sim M) \quad (\sim B, \sim E, \sim A, \sim J, M) \quad (\sim B, \sim E, \sim A, \sim J, \sim M) \quad (\sim B, \sim E, \sim A, J, \sim M) \quad (\sim B, \sim E, \sim A, \sim J, \sim M) \quad (\sim B, \sim E, \sim A, J, M) \quad (\sim B, \sim E, \sim A, \sim J, M) \quad (\sim B, \sim E, \sim A, \sim J, \sim M) \quad (B, E, A, J, M) \quad (\sim B, \sim E, \sim A, J, \sim M) \quad (\sim B, \sim E, \sim A, \sim J, \sim M) \]

Count \(|A|\) and \(|\sim A|\) in dataset where \(B = false, E = true\),

\[P(A \mid \sim B, E) = \frac{|A|}{|A| + |\sim A|}\]

\[
\begin{align*}
P(B) &= 0.001 \\
P(E) &= 0.002 \\
P(A \mid B, E) &= 0.95 \\
P(A \mid B, \sim E) &= 0.94 \\
P(A \mid \sim B, E) &= 0.29 \\
P(A \mid \sim B, \sim E) &= 0.001 \\
P(J \mid A) &= 0.9 \\
P(J \mid \sim A) &= 0.05 \\
P(M \mid A) &= 0.7 \\
P(M \mid \sim A) &= 0.01
\end{align*}
\]
Parameter (CPT) Learning for BNs (cont.)

\[ P(B \mid \neg B, \neg E) = \frac{|A|}{(|A| + |\neg A|)} \]

\[ P(B) = 0.001 \quad P(E) = 0.002 \]

\[ P(A \mid B, E) = 0.95 \quad P(A \mid B, \neg E) = 0.94 \]
\[ P(A \mid \neg B, E) = 0.29 \quad P(A \mid \neg B, \neg E) = 0.001 \]

\[ P(J \mid A) = 0.9 \quad P(M \mid A) = 0.7 \]
\[ P(J \mid \neg A) = 0.05 \quad P(M \mid \neg A) = 0.01 \]

...
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)

► ‘Unseen event’ problem

► Going back to:

Count \(|A|\) and \(\neg A\) in dataset

where \(B = true, E = true\),

\[ P(A \mid B, E) = \frac{|A|}{(|A| + |\neg A|)} \]

► 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

- \( P(X = x \mid \text{parents}(X)) = \) (frequency of \( x \) 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
‘Add one’ smoothing: add 1 to all counts

e.g. Count $|A|$, $|\neg A|$ in dataset where $B = true$, $E = true$

$P(A | B, E) = (|A|+1) / (|A|+1 + |\neg A|+1)$

If $|A| = 1$, $|\neg A| = 0$:

- without smoothing: $P(A | B, E) = 1$, $P(\neg | B, E) = 0$
- with smoothing: $P(A | B, E) = 0.67$, $P(\neg | B, E) = 0.33$

If $|A| = 100$, $|\neg A| = 0$:

- without smoothing: $P(A | B, E) = 1$, $P(\neg | B, E) = 0$
- with smoothing: $P(A | B, E) = 0.99$, $P(\neg | B, E) = 0.01$

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.)

Suppose we have a 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.)

\[
P(J) = \frac{|\text{juniors}|}{|\text{people}|}
\]

\[
P(C | J) = \frac{|\text{juniors who saw A}|}{|\text{juniors}|}
\]

\[
P(C | \neg J) = \frac{|\text{non-juniors who saw A}|}{|\text{non-juniors}|}
\]

\[
P(Z | J) = \frac{|\text{juniors}|}{|\text{people}|}
\]

\[
P(Z | \neg J) = \frac{|\text{non-juniors}|}{|\text{non-juniors}|}
\]

\[
P(A | J) = \frac{|\text{juniors who saw A}|}{|\text{juniors}|}
\]

\[
P(A | \neg J) = \frac{|\text{non-juniors who saw A}|}{|\text{non-juniors}|}
\]

\[J: \text{ junior}\]
\[C: \text{ coat}\]
\[Z: 53706\]
\[A: \text{ Avatar}\]
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.)

- **Input (Evidence, x)**: C, Z, A
- **Output (Query, y)**: J?

\[
P(J | C, Z, A) = \frac{P(J, C, Z, A)}{P(C, Z, A)}
\]

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

\[
\]

\[
P(\neg J, C, Z, A) = P(\neg J)P(C | \neg J)P(Z | \neg J)P(A | \neg J)
\]

J: junior
C: coat
Z: 53706
A: Avatar
Naïve Bayes Classifiers have a special structure:

- a “class” node $y$ at the root
- evidence nodes $x$ (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