Lecture 16: Bayes Nets Inference
3/20/2014

Kristen Grauman
UT Austin

Slides courtesy of Dan Klein, UC Berkeley
Survey feedback - thank you!

- Reading/exercise deadline time
- Web page ease of use
- Programming assignments
  - More project debriefing after deadline
  - Contest rankings beyond top 3
  - Some would like less skeleton, more creativity
  - Python programming standards
Survey feedback - thank you!

- Lecture slides – include answers
- Office hours
- Examples in class lecture
- Textbook
Announcements

- Reading/exercise assignments for next week posted – choose one of the 2 exercises and provide reading response
- PS4 out next week
Bayes’ Net Semantics

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over $X$, one for each combination of parents’ values
    \[ P(X|a_1 \ldots a_n) \]
- Bayes’ nets implicitly encode joint distributions
  - As a product of local conditional distributions
    \[ P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]
Example: Alarm Network

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
</tr>
<tr>
<td>-b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>-e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

P(+b, -e, +a, -j, +m) =

P(+b) P(-e) P(+a | +b, -e) P(-j | +a) P(+m | +a) =

0.001 x 0.998 x 0.94 x 0.1 x 0.7
### Example: Alarm Network

#### Probabilities

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
</tr>
<tr>
<td>−b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>−e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

#### Conditional Probabilities

| A | J | P(J|A) |
|---|---|-------|
| +a | +j | 0.9 |
| +a | −j | 0.1 |
| −a | +j | 0.05 |
| −a | −j | 0.95 |

| A | M | P(M|A) |
|---|---|-------|
| +a | +m | 0.7 |
| +a | −m | 0.3 |
| −a | +m | 0.01 |
| −a | −m | 0.99 |

| B | E | A | P(A|B,E) |
|---|---|---|---------|
| +b | +e | +a | 0.95 |
| +b | +e | −a | 0.05 |
| +b | −e | +a | 0.94 |
| +b | −e | −a | 0.06 |
| −b | +e | +a | 0.29 |
| −b | +e | −a | 0.71 |
| −b | −e | +a | 0.001 |
| −b | −e | −a | 0.999 |
Bayes’ Nets

- Representation
- Conditional independences
- Probabilistic inference
  - Enumeration (exact, exponential complexity)
  - Variable elimination (exact, worst-case exponential complexity, often better)
  - Inference is NP-complete
  - Sampling (approximate)
- Learning Bayes’ Nets from data
Inference

- Inference: calculating some useful quantity from a joint probability distribution

- Examples:
  - Posterior probability:
    \[ P(Q|E_1 = e_1, \ldots E_k = e_k) \]
  - Most likely explanation:
    \[ \arg\max_q P(Q = q|E_1 = e_1 \ldots) \]
Recognizing objects in context

Scene

$N_{\text{car}}$

$P(N_{\text{car}} \mid S = \text{street})$

$P(N_{\text{car}} \mid S = \text{park})$

Scene “gist” features

Antonio Torralba
Inference by Enumeration

- Given unlimited time, inference in BNs is easy
- Recipe:
  - State the marginal probabilities you need
  - Figure out ALL the atomic probabilities you need
  - Calculate and combine them
Recall: Inference by Enumeration

- **General case:**
  - Evidence variables: $E_1 \ldots E_k = e_1 \ldots e_k$
  - Query* variable: $Q$
  - Hidden variables: $H_1 \ldots H_r$

- **We want:** $P(Q|e_1 \ldots e_k)$

1. Select the entries consistent with the evidence
2. Sum out H to get joint of Query and evidence:

   \[
P(Q, e_1 \ldots e_k) = \sum_{h_1 \ldots h_r} P(Q, h_1 \ldots h_r, e_1 \ldots e_k) \]

   \[
   \] 

   $X_1, X_2, \ldots X_n$

   All variables

3. Normalize

   \[
   Z = \sum_q P(Q, e_1, \ldots, e_k)
   \]

   \[
   P(Q|e_1, \ldots, e_k) = \frac{1}{Z} \sum_q P(Q, e_1, \ldots, e_k)
   \]

   * Works fine with multiple query variables, too
Example: Enumeration

\[ P(+b | +j, +m) = \frac{P(+b, +j, +m)}{P(+j, +m)} \]

\[ P(+b, +j, +m) = \]
\[ P(+b)P(+e)P(+a | +b, +e)P(+j | +a)P(+m | +a) + \]
\[ P(+b)P(+e)P(-a | +b, +e)P(+j | -a)P(+m | -a) + \]
\[ P(+b)P(-e)P(+a | +b, -e)P(+j | +a)P(+m | +a) + \]
\[ P(+b)P(-e)P(-a | +b, -e)P(+j | -a)P(+m | -a) \]
Inference by Enumeration?
Inference by Enumeration vs. Variable Elimination

- Why is inference by enumeration so slow?
  - You join up the whole joint distribution before you sum out the hidden variables

- Idea: interleave joining and marginalizing!
  - Called “Variable Elimination”
  - Still NP-hard, but usually much faster than inference by enumeration
Factor Zoo I

- **Joint distribution**: $P(X,Y)$
  - Entries $P(x,y)$ for all $x$, $y$
  - Sums to 1

- **Selected joint**: $P(x,Y)$
  - A slice of the joint distribution
  - Entries $P(x,y)$ for fixed $x$, all $y$
  - Sums to $P(x)$

- **Note**: Number of capitals => size of the table
Factor Zoo II

- Family of conditionals: $P(X | Y)$
  - Multiple conditionals
  - Entries $P(x | y)$ for all $x, y$
  - Sums to $|Y|$

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.8</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

- Single conditional: $P(Y | x)$
  - Entries $P(y | x)$ for fixed $x$, all $y$
  - Sums to 1

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Factor Zoo III

- Specified family: $P(y \mid X)$
  - Entries $P(y \mid x)$ for fixed $y$, but for all $x$
  - Sums to … who knows!

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

$P(rain \mid T)$

$P(rain \mid hot)$

$P(rain \mid cold)$
In general, when we write $P(Y_1 \ldots Y_N \mid X_1 \ldots X_M)$

- It is a “factor,” a multi-dimensional array
- Its values are all $P(y_1 \ldots y_N \mid x_1 \ldots x_M)$
- Any assigned $X$ or $Y$ is a dimension missing (selected) from the array
Example: Traffic Domain

- **Random Variables**
  - **R**: Raining
  - **T**: Traffic
  - **L**: Late for class!

**Probability Tables**

- **$P(R)$**
  - $+r$: 0.1
  - $-r$: 0.9

- **$P(T|R)$**
  - $+r$: $+t$: 0.8
  - $+r$: $-t$: 0.2
  - $-r$: $+t$: 0.1
  - $-r$: $-t$: 0.9

- **$P(L|R)$**
  - $+t$: $+l$: 0.3
  - $+t$: $-l$: 0.7
  - $-t$: $+l$: 0.1
  - $-t$: $-l$: 0.9
Variable Elimination Outline

- Track objects called factors
- Initial factors are local CPTs (one per node)

\[
P(R) \quad P(T|R) \quad P(L|T)
\]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td></td>
<td></td>
<td>+t</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td></td>
<td></td>
<td>-t</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td></td>
<td>+l</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Any known values are selected
- E.g. if we know \( L = +\ell \), the initial factors are

\[
P(R) \quad P(T|R) \quad P(\ell|T)
\]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td></td>
<td></td>
<td>+t</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td></td>
<td></td>
<td>-t</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td></td>
<td>+l</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- VE: Alternately join factors and eliminate variables
Operation 1: Join Factors

- First basic operation: **joining factors**
- Combining factors:
  - Get all factors over the joining variable
  - Build a new factor over the union of the variables involved

- Example: Join on R

\[
P(R) \times P(T|R) \rightarrow P(R,T)
\]

\[
\begin{array}{c|c}
+r & 0.1 \\
-r & 0.9 \\
\end{array}
\begin{array}{c|c|c}
+r & +t & 0.8 \\
+r & -t & 0.2 \\
-r & +t & 0.1 \\
-r & -t & 0.9 \\
\end{array}
\begin{array}{c|c|c}
+r & +t & 0.08 \\
+r & -t & 0.02 \\
-r & +t & 0.09 \\
-r & -t & 0.81 \\
\end{array}
\]

- Computation for each entry: pointwise products

\[
\forall r, t : P(r, t) = P(r) \cdot P(t|r)
\]
Example: Multiple Joins

\[
P(R) = \begin{align*}
 & +r & 0.1 \\
 & -r & 0.9 
\end{align*}
\]

\[
P(T | R) = \begin{align*}
 & +r & +t & 0.8 \\
 & +r & -t & 0.2 \\
 & -r & +t & 0.1 \\
 & -r & -t & 0.9 
\end{align*}
\]

\[
P(L | T) = \begin{align*}
 & +t & +l & 0.3 \\
 & +t & -l & 0.7 \\
 & -t & +l & 0.1 \\
 & -t & -l & 0.9 
\end{align*}
\]

Join R

\[
P(R, T) = \begin{align*}
 & +r & +t & 0.08 \\
 & +r & -t & 0.02 \\
 & -r & +t & 0.09 \\
 & -r & -t & 0.81 
\end{align*}
\]

Join T

\[
P(R, T, L) = \begin{align*}
 & +r & +t & +l & 0.024 \\
 & +r & +t & -l & 0.056 \\
 & +r & -t & +l & 0.002 \\
 & +r & -t & -l & 0.018 \\
 & -r & +t & +l & 0.027 \\
 & -r & +t & -l & 0.063 \\
 & -r & -t & +l & 0.081 \\
 & -r & -t & -l & 0.729 
\end{align*}
\]
Operation 2: Eliminate

- Second basic operation: marginalization
- Take a factor and sum out a variable
  - Shrinks a factor to a smaller one
  - A projection operation

Example:

\[ P(R, T) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>-t</td>
<td>0.02</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td>0.09</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td>0.81</td>
</tr>
</tbody>
</table>

\[ \text{sum } R \]

\[ P(T) \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>0.17</td>
</tr>
<tr>
<td>-t</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Multiple Elimination

$P(R, T, L)$

<table>
<thead>
<tr>
<th>+r</th>
<th>+t</th>
<th>+l</th>
<th>0.024</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td>-l</td>
<td>0.056</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td>+l</td>
<td>0.002</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td>-l</td>
<td>0.018</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td>+l</td>
<td>0.027</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td>-l</td>
<td>0.063</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td>+l</td>
<td>0.081</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td>-l</td>
<td>0.729</td>
</tr>
</tbody>
</table>

Sum out R

$P(T, L)$

<table>
<thead>
<tr>
<th>+t</th>
<th>+l</th>
<th>0.051</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>-l</td>
<td>0.119</td>
</tr>
<tr>
<td>-t</td>
<td>+l</td>
<td>0.083</td>
</tr>
<tr>
<td>-t</td>
<td>-l</td>
<td>0.747</td>
</tr>
</tbody>
</table>

Sum out T

$P(L)$

<table>
<thead>
<tr>
<th>+l</th>
<th>0.134</th>
</tr>
</thead>
<tbody>
<tr>
<td>-l</td>
<td>0.886</td>
</tr>
</tbody>
</table>
Marginalizing early! (aka VE)

\[
\begin{align*}
P(R) & = \begin{array}{c|cc}
+ r & 0.1 \\
- r & 0.9 \\
\end{array} \\

P(T|R) & = \begin{array}{c|cc}
+ r & + t & 0.8 \\
+ r & - t & 0.2 \\
- r & + t & 0.1 \\
- r & - t & 0.9 \\
\end{array} \\
P(R,T) & = \begin{array}{c|cc}
+ r & + t & 0.08 \\
+ r & - t & 0.02 \\
- r & + t & 0.09 \\
- r & - t & 0.81 \\
\end{array} \\
Sum\ out\ R & \quad \rightarrow \quad P(T) & = \begin{array}{c}
+ t & 0.17 \\
- t & 0.83 \\
\end{array} \\
Join\ T & \quad \rightarrow \quad P(T,L) & = \begin{array}{c|cc}
+ t & + l & 0.051 \\
+ t & - l & 0.119 \\
- t & + l & 0.083 \\
- t & - l & 0.747 \\
\end{array} \\
Sum\ out\ T & \quad \rightarrow \quad P(L) & = \begin{array}{c}
+ l & 0.134 \\
- l & 0.866 \\
\end{array}
\end{align*}
\]
If evidence, start with factors that select that evidence

- No evidence uses these initial factors:

\[
\begin{align*}
P(R) & \quad P(T|R) & \quad P(L|T) \\
+r & 0.1 & +r & +t & 0.8 & +t & +l & 0.3 \\
-r & 0.9 & +r & -t & 0.2 & +t & -l & 0.7 \\
- & +t & 0.1 & -r & +t & 0.9 & -t & +l & 0.1 \\
- & -t & 0.9 & -r & -t & 0.9 & -t & -l & 0.9
\end{align*}
\]

- Computing \(P(L|+r)\), the initial factors become:

\[
\begin{align*}
P(+r) & \quad P(T|+r) & \quad P(L|T) \\
+r & 0.1 & +r & +t & 0.8 & +t & +l & 0.3 \\
- & +t & 0.2 & +r & -t & 0.2 & +t & -l & 0.7 \\
- & +t & 0.2 & -r & +t & 0.9 & -t & +l & 0.1 \\
- & -t & 0.9 & -r & -t & 0.9 & -t & -l & 0.9
\end{align*}
\]

- We eliminate all vars other than query + evidence
Result will be a selected joint of query and evidence

- E.g. for $P(L \mid +r)$, we’d end up with:

\[
P(+r, L)
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+l</td>
<td>0.026</td>
</tr>
<tr>
<td>+r</td>
<td>-l</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Normalize

\[
P(L \mid +r)
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+l</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>-l</td>
<td>0.74</td>
</tr>
</tbody>
</table>

To get our answer, just normalize this!

That’s it!
General Variable Elimination

- **Query:** $P(Q|E_1 = e_1, \ldots E_k = e_k)$

- **Start with initial factors:**
  - Local CPTs (but instantiated by evidence)

- **While there are still hidden variables (not Q or evidence):**
  - Pick a hidden variable $H$
  - Join all factors mentioning $H$
  - Eliminate (sum out) $H$

- **Join all remaining factors and normalize**
Example

\[ P(B|j, m) \propto P(B, j, m) \]

| P(B) | P(E) | P(A|B, E) | P(j|A) | P(m|A) |

Choose A

\[
\begin{align*}
P(A|B, E) \\
P(j|A) \\
P(m|A)
\end{align*}
\]

\[ \times \]

\[
\begin{align*}
P(j, m, A|B, E) \quad \sum \quad P(j, m|B, E)
\end{align*}
\]

| P(B) | P(E) | P(j, m|B, E) |
Query:  $P(B|j, m)$

Example (continued)

Choose E

$$P(E) \times P(j, m|B, E) \sum P(j, m|B)$$

Finish with B

$$P(B) \times P(j, m, B) \rightarrow \text{Normalize} \rightarrow P(B|j, m)$$
### Same example in equations

\[ P(B|j, m) \propto P(B, j, m) \]

| \( P(B) \) | \( P(E) \) | \( P(A|B, E) \) | \( P(j|A) \) | \( P(m|A) \) |
|----------------|----------------|----------------|----------------|----------------|
| \( P(B|j, m) \) | \( P(B, j, m) \) |                           |                           |                           |
| \( = \sum_{e,a} P(B, j, m, e, a) \) |                           |                           |                           |                           |
| \( = \sum_{e,a} P(B)P(e)P(a|B, e)P(j|a)P(m|a) \) |                           |                           |                           |                           |
| \( = \sum_{e} P(B)P(e)\sum_{a} P(a|B, e)P(j|a)P(m|a) \) |                           |                           |                           |                           |
| \( = \sum_{e} P(B)P(e)f_1(B, e, j, m) \) |                           |                           |                           |                           |
| \( = P(B)\sum_{e} P(e)f_1(B, e, j, m) \) |                           |                           |                           |                           |
| \( = P(B)f_2(B, j, m) \) |                           |                           |                           |                           |

- **marginal can be obtained from joint by summing out**
- **use Bayes’ net joint distribution expression**
- **use \( x*(y+z) = xy + xz \)**
- **joining on \( a \), and then summing out gives \( f_1 \)**
- **\( x*(y+z) = xy + xz \)**
- **joining on \( e \), and then summing out gives \( f_2 \)**

We are exploiting:

\[ uwz + uwy + uxy + uxz + vwy + vwz + vxy + vxz = (u + v)(w + x)(y+z) \]
Another variable elimination example

Query: \( P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \)

Start by inserting evidence, which gives the following initial factors:

\[
p(Z)p(X_1|Z)p(X_2|Z)p(X_3|Z)p(y_1|X_1)p(y_2|X_2)p(y_3|X_3)
\]

Eliminate \( X_1 \), this introduces the factor \( f_1(Z, y_1) = \sum_{x_1} p(x_1|Z)p(y_1|x_1) \), and we are left with:

\[
p(Z)f_1(Z, y_1)p(X_2|Z)p(X_3|Z)p(y_2|X_2)p(y_3|X_3)
\]

Eliminate \( X_2 \), this introduces the factor \( f_2(Z, y_2) = \sum_{x_2} p(x_2|Z)p(y_2|x_2) \), and we are left with:

\[
p(Z)f_1(Z, y_1)f_2(Z, y_2)p(X_3|Z)p(y_3|X_3)
\]

Eliminate \( Z \), this introduces the factor \( f_3(y_1, y_2, X_3) = \sum_z p(z)f_1(z, y_1)f_2(z, y_2)p(X_3|z) \), and we are left:

\[
p(y_3|X_3), f_3(y_1, y_2, X_3)
\]

No hidden variables left. Join the remaining factors to get:

\[
f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3) f_3(y_1, y_2, X_3).
\]

Normalizing over \( X_3 \) gives \( P(X_3|y_1, y_2, y_3) \).
Another variable elimination example

- Computational complexity depends on largest factor being generated.
- Size of factor = number of entries in table
- In previous example, assuming all binary variables, all factors are of size 2 – they all have only one variable (Z, Z, and X3, respectively)
Quiz: Variable elimination ordering

For the query $P(X_n \mid y_1, \ldots, y_n)$, what would be a **good** and **bad** ordering for elimination?
VE: Computational and space complexity

- Determined by the largest factor
- Elimination ordering can greatly affect the size of the largest factor
  - e.g., previous example, $2^n$ vs $2^2$.
- Does there always exist an ordering that’s good?
  - No.
Recap: Bayes’ Nets

- Representation
- Conditional independences
  - Probabilistic inference
    - Enumeration (exact, exponential complexity)
    - Variable elimination (exact, worst-case exponential complexity, often better)
  - Inference is NP-complete
    - Sampling (approximate)
- Learning Bayes’ Nets from data