Bayes Nets for representing and reasoning about uncertainty

Andrew W. Moore
Associate Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew’s tutorials: http://www.cs.cmu.edu/~awm/tutorials.

Comments and corrections gratefully received.

What we’ll discuss

• Recall the numerous and dramatic benefits of Joint Distributions for describing uncertain worlds
• Reel with terror at the problem with using Joint Distributions
• Discover how Bayes Net methodology allows us to built Joint Distributions in manageable chunks
• Discover there’s still a lurking problem…
• …Start to solve that problem
Ways to deal with Uncertainty

• Three-valued logic: True / False / Maybe
• Fuzzy logic (truth values between 0 and 1)
• Non-monotonic reasoning (especially focused on Penguin informatics)
• Dempster-Shafer theory (and an extension known as quasi-Bayesian theory)
• Possibilistic Logic
• Probability

Discrete Random Variables

• A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
• Examples
  • A = The US president in 2023 will be male
  • A = You wake up tomorrow with a headache
  • A = You have Ebola
Probabilities

- We write $P(A)$ as “the fraction of possible worlds in which $A$ is true”
- We could at this point spend 2 hours on the philosophy of this.
- But we won’t.

Visualizing $A$

- Event space of all possible worlds
- Its area is 1
- $P(A) =$ Area of reddish oval

Worlds in which $A$ is False

Worlds in which $A$ is True
Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of $A$ can't get any smaller than 0

And a zero area would mean no world could ever have $A$ true

---

Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of $A$ can't get any bigger than 1

And an area of 1 would mean all worlds will have $A$ true
Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Simple addition and subtraction
These Axioms are Not to be Trifled With

- There have been attempts to do different methodologies for uncertainty
  - Fuzzy Logic
  - Three-valued logic
  - Dempster-Shafer
  - Non-monotonic reasoning

- But the axioms of probability are the only system with this property:
  If you gamble using them you can’t be unfairly exploited by an opponent using some other system [di Finetti 1931]

Theorems from the Axioms

- 0 <= P(A) <= 1, P(True) = 1, P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)
  From these we can prove:
  \[ P(\text{not } A) = P(\sim A) = 1 - P(A) \]

- How?
Side Note

• I am inflicting these proofs on you for two reasons:
  1. These kind of manipulations will need to be second nature to you if you use probabilistic analytics in depth
  2. Suffering is good for you

Another important theorem

• $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
• $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  From these we can prove:
  $P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B)$

• How?
Conditional Probability

- $P(A|B) = \text{Fraction of worlds in which } B \text{ is true that also have } A \text{ true}$

H = “Have a headache”
F = “Coming down with Flu”

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

“Headaches are rare and flu is rarer, but if you’re coming down with ‘flu there’s a 50-50 chance you’ll have a headache.”

P(H|F) = Fraction of flu-inflicted worlds in which you have a headache

= $\frac{\#\text{worlds with flu and headache}}{\#\text{worlds with flu}} = \frac{\text{Area of “H and F” region}}{\text{Area of “F” region}}$

= $\frac{P(H \cap F)}{P(F)}$
Definition of Conditional Probability

\[ P(A \cap B) \]
\[ P(A|B) = \frac{\text{---------}}{P(B)} \]

Corollary: The Chain Rule

\[ P(A \cap B) = P(A|B) P(B) \]

Bayes Rule

\[ \frac{P(A \cap B)}{P(B|A)} = \frac{P(A|B) P(B)}{P(A)} = \frac{\text{---------}}{P(A)} \]

This is Bayes Rule

An easy fact about Multivalued Random Variables:

• Using the axioms of probability...
  \[ 0 \leq P(A) \leq 1, \quad P(\text{True}) = 1, \quad P(\text{False}) = 0 \]
  \[ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \]

• And assuming that A obeys...
  \[ P(A = v_i \land A = v_j) = 0 \text{ if } i \neq j \]
  \[ P(A = v_1 \lor A = v_2 \lor A = v_k) = 1 \]

• It’s easy to prove that
  \[ P(A = v_1 \lor A = v_2 \lor A = v_i) = \sum_{j=1}^{i} P(A = v_j) \]

An easy fact about Multivalued Random Variables:

• Using the axioms of probability...
  \[ 0 \leq P(A) \leq 1, \quad P(\text{True}) = 1, \quad P(\text{False}) = 0 \]
  \[ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \]

• And assuming that A obeys...
  \[ P(A = v_i \land A = v_j) = 0 \text{ if } i \neq j \]
  \[ P(A = v_1 \lor A = v_2 \lor A = v_k) = 1 \]

• It’s easy to prove that
  \[ P(A = v_1 \lor A = v_2 \lor A = v_i) = \sum_{j=1}^{i} P(A = v_j) \]

• And thus we can prove
  \[ \sum_{j=1}^{k} P(A = v_j) = 1 \]
Another fact about Multivalued Random Variables:

- Using the axioms of probability...
  
  0 \leq P(A) \leq 1, P(\text{True}) = 1, P(\text{False}) = 0
  
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)
- And assuming that A obeys...
  
  \[ P(A = v_i \land A = v_j) = 0 \text{ if } i \neq j \]
  
  \[ P(A = v_1 \lor A = v_2 \lor A = v_k) = 1 \]
- It’s easy to prove that
  
  \[ P(B \land [A = v_1 \lor A = v_2 \lor A = v_j]) = \sum_{j=1}^{i} P(B \land A = v_j) \]

And thus we can prove

\[ P(B) = \sum_{j=1}^{k} P(B \land A = v_j) \]
Useful Easy-to-prove facts

\[ P(A \mid B) + P(\neg A \mid B) = 1 \]

\[ \sum_{k=1}^{n_A} P(A = v_k \mid B) = 1 \]

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have \(2^M\) rows).

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

2. For each combination of values, say how probable it is.
The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have $2^M$ rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

**Example: Boolean variables A, B, C**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Using the Joint

Once you have the JD you can ask for the probability of any logical expression involving your attribute

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

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>P(wealth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0.40.5-</td>
<td>poor</td>
<td>0.253122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.024595</td>
</tr>
<tr>
<td></td>
<td>v1.40.5+</td>
<td>poor</td>
<td>0.0421760</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>v0.40.5-</td>
<td>poor</td>
<td>0.331331</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
</tr>
<tr>
<td></td>
<td>v1.40.5+</td>
<td>poor</td>
<td>0.134106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
</tr>
</tbody>
</table>
Using the Joint

\[ P(\text{Poor Male}) = 0.4654 \]

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

Using the Joint

\[ P(\text{Poor}) = 0.7604 \]

\[ P(E) = \sum_{\text{rows matching } E} P(\text{row}) \]
Inference with the Joint

\[ P(E_1 | E_2) = \frac{P(E_1 \land E_2)}{P(E_2)} = \frac{\sum P(\text{row})}{\sum P(\text{row})} \]

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.253122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0245895</td>
</tr>
<tr>
<td>v1:40.5+</td>
<td></td>
<td>poor</td>
<td>0.0421766</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.331313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
</tr>
<tr>
<td>v1:40.5+</td>
<td></td>
<td>poor</td>
<td>0.134106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
</tr>
</tbody>
</table>

Inference with the Joint

\[ P(E_1 | E_2) = \frac{P(E_1 \land E_2)}{P(E_2)} = \frac{\sum P(\text{row})}{\sum P(\text{row})} \]

\[ P(\text{Male} | \text{Poor}) = \frac{0.4654}{0.7604} = 0.612 \]
Joint distributions

• Good news
  Once you have a joint distribution, you can ask important questions about stuff that involves a lot of uncertainty

• Bad news
  Impossible to create for more than about ten attributes because there are so many numbers needed when you build the damn thing.

Using fewer numbers

Suppose there are two events:
• M: Manuela teaches the class (otherwise it’s Andrew)
• S: It is sunny
The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we'll have to invent those four numbers. OR WILL WE??
• We don’t have to specify with bottom level conjunctive events such as P(~M^S) IF…
• …instead it may sometimes be more convenient for us to specify things like: P(M), P(S).
But just P(M) and P(S) don’t derive the joint distribution. So you can’t answer all questions.
Using fewer numbers

Suppose there are two events:
- M: Manuela teaches the class (otherwise it’s Andrew)
- S: It is sunny

The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we’ll have to invent those four numbers. OR WILL WE??
- We don't have to specify with bottom level conjunctive events such as P(~M^S) IF…
- … instead it may sometimes be more convenient for us to specify things like: P(M), P(S).

But just P(M), P(S) don’t derive the joint distribution. So what extra assumption can you make?

Independence

“The sunshine levels do not depend on and do not influence who is teaching.”

This can be specified very simply:

\[ P(S \mid M) = P(S) \]

This is a powerful statement!

It required extra domain knowledge. A different kind of knowledge than numerical probabilities. It needed an understanding of causation.
Independence

From \( P(S \mid M) = P(S) \), the rules of probability imply: (can you prove these?)

- \( P(\neg S \mid M) = P(\neg S) \)
- \( P(M \mid S) = P(M) \)
- \( P(M \wedge S) = P(M) P(S) \)
- \( P(\neg M \wedge S) = P(\neg M) P(S), \quad (PM^{\neg S}) = P(M)P(\neg S), \quad P(\neg M^{\neg S}) = P(\neg M)P(\neg S) \)

And in general:

\[
P(M=u \wedge S=v) = P(M=u) P(S=v)
\]

for each of the four combinations of

- \( u=True/False \)
- \( v=True/False \)
Independence

We’ve stated:

\[
\begin{align*}
P(M) &= 0.6 \\
P(S) &= 0.3 \\
P(S \mid M) &= P(S)
\end{align*}
\]

From these statements, we can derive the full joint pdf.

<table>
<thead>
<tr>
<th>M</th>
<th>S</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>

And since we now have the joint pdf, we can make any queries we like.

A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.
A more interesting case

• M : Manuela teaches the class
• S : It is sunny
• L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

Let’s begin with writing down knowledge we’re happy about:

\[ P(S \mid M) = P(S), \quad P(S) = 0.3, \quad P(M) = 0.6 \]

Lateness is not independent of the weather and is not independent of the lecturer.

We already know the Joint of S and M, so all we need now is

\[ P(L \mid S=u, M=v) \]

in the 4 cases of u/v = True/False.
A more interesting case

• M : Manuela teaches the class
• S : It is sunny
• L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

\[
\begin{align*}
P(S | M) &= P(S) \\
P(S) &= 0.3 \\
P(M) &= 0.6
\end{align*}
\]

\[
\begin{align*}
P(L | M \land S) &= 0.05 \\
P(L | M \land \neg S) &= 0.1 \\
P(L | \neg M \land S) &= 0.1 \\
P(L | \neg M \land \neg S) &= 0.2
\end{align*}
\]

Now we can derive a full joint p.d.f. with a “mere” six numbers instead of seven.*

*Savings are larger for larger numbers of variables.

---

A more interesting case

• M : Manuela teaches the class
• S : It is sunny
• L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

\[
\begin{align*}
P(S | M) &= P(S) \\
P(S) &= 0.3 \\
P(M) &= 0.6
\end{align*}
\]

\[
\begin{align*}
P(L | M \land S) &= 0.05 \\
P(L | M \land \neg S) &= 0.1 \\
P(L | \neg M \land S) &= 0.1 \\
P(L | \neg M \land \neg S) &= 0.2
\end{align*}
\]

Question: Express

\[P(L=x \land M=y \land S=z)\]

in terms that only need the above expressions, where \(x, y\) and \(z\) may each be True or False.
A bit of notation

\[ P(S \mid M) = P(S) \]
\[ P(S) = 0.3 \]
\[ P(M) = 0.6 \]

\[ P(L \mid M \land S) = 0.05 \]
\[ P(L \mid M \land \sim S) = 0.1 \]
\[ P(L \mid \sim M \land S) = 0.1 \]
\[ P(L \mid \sim M \land \sim S) = 0.2 \]

Read the absence of an arrow between S and M to mean “it would not help me predict M if I knew the value of S.”

Read the two arrows into L to mean that if I want to know the value of L it may help me to know M and to know S.

This kind of stuff will be thoroughly formalized later.
An even cuter trick

Suppose we have these three events:
• M : Lecture taught by Manuela
• L : Lecturer arrives late
• R : Lecture concerns robots

Suppose:
• Andrew has a higher chance of being late than Manuela.
• Andrew has a higher chance of giving robotics lectures.

What kind of independence can we find?

How about:
• \( P(L \mid M) = P(L) \) ?
• \( P(R \mid M) = P(R) \) ?
• \( P(L \mid R) = P(L) \) ?

Conditional independence

Once you know who the lecturer is, then whether they arrive late doesn’t affect whether the lecture concerns robots.

\[
P(R \mid M,L) = P(R \mid M) \text{ and } P(R \mid ~M,L) = P(R \mid ~M)
\]

We express this in the following way:

“R and L are conditionally independent given M”
Conditional Independence formalized

R and L are conditionally independent given M if for all x,y,z in {T,F}:

\[ P(R=x \mid M=y \land L=z) = P(R=x \mid M=y) \]

More generally:

Let S1 and S2 and S3 be sets of variables.

Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 if for all assignments of values to the variables in the sets,

\[ P(S_1's \text{ assignments} \mid S_2's \text{ assignments} \land S_3's \text{ assignments}) = P(S_1's \text{ assignments} \mid S_3's \text{ assignments}) \]

Example:

"Shoe-size is conditionally independent of Glove-size given height weight and age"

means

for all s,g,h,w,a

\[ P(\text{ShoeSize}=s\mid \text{Height}=h,\text{Weight}=w,\text{Age}=a) = P(\text{ShoeSize}=s\mid \text{Height}=h,\text{Weight}=w,\text{Age}=a,\text{GloveSize}=g) \]
Example:

R and L are conditionally independent given M if for all x, y, z in \{T, F\}:

\[
P(R = x \mid M = y \land L = z) = P(R = x \mid M = y)
\]

More generally:

Let S1 and S2 and S3 be sets of variables.

Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 if for all assignments of values to the variables in the sets,

\[
P(S_1's \text{ assignments} \mid S_2's \text{ assignments} \land S_3's \text{ assignments}) = P(S_1's \text{ assignments} \mid S_3's \text{ assignments})
\]

“Shoe-size is conditionally independent of Glove-size given height weight and age” does not mean for all s, g, h

\[
P(\text{ShoeSize} = s | \text{Height} = h) = P(\text{ShoeSize} = s | \text{Height} = h, \text{GloveSize} = g)
\]

Conditional independence

We can write down \(P(M)\). And then, since we know L is only directly influenced by M, we can write down the values of \(P(L \mid M)\) and \(P(L \mid \neg M)\) and know we’ve fully specified L’s behavior. Ditto for R.

\[
\begin{align*}
P(M) &= 0.6 \\
P(L \mid M) &= 0.085 \\
P(L \mid \neg M) &= 0.17 \\
P(R \mid M) &= 0.3 \\
P(R \mid \neg M) &= 0.6
\end{align*}
\]

‘R and L conditionally independent given M’
Conditional independence

P(M) = 0.6
P(L | M) = 0.085
P(L | ~M) = 0.17
P(R | M) = 0.3
P(R | ~M) = 0.6

Conditional Independence:
P(R | M, L) = P(R | M),
P(R | ~M, L) = P(R | ~M)

Again, we can obtain any member of the Joint prob dist that we desire:
P(L=x ^ R=y ^ M=z) =

Assume five variables

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny

• T only directly influenced by L (i.e. T is conditionally independent of R, M, S given L)
• L only directly influenced by M and S (i.e. L is conditionally independent of R given M & S)
• R only directly influenced by M (i.e. R is conditionally independent of L, S, given M)
• M and S are independent
Making a Bayes net

Step One: add variables.
• Just choose the variables you’d like to be included in the net.

Step Two: add links.
• The link structure must be acyclic.
• If node X is given parents Q₁,Q₂,…Qₙ you are promising that any variable that’s a non-descendent of X is conditionally independent of X given {Q₁,Q₂,…Qₙ}.

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny
Step Three: add a probability table for each node.
- The table for node X must list P(X|Parent Values) for each possible combination of parent values.

- Two unconnected variables may still be correlated
- Each node is conditionally independent of all non-descendants in the tree, given its parents.
- You can deduce many other conditional independence relations from a Bayes net. See the next lecture.