Artificial Intelligence

Lesson 11
(From Russell & Norvig)

168
Ram Meshulam 2004

Conditional probability

• Conditional or posterior probabilities
e.g., \( P(\text{cavity} \mid \text{toothache}) = 0.8 \)
i.e., given that toothache is all I know

• Notation for conditional distributions:
\( P(\text{Cavity} \mid \text{Toothache}) = 2\)-element vector of 2-element vectors

• If we know more, e.g., cavity is also given, then we have
\( P(\text{cavity} \mid \text{toothache}, \text{cavity}) = 1 \)

• New evidence may be irrelevant, allowing simplification, e.g.,
\( P(\text{cavity} \mid \text{toothache}, \text{sunny}) = P(\text{cavity} \mid \text{toothache}) = 0.8 \)

• This kind of inference, sanctioned by domain knowledge, is crucial

Inference by enumeration

• Start with the joint probability distribution:

<table>
<thead>
<tr>
<th>cavity</th>
<th>~cavity</th>
<th>toothache</th>
<th>~toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>tooth</td>
<td>~tooth</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>cavity</td>
<td>~cavity</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>~cavity</td>
<td>tooth</td>
<td>.016</td>
<td>.054</td>
</tr>
<tr>
<td>~cavity</td>
<td>~cavity</td>
<td>.144</td>
<td>.576</td>
</tr>
</tbody>
</table>

• Can also compute conditional probabilities:
\[
P(\sim\text{cavity} \mid \text{toothache}) = \frac{P(\sim\text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]

Independence

• \( A \) and \( B \) are independent iff
\( P(A|B) = P(A) \) or \( P(B|A) = P(B) \) or \( P(A, B) = P(A)P(B) \)

\( P(\text{Toothache, Catch, Cavity, Weather}) \)
\( = P(\text{Toothache, Catch, Cavity})P(\text{Weather}) \)

• 32 entries reduced to 12; for \( n \) independent biased coins, \( O(2^n) \rightarrow O(n) \)
• Absolute independence powerful but rare
• Dentistry is a large field with hundreds of variables, none of which are independent. What to do?
Conditional independence

- $P(\text{Toothache}, \text{Cavity}, \text{Catch})$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  \begin{equation}
  P(\text{catch} | \text{toothache}, \text{cavity}) = P(\text{catch} | \text{cavity})
  \end{equation}
- The same independence holds if I haven't got a cavity:
  \begin{equation}
  P(\text{catch} | \text{toothache}, \neg \text{cavity}) = P(\text{catch} | \neg \text{cavity})
  \end{equation}
- Catch is conditionally independent of Toothache given Cavity:
  \begin{equation}
  P(\text{Catch} | \text{Toothache}, \text{Cavity}) = P(\text{Catch} | \text{Cavity})
  \end{equation}

- Equivalent statements:
  \begin{equation}
  P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})
  \end{equation}

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- It describes how variables interact locally
- Local interactions chain together to give global, indirect interactions

- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:
    \begin{equation}
    P(X_i | \text{Parents}(X_i)) - \text{conditional probability table (CPT)}
    \end{equation}

Example 1

- Topology of network encodes conditional independence assertions:
  \begin{itemize}
  \item Weather is independent of the other variables
  \item Toothache and Catch are conditionally independent given Cavity
  \item It is usually easy for a domain expert to decide what direct influences exist
  \end{itemize}

Example 2

- $N$ independent coin flips:
  \begin{itemize}
  \item No interactions between variables: absolute independence
  \item Does every Bayes Net can represent every full joint?
  \item No. For example, Only distributions whose variables are absolutely independent can be represented by a Bayes’ net with no arcs.
  \end{itemize}
Calculation of Joint Probability

- Given its parents, each node is conditionally independent of everything except its descendants
- Thus,

\[ P(x_1 \land x_2 \land \ldots \land x_n) = \prod_{i=1}^{1,\ldots,n} P(x_i | \text{parents}(X_i)) \]

⇒ full joint distribution table
- Every BN over a domain implicitly represents some joint distribution over that domain

Example 3

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.

Example 3 contd.

Answering queries

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
  - Do the same to calculate \( P(\neg b|\neg j, \neg m) \) and normalize
  - Worst case, for a network with \( n \) Boolean variables, \( O(n^2) \).
Laziness and Ignorance

- The probabilities actually summarize a potentially infinite set of circumstances in which the alarm might fail to go off
  - high humidity
  - power failure
  - dead battery
  - cut wires
  - a dead mouse stuck inside the bell
- John or Mary might fail to call and report it
  - out to lunch
  - on vacation
  - temporarily deaf
  - passing helicopter

Compactness

- A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values
- Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1-p$)
- If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)
- We utilize the property of **locally structured system**

Casualty?

- **Rain** causes **Traffic**
- Let’s build the joint:

<table>
<thead>
<tr>
<th>$P(R)$</th>
<th>$P(T,R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>1/4</td>
</tr>
<tr>
<td>$\neg r$</td>
<td>3/4</td>
</tr>
</tbody>
</table>


| $P(T|R)$ |
|----------|
| $r$     |
| $-r$    |
| $t$     | 3/16     |
| $-t$    | 1/16     |


<table>
<thead>
<tr>
<th>$P(T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
</tr>
<tr>
<td>$\neg r$</td>
</tr>
<tr>
<td>$t$</td>
</tr>
<tr>
<td>$-t$</td>
</tr>
</tbody>
</table>


| $P(R|T)$ |
|----------|
| $r$     |
| $\neg r$ |
| $t$     | 3/4     |
| $-t$    | 1/4     |


<table>
<thead>
<tr>
<th>$P(R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
</tr>
<tr>
<td>$\neg r$</td>
</tr>
<tr>
<td>$t$</td>
</tr>
<tr>
<td>$-t$</td>
</tr>
</tbody>
</table>


| $P(T|R)$ |
|----------|
| $r$     |
| $\neg r$ |
| $t$     | 1/3     |
| $-t$    | 2/3     |


<table>
<thead>
<tr>
<th>$P(R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
</tr>
<tr>
<td>$\neg r$</td>
</tr>
<tr>
<td>$t$</td>
</tr>
<tr>
<td>$-t$</td>
</tr>
</tbody>
</table>


Reverse Casualty?

<table>
<thead>
<tr>
<th>$P(T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
</tr>
<tr>
<td>$\neg r$</td>
</tr>
<tr>
<td>$t$</td>
</tr>
<tr>
<td>$-t$</td>
</tr>
</tbody>
</table>


| $P(R|T)$ |
|----------|
| $r$     |
| $\neg r$ |
| $t$     | 3/4     |
| $-t$    | 1/4     |


<table>
<thead>
<tr>
<th>$P(R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
</tr>
<tr>
<td>$\neg r$</td>
</tr>
<tr>
<td>$t$</td>
</tr>
<tr>
<td>$-t$</td>
</tr>
</tbody>
</table>
Casualty?

• What do the arrows really mean?
• Topology may happen to encode causal structure
• Topology really encodes conditional independencies

• When Bayes’ nets reflect the true causal patterns:
  – Often simpler (nodes have fewer parents)
  – Often easier to think about
  – Often easier to elicit from experts
• BNs need not actually be causal
  – Sometimes no causal net exists over the domain
  – E.g. consider the variables Traffic and RoofDrips
  – End up with arrows that reflect correlation, not causation

Example 2, Again

Consider the following 2 orders for insertion:
• (a) MaryCalls, JohnCalls, Alarm, Burglary, Earthquake
  – Since, $P(\text{Burglary}|\text{Alarm, JohnCalls, MaryCalls}) = P(\text{Burglary}|\text{Alarm})$
• (b) Mary Calls, JohnCalls, Earthquake, Burglary, Alarm.

Connection Types

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual chain</td>
<td>$X \rightarrow Y \rightarrow Z$</td>
<td>Not necessarily</td>
<td>Yes</td>
</tr>
<tr>
<td>Common Cause</td>
<td>$Y \rightarrow X \rightarrow Z$</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Common Effect</td>
<td>$X \rightarrow Y \rightarrow Z$</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Test Question

$P(H=\text{true}) = 0.1$

| $H$ | $P(G=\text{true}|H)$ |
|-----|----------------------|
| F   | 0.4                  |
| F   | 0.8                  |

$H$ - Hardworking
$G$ - Good Grader
$R$ - Excellent Recommendation
$J$ - Landed a good Job
What can be inferred?

i: \( P(H, G) = P(H) \cdot P(G) \) ✗

ii: \( P(J|H, R) = P(J|R) \) ✓

iii: \( P(\overline{J}) = P(\overline{J}|H) \) ✗

Q: What is the value of \( P(H, G, \overline{R}, \overline{J}) \)?

A: \[
P(H, G, \overline{R}, \overline{J}) = P(H) \cdot P(G|H) \cdot P(\overline{R}|H, G) \cdot P(\overline{J}| \overline{R}) = 0.1 \cdot 0.4 \cdot 0.2 \\
* 0.8 = 0.0064
\]

Q: What if we want to add another parameter, \( C = \) Has The Right Connections?

Answer

\[
P(\overline{C} | \text{true}) = 0.1
\]

Reachability (the Bayes Ball)

- Shade evidence nodes
- Start at source node
- Try to reach target by search
- States: node, along with previous arc
- Successor function:
  - Unobserved nodes:
    - To any child
    - To any parent if coming from a child
  - Observed nodes:
    - From parent to parent
- If you can’t reach a node, it’s conditionally independent of the start node

Example

- L ind. \( T' \), given \( T \)?
  - Yes
- L ind. \( B \)?
  - Yes
- L ind. \( B \), given \( T \)?
  - No
- L ind. \( B \), given \( T' \)?
  - No
- L ind. \( B \), given \( T \) and \( R \)?
  - Yes
Naïve Bayes

- **Conditional Independence Assumption**: features are independent of each other given the class:
  \[ P(X_1, \ldots, X_n | C) = P(X_1 | C) \cdot P(X_2 | C) \cdot \cdots \cdot P(X_n | C) \]

- What can we model with naïve bayes?
  - Any process where,
    - Each cause has lots of “independent” effects
    - Easy to estimate the CPT for each effect
  - We want to reason about the probability of different causes given observed effects

**Naive Bayes Classifiers**

Task: Classify a new instance \( D \) based on a tuple of attribute values into one of the classes \( c_j \in C \)

\[
D = \langle x_1, x_2, \ldots, x_n \rangle
\]

\[
c_{MAP} = \arg\max_{c \in C} P(c | x_1, x_2, \ldots, x_n)
\]

\[
= \arg\max_{c \in C} \frac{P(x_1, x_2, \ldots, x_n | c)P(c)}{P(x_1, x_2, \ldots, x_n)}
\]

\[
= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n | c)P(c)
\]

**Summary**

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct