Belief networks

Chapter 15.1–2

Independence

Two random variables $A$ and $B$ are (absolutely) independent if

$$P(A|B) = P(A)$$

or

$$P(A, B) = P(A)P(B)$$

e.g., $A$ and $B$ are two coin tosses

If $n$ Boolean variables are independent, the full joint is

$$P(X_1, \ldots, X_n) = \Pi_{i=1}^n P(X_i)$$

hence can be specified by just $n$ numbers

Absolute independence is a very strong requirement, seldom met

Conditional independence

Consider the dentist problem with three random variables:

-Toothache, Cavity, Catch (steel probe catches in my tooth)

The full joint distribution 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:

1. $P(\text{Catch} | \text{Toothache}, \text{Cavity}) = P(\text{Catch} | \text{Cavity})$

i.e., Catch is conditionally independent of Toothache given Cavity

The same independence holds if I haven’t got a cavity:

2. $P(\text{Catch} | \text{Toothache}, \neg \text{Cavity}) = P(\text{Catch} \neg \text{Cavity})$

Equivalent statements to (1)

1a) $P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})$ Why??

1b) $P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity})P(\text{Catch} | \text{Cavity})$ Why??

Full joint distribution can now be written as

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

i.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)

Equivalent statements to (1)

1a) $P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})$ Why??

1b) $P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity})P(\text{Catch} | \text{Cavity})$ Why??

$P(\text{Toothache}, \text{Catch} | \text{Cavity})$

- $P(\text{Toothache} | \text{Catch}, \text{Cavity})P(\text{Catch} | \text{Cavity})$ (product rule)

- $P(\text{Toothache} | \text{Cavity})P(\text{Catch} | \text{Cavity})$ (from 1a)
Belief networks

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.

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

In the simplest case, conditional distribution represented as a conditional probability table (CPT)

Example

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: Burglar, Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects "causal" knowledge:

```
  Burglar   \\
  |         \Earthquake
  |         /  
  |     Alarm
  |     /   
  |    JohnCalls
  |    /     
  |   MaryCalls
```

Note: \( \leq k \) parents \( \Rightarrow O(d^n) \) numbers vs. \( O(d^k) \)

Semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

\[ P(X_1, \ldots, X_n) = \Pi_{i=1}^n P(X_i|\text{Parents}(X_i)) \]

e.g., \( P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) \) is given by?

\[ = P(-\neg B)P(-E)P(A | -B \wedge -E)P(J | A)P(M | A) \]

"Local" semantics: each node is conditionally independent of its non-descendants given its parents.

Theorem: Local semantics \( \Leftrightarrow \) global semantics

Markov blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents

Constructing belief networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables \( X_1, \ldots, X_n \)
2. For \( i = 1 \) to \( n \)
   - add \( X_i \) to the network
   - select parents from \( X_1, \ldots, X_{i-1} \) such that
   \[ P(X_i|\text{Parents}(X_i)) = P(X_i|X_1, \ldots, X_{i-1}) \]

This choice of parents guarantees the global semantics:

\[ P(X_1, \ldots, X_n) = \Pi_{i=1}^n P(X_i|\text{Parents}(X_i)) \]

(chain rule)

\[ = \Pi_{i=1}^n P(X_i|\text{Parents}(X_i)) \] by construction
Example

Suppose we choose the ordering \( M, J, A, B, E \)

\[
P(J|M) = P(J) ?
\]

\[
\]

- Mary Calls
- John Calls

- burglary
- Earthquake
- Burglary

- No
- Yes

- No
- Yes

Example: Car diagnosis

Initial evidence: engine won’t start
Testable variables (thin ovals), diagnosis variables (thick ovals)
Hidden variables (shaded) ensure sparse structure, reduce parameters
Example: Car insurance

Predict claim costs (medical, liability, property)
given data on application form (other unshaded nodes)

Compact conditional distributions

Noisy-OR distributions model multiple noninteracting causes
1) Parents $U_1, \ldots, U_k$ include all causes (can add leak node)
2) Independent failure probability $q_i$ for each cause alone
   $P(X|U_1, \ldots, U_k, \neg U_j, \ldots, \neg U_k) = 1 - \Pi_{i} q_i$

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>P(Fever)</th>
<th>P(\neg Fever)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.98</td>
<td>0.02</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.88</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Number of parameters linear in number of parents

Compact conditional distributions contd.

CPT grows exponentially with no. of parents
CPT becomes infinite with continuous-valued parent or child
Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:
$X = f(Parents(X))$ for some function $f$

E.g., Boolean functions
NorthAmerican $\Rightarrow$ Canadian $\lor$ US $\lor$ Mexican

E.g., numerical relationships among continuous variables
$
\frac{\partial Level}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation}
$

Hybrid (discrete+continuous) networks

Discrete (Subsidy? and Buys?); continuous (Harvest and Cost)

Continuous child variables

Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the linear Gaussian model, e.g.:
$P(Cost = c|Harvest = h, \text{Subsidy?} = \text{true})$
$= N(a_h + b_h c, \sigma_c)$
$= \frac{1}{\sigma_c \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{c - (a_h + b_h c)}{\sigma_c} \right)^2 \right)$

Mean $Cost$ varies linearly with $Harvest$, variance is fixed
Linear variation is unreasonable over the full range
but works OK if the likely range of $Harvest$ is narrow

Continuous child variables

All-continuous network with LG distributions
   $\Rightarrow$ full joint is a multivariate Gaussian

Discrete+continuous LG network is a conditional Gaussian network i.e.,
a multivariate Gaussian over all continuous variables for each combination of discrete variable values
Discrete variable w/ continuous parents

Probability of $\text{Buys}^? \text{ given } \text{Cost}$ should be a "soft" threshold:

![Graph showing the probability of Buys given Cost]

Probit distribution uses integral of Gaussian:

$$\Phi(x) = \int_{-\infty}^{x} N(0,1) dx$$

$$P(\text{Buys}^? = \text{true} \mid \text{Cost} = c) = \Phi((-c + \mu)/\sigma)$$

Can view as hard threshold whose location is subject to noise.

Discrete variable contd.

Sigmoid (or logit) distribution also used in neural networks:

$$P(\text{Buys}^? = \text{true} \mid \text{Cost} = c) = \frac{1}{1 + \exp(-2(-c+c_0)/\sigma)}$$

Sigmoid has similar shape to probit but much longer tails:

![Graph showing the sigmoid function]