Bayesian networks

Chapter 14.1–3

Outline

◊ Syntax
◊ Semantics
◊ Parameterized distributions

Bayesian 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 \(\Rightarrow\) "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) giving the distribution over \(X_i\) for each combination of parent values

Example

Topology of network encodes conditional independence assertions:

\[
\begin{array}{c}
\text{Weather} \\
\text{Cavity} \\
\text{Toothache} \quad \text{Catch}
\end{array}
\]

\textit{Weather} is independent of the other variables
\textit{Toothache} and \textit{Catch} are conditionally independent given \textit{Cavity}

Example contd.

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: \textit{Burglar}, \textit{Earthquake}, \textit{Alarm}, \textit{JohnCalls}, \textit{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.

\[
\begin{array}{c|c|c|c|c|c}
\text{Burglar} & \text{Earthquake} & \text{Alarm} & \text{JohnCalls} & \text{MaryCalls} \\
\hline
\text{T} & \text{T} & \text{T} & \text{.95} & \text{.90} & \text{.70} \\
\text{T} & \text{T} & \text{F} & \text{.94} & \text{.05} & \text{.70} \\
\text{F} & \text{T} & \text{T} & \text{.29} & \text{.05} & \text{.01} \\
\text{F} & \text{F} & \text{F} & \text{.001} & \text{.01} & \text{.01}
\end{array}
\]
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 $X_i = \text{true}$ (the number for $X_i = \text{false}$ 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. $2^5 = 31$).

Local semantics

Local semantics: each node is conditionally independent of its nondescendants given its parents.

Markov blanket

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

Constructing Bayesian 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:

   \[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \quad \text{(chain rule)} \]

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

Global semantics

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

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

For example:

\[ P(j \land m \land a \land \neg b \land \neg e) = P(j) P(m) P(a) P(\neg b) P(\neg e) \]

\[ = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \]

\[ \approx 0.000063 \]

Global semantics

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

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

For example:

\[ P(j \land m \land a \land \neg b \land \neg e) = P(j) P(m) P(a) P(\neg b) P(\neg e) \]

\[ = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \]

\[ \approx 0.000063 \]
Example

Suppose we choose the ordering $M, J, A, B, E$

$P(J|M) = P(J)?$  No
$P(A|J, M) = P(A|J)?$  No
$P(A|J, M) = P(A)?$  No
$P(B|A, J, M) = P(B|A)?$  Yes
$P(B|A, J, M) = P(B)?$  No

$P(E|B, A, J, M) = P(E|A, B)?$  Yes

Example contd.

Deciding conditional independence is hard in noncausal directions
(Causal models and conditional independence seem hardwired for humans!)
Assessing conditional probabilities is hard in noncausal directions
Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed
Example: Car diagnosis

Initial evidence: car won’t start
Testable variables (green), “broken, so fix it” variables (orange)
Hidden variables (gray) ensure sparse structure, reduce parameters

Compact conditional distributions contd.

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
   \[ P(X|[U_1 \ldots U_j, \neg U_{j+1}, \ldots \neg U_k]) = 1 - \prod_{i=1}^k q_i \]

<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 = 0.2 \times 0.1</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>T</td>
<td>F</td>
<td>0.94</td>
<td>0.06 = 0.6 \times 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.88</td>
<td>0.12 = 0.6 \times 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 \times 0.2 \times 0.1</td>
</tr>
</tbody>
</table>

Number of parameters linear in number of parents

Example: Car insurance

SocioEcon Age GoodStudent ExtraCar Mileage VehicleYear RiskAversion SeniorTrain DrivingSkill MakeModel DrivingHist DrivQuality Antilock Airbag CarValue HomeBase AntiTheft Theft OwnDamage PropertyCost LiabilityCost MedicalCost Cushioning Ruggedness Accident OtherCost OwnCost

Hybrid (discrete+continuous) networks

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

Option 1: discretization—possibly large errors, large CPTs
Option 2: finitely parameterized canonical families
1) Continuous variable, discrete+continuous parents (e.g., Cost)
2) Discrete variable, continuous parents (e.g., Buys?)

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.,:
\[ P(Cost = c|Harvest = h, Subsidy? = true) = N(\alpha h + b, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{1}{2} \frac{(c-\alpha h - b)^2}{\sigma^2} \right) \]

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

Compact conditional distributions

CPT grows exponentially with number of parents
CPT becomes infinite with continuous-valued parent or child
Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:
\[ X = f(\text{Parents}(X)) \text{ for some function } f \]

E.g., Boolean functions
\[ \text{NorthAmerican } \Rightarrow \text{Canadian} \lor \text{US} \lor \text{Mexican} \]

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

All-continuous network with LG distributions
⇒ full joint distribution is a multivariate Gaussian

Discrete-i-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 \textit{Buys} given \textit{Cost} should be a “soft” threshold:

\textbf{Probit} distribution uses integral of Gaussian:
\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt
\]
\[
P(\text{Buys} = \text{true} | \text{Cost} = c) = \Phi(-c + \mu / \sigma)
\]

Summary

Bayes nets provide a natural representation for (causally induced) conditional independence

- Topology + CPTs = compact representation of joint distribution
- Generally easy for (non)experts to construct
- Canonical distributions (e.g., noisy-OR) = compact representation of CPTs
- Continuous variables ⇒ parameterized distributions (e.g., linear Gaussian)

Why the probit?

1. It’s sort of the right shape
2. Can view as hard threshold whose location is subject to noise