Extended Bayesian networks

Peter Antal
antal@mit.bme.hu
Outline

- Reminder
- Bayesian network extensions
  - Canonical local models
  - Decision tree/graph local models
  - Dynamic Bayesian networks
Independence, Conditional independence

\(I_p(X; Y \mid Z)\) or \((X \perp Y \mid Z)_p\) denotes that \(X\) is independent of \(Y\) given \(Z\) defined as follows:

for all \(x, y\) and \(z\) with \(P(z) > 0\):

\[P(x; y \mid z) = P(x \mid z) \cdot P(y \mid z)\]

(Almost) alternatively, \(I_p(X; Y \mid Z)\) iff

\[P(X \mid Z, Y) = P(X \mid Z)\]

for all \(z, y\) with \(P(z, y) > 0\).

Other notations: \(D_p(X; Y \mid Z) = \text{def} = \neg I_p(X; Y \mid Z)\)

Direct dependence: \(D_p(X; Y \mid V / \{X, Y\})\)
The independence model of a distribution

The independence map (model) $M$ of a distribution $P$ is the set of the valid independence triplets:

$$M_P = \{ I_{P,1}(X_1; Y_1 | Z_1), \ldots, I_{P,K}(X_K; Y_K | Z_K) \}$$

If $P(X,Y,Z)$ is a Markov chain, then

$$M_P = \{ D(X;Y), D(Y;Z), I(X;Z|Y) \}$$

Normally/almost always: $D(X;Z)$

Exceptionally: $I(X;Z)$
Bayesian networks: three facets

1. Causal model

2. Graphical representation of (in)dependencies

3. Concise representation of joint distributions

\[ P(M, O, D, S, T) = P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M) \]
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 ≈ "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
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.

- **Burglary**
  - \( P(B) = 0.01 \)

- **Earthquake**
  - \( P(E) = 0.002 \)

- **Alarm**
  - Probabilities for different combinations of Burglary and Earthquake:
    | B | E | P(A|B,E) |
    |---|---|---------|
    | T | T | 0.95    |
    | T | F | 0.94    |
    | F | T | 0.29    |
    | F | F | 0.001   |

- **JohnCalls**
  - Probabilities for John calling given Alarm:
    | A | P(J|A) |
    |---|-------|
    | T | 0.90  |
    | F | 0.05  |

- **MaryCalls**
  - Probabilities for Mary calling given Alarm:
    | A | P(M|A) |
    |---|-------|
    | T | 0.70  |
    | F | 0.01  |
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. $2^5-1 = 31$).
Tfh: 5 szülő csomópont bináris értékű
2 szülő csomópont 3-as értékű
1 szülő csomópont 4-es értékű és az eredmény csomópont 5-ös értékű ?????
A multinomiális általános eset II.

<table>
<thead>
<tr>
<th>Sz1</th>
<th>Sz2</th>
<th>Sz3</th>
<th>Sz4</th>
<th>Sz5</th>
<th>Sz6</th>
<th>Sz7</th>
<th>Sz8</th>
<th>Kimeneti változó</th>
<th>e1</th>
<th>e2</th>
<th>e3</th>
<th>e4</th>
<th>e5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>e1</td>
<td>e1</td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e2</td>
<td>e2</td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e3</td>
<td>e3</td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e1</td>
<td></td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e2</td>
<td></td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e3</td>
<td></td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e4</td>
<td></td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
</tbody>
</table>

Minden kombináció

$2^5 \times 3^2 \times 4$ szülői feltétel van (FVT sor) és 4 (független érték) (FVT oszlop) = összesen: $(32 \times 9 \times 4) \times 4 = 4608$
eyüttes eloszláshoz kell: $2^5 \times 3^2 \times 4 \times 5 – 1 = 5759$
Constructing Bayesian networks

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
     \[ P(X_i \mid \text{Parents}(X_i)) = P(X_i \mid X_1, \ldots, X_{i-1}) \]

This choice of parents guarantees:

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \\
= \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))
\]  

\(\text{//}(\text{chain rule})\)

\(\text{//}(\text{by construction})\)
Effect of ordering

- Construct a general BN for the example using the ordering M, J, A, B, E.
- Construct a Naïve–BN for a reverse ordering when the central variable \( Y \) is the last one (and not the first).
The full joint distribution is defined as the product of the local conditional distributions:

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

e.g., \(P(j \land m \land a \land \neg b \land \neg e)\)

\[
= P(j \mid a) P(m \mid a) P(a \mid \neg b, \neg e) P(\neg b) P(\neg e)
\]
Context-specific independence

$I_p(X;Y|Z=z)$ or $(X \perp Y|Z=z)_p$ denotes that $X$ is independent of $Y$ for a specific value $z$ of $Z$:

for $z$ and for all $x,y$: $P(x;y|z) = P(x|z) P(y|z)$


Learning decision trees

Problem: decide whether to wait for a table at a restaurant, based on the following attributes:
1. Alternate: is there an alternative restaurant nearby?
2. Bar: is there a comfortable bar area to wait in?
3. Fri/Sat: is today Friday or Saturday?
4. Hungry: are we hungry?
5. Patrons: number of people in the restaurant (None, Some, Full)
6. Price: price range ($, $$, $$$)
7. Raining: is it raining outside?
8. Reservation: have we made a reservation?
9. Type: kind of restaurant (French, Italian, Thai, Burger)
10. WaitEstimate: estimated waiting time (0–10, 10–30, 30–60, >60)
Examples described by attribute values (Boolean, discrete, continuous)

E.g., situations where I will/won't wait for a table:

- Classification of examples is positive (T) or negative (F)
Decision trees

- One possible representation for hypotheses
- E.g., here is the “true” tree for deciding whether to wait:
Decision trees can express any function of the input attributes.
E.g., for Boolean functions, truth table row $\rightarrow$ path to leaf:

![Decision Tree Example]

- Trivially, there is a consistent decision tree for any training set with one path to leaf for each example (unless $f$ nondeterministic in $x$) but it probably won't generalize to new examples
- Prefer to find more compact decision trees
Hypothesis spaces

How many distinct decision trees with $n$ Boolean attributes?

$= \text{number of Boolean functions} \quad = \text{number of distinct truth tables with } 2^n \text{ rows} = 2^{2n}$

- E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees
Hypothesis spaces

How many distinct decision trees with \( n \) Boolean attributes?

\[
\begin{align*}
\text{= number of Boolean functions} \\
\text{= number of distinct truth tables with } 2^n \text{ rows } = 2^{2^n}
\end{align*}
\]

- E.g., with 6 Boolean attributes, there are
  \( 18,446,744,073,709,551,616 \) trees

How many purely conjunctive hypotheses (e.g., \( \text{Hungry} \land \neg \text{Rain} \))?

- Each attribute can be in (positive), in (negative), or out
  \( \Rightarrow 3^n \) distinct conjunctive hypotheses

- More expressive hypothesis space
  - increases chance that target function can be expressed
  - increases number of hypotheses consistent with training set
    \( \Rightarrow \) may get worse predictions
Decision trees, decision graphs

Decision tree: Each internal node represent a (univariate) test, the leaves contains the conditional probabilities given the values along the path.

Decision graph: If conditions are equivalent, then subtrees can be merged. E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)

A.I.: BN homework guide
**Noisy-OR**

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

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

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>$P(\text{Fever})$</th>
<th>$P(\neg\text{Fever})$</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 × 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>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06 = 0.6 × 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12 = 0.6 × 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 × 0.2 × 0.1</td>
</tr>
</tbody>
</table>

Number of parameters **linear** in number of parents
Dynamic Bayesian networks

$X_t$, $E_t$ contain arbitrarily many variables in a replicated Bayes net

http://phoenix.mit.bme.hu:49080/kgt/
DBNs vs. HMMs

Every HMM is a single-variable DBN; every discrete DBN is an HMM

\[
\begin{align*}
X_t & \rightarrow X_{t+1} \\
Y_t & \rightarrow Y_{t+1} \\
Z_t & \rightarrow Z_{t+1}
\end{align*}
\]

Sparse dependencies $\Rightarrow$ exponentially fewer parameters;
  e.g., 20 state variables, three parents each
  DBN has $20 \times 2^3 = 160$ parameters, HMM has $2^{20} \times 2^{20} \approx 10^{12}$
Inferring independencies from structure: $d$–separation

$I_G(X;Y|Z)$ denotes that $X$ is $d$–separated (directed separated) from $Y$ by $Z$ in directed graph $G$. 

(1) $X$ (2) $Z$ (3) $Y$
Definition 7  A distribution \( P(X_1, \ldots, X_n) \) obeys the global Markov condition w.r.t. DAG \( G \), if

\[
\forall X, Y, Z \subseteq U \ (X \perp Y|Z)_G \Rightarrow (X \perp Y|Z)_P ,
\]

(9)

where \( (X \perp Y|Z)_G \) denotes that \( X \) and \( Y \) are d-separated by \( Z \), that is if every path \( p \) between a node in \( X \) and a node in \( Y \) is blocked by \( Z \) as follows

1. either path \( p \) contains a node \( n \) in \( Z \) with non-converging arrows (i.e. \( \rightarrow n \rightarrow \) or \( \leftarrow n \rightarrow \)),

2. or path \( p \) contains a node \( n \) not in \( Z \) with converging arrows (i.e. \( \rightarrow n \leftarrow \)) and none of its descendants of \( n \) is in \( Z \).
Conditional independencies allows:
- efficient representation of the joint probabilility distribution,
- efficient inference to compute conditional probabilities.

Bayesian networks use directed acyclic graphs to represent
- conditional independencies,
- conditional probability distributions,
- causal mechanisms.

Design of variables and order of the variables can drastically influence structure

Suggested reading:
- Charniak: Bayesian networks without tears, 1991