Bayes Nets II: Independence Day

CS 5300 / CS 6300
Artificial Intelligence
Spring 2010
Hal Daumé III
hal@cs.utah.edu

www.cs.utah.edu/~hal/courses/2010S_AI

Many slides courtesy of Dan Klein, Stuart Russell, or Andrew Moore
Reasoning Patterns and D-Separation

Sargur Srihari
srihari@cedar.buffalo.edu
Example Bayes’ Net
Bayes’ Nets

- A Bayes’ net is an efficient encoding of a probabilistic model of a domain

- Questions we can ask:
  - Inference: given a fixed BN, what is $P(X \mid e)$?
  - Representation: given a fixed BN, what kinds of distributions can it encode?
  - Modeling: what BN is most appropriate for a given domain?
Bayes’ Net Semantics

➢ A Bayes’ net:
  ➢ A set of nodes, one per variable \( X \)
  ➢ A directed, acyclic graph
  ➢ A conditional distribution of each variable conditioned on its parents (the parameters \( \theta \))

\[
P(X | a_1 \ldots a_n)
\]

➢ Semantics:
  ➢ A BN **defines** a joint probability distribution over its variables:

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

➢ We can take a Bayes’ net and build any entry from the full joint distribution it encodes

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

➢ Typically, there’s no reason to build ALL of it
➢ We build what we need on the fly

➢ To emphasize: every BN over a domain implicitly defines a joint distribution over that domain, specified by local probabilities and graph structure
Example: Coin Flips

Only distributions whose variables are absolutely independent can be represented by a Bayes’ net with no arcs.
Example: Traffic

\[ P(R) \]

<table>
<thead>
<tr>
<th></th>
<th>1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td></td>
</tr>
<tr>
<td>\neg r</td>
<td>3/4</td>
</tr>
</tbody>
</table>

\[ P(T|R) \]

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>\neg t</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>3/4</td>
<td>1/4</td>
</tr>
<tr>
<td>\neg r</td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

\[ P(r, \neg t) = \]
Example: Alarm Network

P(b, e, ¬a, j, m) =

| B | E | P(A|B,E) |
|---|---|----------|
| T | T | .95      |
| T | F | .94      |
| F | T | .29      |
| F | F | .001     |

P(B) = .001
P(E) = .002

| A | P(J|A) |
|---|-------|
| T | .90   |
| F | .05   |

| A | P(M|A) |
|---|-------|
| T | .70   |
| F | .01   |

Burglary → Alarm
Earthquake → Alarm
JohnCalls → Alarm
MaryCalls → Alarm
Example: Traffic II

- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame

Diagram:

- L (Low pressure) → R (It rains)
- R (It rains) → B (Ballgame)
- R (It rains) → T (Traffic)
- D (Roof drips) → T (Traffic)
Size of a Bayes’ Net

➢ How big is a joint distribution over N Boolean variables?

➢ How big is an N-node net if nodes have k parents?

➢ Both give you the power to calculate \( P(X_1, X_2, \ldots X_n) \)

➢ BNs: Huge space savings!

➢ Also easier to elicit local CPTs

➢ Also turns out to be faster to answer queries (next class)
Bayes’ Nets

➢ So far:
  ➢ What is a Bayes’ net?
  ➢ What joint distribution does it encode?

➢ Next: how to answer queries about that distribution
  ➢ Key idea: conditional independence
  ➢ Last class: assembled BNs using an intuitive notion of conditional independence as causality
  ➢ Today: formalize these ideas
  ➢ Main goal: answer queries about conditional independence and influence

➢ After that: how to answer numerical queries (inference)
Bayesian Network: Student Model

Graph and CPDs

Val(I) = \{i^0 = \text{low intelligence}, \ i^1 = \text{high intelligence}\}

Val(D) = \{d^0 = \text{easy}, \ d^1 = \text{hard}\}

Val(G) = \{g^1 = A, \ g^2 = B, \ g^3 = C\}

Val(S) = \{s^0 = \text{low}, \ s^1 = \text{high}\}

Val(L) = \{l^0 = \text{weak}, \ l^1 = \text{strong}\}

\[
P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)
\]
\[
P(i^1, d^0, g^2, s^1, l^0) = P(i^1)P(d^0)P(g^2 \mid i^1, d^0)P(s^1 \mid i^1)P(l^0 \mid g^2)
\]
\[
= 0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8 \cdot 0.4 = 0.004608
\]
Reasoning Patterns

Reasoning about a student George using the model

- **Causal Reasoning**
  - George is interested in knowing as to how likely he is to get a strong letter (based on intelligence, difficulty)?

- **Evidential Reasoning**
  - Recruiter is interested in knowing whether George is intelligent (based on letter, SAT)
Causal Reasoning

1. How likely is George to get a strong letter (knowing nothing else)?
   • $P(l^1) = 0.502$
   • Obtained by summing-out other variables in joint distribution

2. But George is not so intelligent ($i^0$)
   • $P(l^1|i^0) = 0.389$

3. Next we find out ECON101 is easy ($d^0$)
   • $P(l^1|i^0, d^0) = 0.513$

Query is Example of Causal Reasoning:
Predicting downstream effects of factors such as intelligence

Observe how probabilities change as evidence is obtained

$P(D, I, G, S, l^1) = \sum_{D, I, G, S} P(D)P(I)P(G|D, I)P(S|I)P(l^1|G)$
Evidential Reasoning

- Recruiter wants to hire intelligent student
- A priori George is 30% likely to be intelligent
  \[ P(i^1) = 0.3 \]
- Finds that George received grade \( C (g^3) \) in ECON101
  \[ P(i^1 | g^3) = 0.079 \]
- Similarly probability class is difficult goes up from 0.4 to
  \[ P(d^1 | g^3) = 0.629 \]
- If recruiter has lost grade but has letter
  \[ P(i^1 | l^0) = 0.14 \]

- Recruiter has both grade and letter
  \[ P(i^1 | l^0, g^3) = 0.079 \]
  - Same as if he had only grade
  - Letter is immaterial
- Reasoning from effects to causes is called evidential reasoning
Intercausal reasoning

- Recruiter has grade (letter does not matter)
  \[ P(i^1|g^3) = P(i^1|l^0,g^3) = 0.079 \]
- Recruiter receives high SAT score (leads to dramatic increase)
  \[ P(i^1|g^3,s^1) = 0.578 \]
- Intuition:
  - High SAT score outweighs poor grade since low intelligence rarely gets good SAT scores
  - Smart students more likely to get Cs in hard classes
- Probability of class is difficult also goes up from
  \[ P(d^1|g^3) = 0.629 \] to
  \[ P(d^1|g^3,s^1) = 0.76 \]
**Explaining Away**

**An example:**

- Given grade
  \[ P(i^1|l^0,g^3)=0.079 \]
- If we observe ECON101 is a hard class
  \[ P(i^1|g^3,d^1)=0.11 \]
- We have provided partial explanation for George’s performance in ECON101

**Another example:**

- If George gets a B in ECON101
  \[ P(i^1|g^2)=0.175 \]
- If we observe ECON101 is a hard class
  \[ P(i^1|g^2,d^1)=0.34 \]
- We have explained away the poor grade via the difficulty of the class

**Explaining away is one type of intercausal reasoning**

- Different causes of the same effect can interact
- All determined by probability calculation rather than heuristics
Intercausal Reasoning is Common in Human Reasoning

Another example of explaining away

- Binary Variables
- Fever & Sore Throat can be caused by mono and flu
- When flu is diagnosed, probability of mono is reduced (although mono could still be present)
- It provides an alternative explanation of symptoms

\[ P(m^1|s^1) > P(m^1|s^1,f^1) \]
Another Type of Intercausal Reasoning

- Binary Variables
  - Murder (leaf node)
  - Motive and Opportunity are causal nodes
- Binary Variables $X, Y, Z$
- $X$ and $Y$ both increase the probability of Murder
  - $P(z^I|x^I) > P(z^I)$
  - $P(z^I|y^I) > P(z^I)$
- Each of $X$ and $Y$ increase probability of other
  - $P(x^I > z^I) < P(x^I|y^I, z^I)$
  - $P(y^I|z^I) < P(y^I|x^I, z^I)$

Can go in any direction Different from Explaining Away
Dependencies and Independencies

• Crucial for understanding network behavior
• Independence properties are important for answering queries
  – Exploited to reduce computation of inference
  – A distribution $P$ that factorizes over $G$ satisfies $I(G)$
  – Are there other independencies that can be read off directly from $G$?
    • That hold for every $P$ that factorizes over $G$
Conditional Independence

- Reminder: independence
- X and Y are independent if
  \[ \forall x, y \quad P(x, y) = P(x)P(y) \quad \implies \quad X \perp Y \]

- X and Y are conditionally independent given Z
  \[ \forall x, y, z \quad P(x, y|z) = P(x|z)P(y|z) \quad \implies \quad X \perp Y|Z \]

- (Conditional) independence is a property of a distribution
Example: Independence

➢ For this graph, you can fiddle with $\theta$ (the CPTs) all you want, but you won’t be able to represent any distribution in which the flips are dependent!

\[
\begin{align*}
X_1 & \quad X_2 \\
\begin{array}{c|c}
P(X_1) & P(X_2) \\
\hline
h & h \\
t & t
\end{array}
\end{align*}
\]

All distributions
Topology Limits Distributions

- Given some graph topology $G$, only certain joint distributions can be encoded.
- The graph structure guarantees certain (conditional) independences.
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs.
Independence in a BN

- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can calculate using algebra (really tedious)
  - If no, can prove with a counter example
- Example:

  \[
  \begin{array}{c}
  X \\
  \end{array}
  \begin{array}{c}
  Y \\
  \end{array}
  \begin{array}{c}
  Z \\
  \end{array}
  \]

- Question: are X and Z independent?
  - Answer: not *necessarily*, we’ve seen examples otherwise: low pressure causes rain which causes traffic.
  - X can influence Z, Z can influence X (via Y)
  - Addendum: they *could* be independent: how?
Direct Connection between $X$ and $Y$

- $X$ and $Y$ are correlated regardless of any evidence about any other variables
  - E.g., Feature $Y$ and character $X$ are correlated
  - Grade $G$ and Letter $L$ are correlated

- If $X$ and $Y$ are directly connected we can get examples where they influence each other regardless of $Z$
Indirect Connection betwn $X$ and $Y$

- Four cases where $X$ and $Y$ are connected via $Z$
  1. Indirect causal effect
  2. Indirect evidential effect
  3. Common cause
  4. Common effect

- We will see that first three cases are similar while fourth case ($V$-structure) is different
1. Indirect Causal Effect: $X \rightarrow Z \rightarrow Y$

- **Cause** $X$ cannot influence effect $Y$ if $Z$ observed
  - Observed $Z$ blocks influence
- **If Grade is observed then I does not influence L**
  - Intell influences Letter if Grade is unobserved
Causal Chains

➢ This configuration is a “causal chain”

\[ P(x, y, z) = P(x)P(y|x)P(z|y) \]

➢ Is \( X \) independent of \( Z \) given \( Y \)?

\[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} = P(z|y) \]

\( \text{Yes!} \)

➢ Evidence along the chain “blocks” the influence

\( X: \text{Low pressure} \)

\( Y: \text{Rain} \)

\( Z: \text{Traffic} \)
2. Indirect Evidential Effect: \( Y \rightarrow Z \rightarrow X \)

- Evidence \( X \) can influence \( Y \) via \( Z \) only if \( Z \) is unobserved
  - Observed \( Z \) blocks influence
- If Grade unobserved, Letter influences assessment of Intelligence
- Dependency is a symmetric notion
  - \( X \perp Y \) does not hold then \( Y \perp X \) does not hold either

\[ Z = \text{Grade} \]
3. Common Cause: $X \leftrightarrow Z \rightarrow Y$

- $X$ can influence $Y$ if and only if $Z$ is not observed
  - Observed $Z$ blocks
- Grade is correlated with SAT score
- But if Intelligence is observed then SAT provides no additional information
Common Cause

➢ Another basic configuration: two effects of the same cause
➢ Are X and Z independent?
➢ Are X and Z independent given Y?

\[
P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)}
\]

\[
= P(z|y)
\]

Yes!

➢ Observing the cause blocks influence between effects.
4. Common Effect (V-structure) $X \rightarrow Z \leftarrow Y$

- Influence cannot flow on trail $X \rightarrow Z \leftarrow Y$ if $Z$ is not observed
  - Observed $Z$ enables
  - Opposite to previous 3 cases (Observed $Z$ blocks)

- When $G$ not observed $I$ and $D$ are independent
- When $G$ is observed, $I$ and $D$ are correlated

I ind $D | \sim G$
Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: remember the ballgame and the rain causing traffic, no correlation?
    - Still need to prove they must be (try it!)
  - Are X and Z independent given Y?
    - No: remember that seeing traffic put the rain and the ballgame in competition?
  - This is backwards from the other cases
    - Observing the effect enables influence between effects.
Active Trail

- Grade is not observed
- Observe weak letter
  - Which indicates low grade
  - Suffices to correlate $D$ and $I$
- When influence can flow from $X$ to $Y$ via $Z$ then trail $X\rightarrow Z\rightarrow Y$ is active

Summary

Causal trail: $X\rightarrow Z\rightarrow Y$: active iff $Z$ not observed

Evidential Trail: $X\leftarrow Z\leftarrow Y$: active iff $Z$ is not observed

Common Cause: $X\leftarrow Z\rightarrow Y$: active iff $Z$ is not observed

Common Effect: $X\rightarrow Z\leftarrow Y$: active iff either $Z$ or one of its descendants is observed
D-separation definition

• Let $X, Y$ and $Z$ be three sets of nodes in $G$.

• $X$ and $Y$ are d-separated given $Z$ denoted $d_{-}sep_{G}(X:Y|Z)$ if there is no active trail between any node $X \in X$ and $Y \in Y$ given $Z$.

• That is, nodes in $X$ cannot influence nodes in $Y$.

• Provides notion of separation between nodes in a directed graph ("directed" separation).
Independencies from d-separation

• Consider variables pairwise using d-separation

\[ I(G) = \{(D \perp I, S, L | \phi), (I \perp D, S, L | \phi), \]
\[ (G \perp L, S | D, I), (L \perp I, D, S | G), (S \perp D, G, L | I), (D \perp S | I)\} \]

– Also called Markov independencies

• Definition:

\[ I(G) = \{(X \perp Y | Z): d-sep_G(X:Y|Z)\} \]
The General Case

- Any complex example can be analyzed using these three canonical cases

- General question: in a given BN, are two variables independent (given evidence)?

- Solution: analyze the graph
Reachability

➢ Recipe: shade evidence nodes

➢ Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent

➢ Almost works, but not quite
  ➢ Where does it break?
  ➢ Answer: the v-structure at T doesn’t count as a link in a path unless “active”
Reachability (the Bayes’ Ball)

- Correct algorithm:
  - Shade in evidence
  - Start at source node
  - Try to reach target by search
  - States: pair of (node $X$, previous state $S$)

- Successor function:
  - $X$ unobserved:
    - To any child
    - To any parent if coming from a child
  - $X$ observed:
    - From parent to parent
  - If you can’t reach a node, it’s conditionally independent of the start node given evidence
Reachability (D-Separation)

- Question: Are X and Y conditionally independent given evidence variables \( \{Z\} \)?
  - Look for “active paths” from X to Y
  - No active paths = independence!

- A path is active if each triple is either a:
  - Causal chain \( A \rightarrow B \rightarrow C \) where B is unobserved (either direction)
  - Common cause \( A \leftarrow B \rightarrow C \) where B is unobserved
  - Common effect (aka \( v \)-structure) \( A \rightarrow B \leftarrow C \) where B or one of its descendents is observed
Example

\[ A \perp W \]

\[ A \perp W | R \]

Yes

Diagram:
- **aliens**
- **watch**
- **late**
- **report**

The diagram shows a Bayesian network with nodes for aliens, watch, late, and report, with edges indicating conditional dependencies.
Example

\[ L \perp T' | T \quad Yes \]
\[ L \perp B \quad Yes \]
\[ L \perp B | T \]
\[ L \perp B | T' \]
\[ L \perp B | T', R \quad Yes \]
Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I’m sad

- Questions:

\[
T \perp D \\
T \perp D \mid R \\
T \perp D \mid R, S
\]
Causality?

- 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 *Drips*
  - End up with arrows that reflect correlation, not causation

- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - **Topology only guaranteed to encode conditional independence**
Example: Traffic

- Basic traffic net
- Let’s multiply out the joint

\[

\begin{array}{c|c}
  R & P(R) \\
  \hline
  r & 1/4 \\
  \neg r & 3/4 \\
\end{array}
\]

\[

\begin{array}{c|c|c}
  T & P(T|R) \\
  \hline
  r & t & 3/4 \\
  r & \neg t & 1/4 \\
  \neg r & t & 1/2 \\
  \neg r & \neg t & 1/2 \\
\end{array}
\]

\[

\begin{array}{c|c|c}
  T & P(T, R) \\
  \hline
  r & t & 3/16 \\
  r & \neg t & 1/16 \\
  \neg r & t & 6/16 \\
  \neg r & \neg t & 6/16 \\
\end{array}
\]
Example: Reverse Traffic

- Reverse causality?

<table>
<thead>
<tr>
<th>$T$</th>
<th>$R$</th>
</tr>
</thead>
</table>

\[
P(T) = \begin{array}{c|c}
  \text{t} & 9/16 \\
  \sim \text{t} & 7/16 \\
\end{array}
\]

\[
P(R|T) = \begin{array}{c|c|c}
  \text{t} & \text{r} & 1/3 \\
  & \sim \text{r} & 2/3 \\
  \sim \text{t} & \text{r} & 1/7 \\
  & \sim \sim \text{r} & 6/7 \\
\end{array}
\]

\[
P(T, R) = \begin{array}{c|c|c}
  \text{r} & \text{t} & 3/16 \\
  \text{r} & \sim \text{t} & 1/16 \\
  \sim \text{r} & \text{t} & 6/16 \\
  \sim \text{r} & \sim \text{t} & 6/16 \\
\end{array}
\]
Example: Coins

- Extra arcs don’t prevent representing independence, just allow non-independence

\[
\begin{array}{c|c}
X_1 & h & 0.5 \\
& t & 0.5 \\
\end{array}
\quad
\begin{array}{c|c}
X_2 & h & 0.5 \\
& t & 0.5 \\
\end{array}
\quad
\begin{array}{c|c|c|c}
P(X_1) & h & 0.5 \\
& t & 0.5 \\
\end{array}
\quad
\begin{array}{c|c|c|c}
P(X_2|X_1) & h|h & 0.5 \\
& t|h & 0.5 \\
& h|t & 0.5 \\
& t|t & 0.5 \\
\end{array}
\]
Alternate BNs

- MaryCalls
- JohnCalls
- Alarm
- Burglary
- Earthquake

- B
- E
- A
- J
- M
Summary

➢ Bayes nets compactly encode joint distributions

➢ Guaranteed independencies of distributions can be deduced from BN graph structure

➢ The Bayes’ ball algorithm (aka d-separation)

➢ A Bayes’ net may have other independencies that are not detectable until you inspect its specific distribution
Independencies in a BN

- Graph with CPDs is equivalent to a set of independence assertions

\[ P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G) \]

- Local Conditional Independence Assertions (starting from leaf nodes):

  \[ I(G) = \{(L \perp I, D, S \mid G), \quad L \text{ is conditionally independent of all other nodes given parent } G \}
  \]
  \[ (S \perp D, G, L \mid I), \quad S \text{ is conditionally independent of all other nodes given parent } I \]
  \[ (G \perp S \mid D, I), \quad \text{Even given parents, } G \text{ is NOT independent of descendant } L \]
  \[ (I \perp D \mid \phi), \quad \text{Nodes with no parents are marginally independent} \]
  \[ (D \perp I, S \mid \phi) \]

- Parents of a variable shield it from probabilistic influence
  - Once value of parents known, no influence of ancestors
  - Information about descendants can change beliefs about a node
Soundness and Completeness

- Formalizing notion of d-separation
- **Soundness Theorem**
  - If a distribution $P$ factorizes according to $G$ then $I(G) \subseteq I(P)$
- A distribution $P$ is faithful to graph $G$ if any independence in $P$ is reflected in $G$
  - $G$ is then called a Perfect Map
- **Completeness Theorem**
  - Definition of $I(G)$ is the maximal one
- Thus d-separation test precisely characterizes independencies that hold for $P$
Algorithm for d-separation

- Enumerating all trails is inefficient
  - Number of trails is exponential with graph size
- Linear time algorithm has two phases
- Algorithm $\text{Reachable}(G,X,Z)$ returns nodes for $X$
- Phase 1 (simple)
  - Traverse bottom-up from leaves marking all nodes in $Z$ or descendants in $Z$; to enable v-structures
- Phase 2 (subtle)
  - Traverse top-down from $X$ to $Y$ stopping when blocked by a node
I-Equivalence

- Conditional assertion statements can be the same with different structures
- Two graphs $K_1$ and $K_2$ are I-equivalent if $I(K_1) = I(K_2)$
- Skeleton of a BN graph $G$ is an undirected graph with an edge for every edge in $G$
- If two BN graphs have the same set of skeletons and v-structures then they are I-equivalent

Same skeleton
Same v-structure $X \rightarrow Y \leftarrow Z$