Probabilistic Models

- A probabilistic model is a joint distribution over a set of variables

\[ P(X_1, X_2, \ldots X_n) \]

- Given a joint distribution, we can reason about unobserved variables given observations (evidence)
- General form of a query:

\[ P(X_q | x_{e1}, \ldots x_{ek}) \]

- This kind of posterior distribution is also called the belief function of an agent which uses this model.
**Ghostbusters by Enumeration**

- Reminder: ghost is hidden, sensors are noisy
- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top
- Sensor model:
  \[
  P(t | g) = 0.8 \\
  P(t | \neg g) = 0.4 \\
  P(b | g) = 0.4 \\
  P(b | \neg g) = 0.8
  \]
- Joint Distribution

<table>
<thead>
<tr>
<th>T</th>
<th>B</th>
<th>G</th>
<th>P(T,B,G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>b</td>
<td>g</td>
<td>0.16</td>
</tr>
<tr>
<td>t</td>
<td>b</td>
<td>\neg g</td>
<td>0.16</td>
</tr>
<tr>
<td>t</td>
<td>\neg b</td>
<td>g</td>
<td>0.24</td>
</tr>
<tr>
<td>t</td>
<td>\neg b</td>
<td>\neg g</td>
<td>0.04</td>
</tr>
<tr>
<td>\neg t</td>
<td>b</td>
<td>g</td>
<td>0.04</td>
</tr>
<tr>
<td>\neg t</td>
<td>b</td>
<td>\neg g</td>
<td>0.24</td>
</tr>
<tr>
<td>\neg t</td>
<td>\neg b</td>
<td>g</td>
<td>0.06</td>
</tr>
<tr>
<td>\neg t</td>
<td>\neg b</td>
<td>\neg g</td>
<td>0.06</td>
</tr>
</tbody>
</table>

**Independence**

- Two variables are *independent* if:
  \[
  \forall x, y : P(x, y) = P(x)P(y)
  \]
- This says that their joint distribution *factors* into a product two simpler distributions
- Another form:
  \[
  \forall x, y : P(x|y) = P(x)
  \]
- We write: \(X \perp Y\)
- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best “close” to independent
  - What could we assume for \{Weather, Traffic, Cavity, Toothache\}?
Example: Independence

- N fair, independent coin flips:

  \[
  P(X_1) \quad P(X_2) \quad \ldots \quad P(X_n)
  \]

  \[
  \begin{array}{c|c}
  H & 0.5 \\ \hline
  T & 0.5 \\
  \end{array}
  \quad \begin{array}{c|c}
  H & 0.5 \\ \hline
  T & 0.5 \\
  \end{array}
  \ldots
  \quad \begin{array}{c|c}
  H & 0.5 \\ \hline
  T & 0.5 \\
  \end{array}
  \]

  \[P(X_1, X_2, \ldots X_n)\]

  \[
  2^n
  \]

Example: Independence?

\[
P(T)
\]

\[
P_1(T, W)
\]

\[
P_2(T, W)
\]

\[
P(W)
\]
Conditional Independence

- \( P(\text{Toothache}, \text{Cavity}, \text{Catch}) \)

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - \( P(\text{catch} \mid \text{toothache}, \text{cavity}) = P(\text{catch} \mid \text{cavity}) \)

- The same independence holds if I don't have a cavity:
  - \( P(\text{catch} \mid \text{toothache}, \neg\text{cavity}) = P(\text{catch} \mid \neg\text{cavity}) \)

- Catch is \textit{conditionally independent} of Toothache given Cavity:
  - \( P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity}) \)

- Equivalent statements:
  - \( P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \)
  - \( P(\text{Toothache}, \text{Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) P(\text{Catch} \mid \text{Cavity}) \)
  - One can be derived from the other easily

Conditional Independence

- Unconditional (absolute) independence is very rare (why?)

- Conditional independence is our most basic and robust form of knowledge about uncertain environments:
  \[
  \forall x, y, z : P(x, y|z) = P(x|z)P(y|z) \\
  \forall x, y, z : P(x|z, y) = P(x|z)
  \]

- What about this domain:
  - Traffic
  - Umbrella
  - Raining

- What about fire, smoke, alarm?
The Chain Rule

- Can always factor any joint distribution as an incremental product of conditional distributions

\[ P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots \]

\[ P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i|X_1 \ldots X_{i-1}) \]

- Why is the chain rule true? *Hint: product rule + induction*

- This actually claims nothing…

- What are the sizes of the tables we supply?

The Chain Rule II

- Trivial decomposition:

\[ P(\text{Traffic, Rain, Umbrella}) = \]

\[ P(\text{Rain})P(\text{Traffic|Rain})P(\text{Umbrella|Rain, Traffic}) \]

- With assumption of conditional independence:

\[ P(\text{Traffic, Rain, Umbrella}) = \]

\[ P(\text{Rain})P(\text{Traffic|Rain})P(\text{Umbrella|Rain}) \]

- Bayes’ nets / graphical models help us express conditional independence assumptions
Ghostbusters Chain Rule

- Each sensor depends only on where the ghost is.
- That means, the two sensors are conditionally independent, given the ghost position.
- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top

\[ P(T,B,G) = P(G) P(T|G) P(B|G) \]

<table>
<thead>
<tr>
<th>T</th>
<th>B</th>
<th>G</th>
<th>P(T,B,G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>b</td>
<td>g</td>
<td>0.16</td>
</tr>
<tr>
<td>t</td>
<td>b</td>
<td>¬g</td>
<td>0.16</td>
</tr>
<tr>
<td>t</td>
<td>¬b</td>
<td>g</td>
<td>0.24</td>
</tr>
<tr>
<td>t</td>
<td>¬b</td>
<td>¬g</td>
<td>0.04</td>
</tr>
<tr>
<td>¬t</td>
<td>b</td>
<td>g</td>
<td>0.04</td>
</tr>
<tr>
<td>¬t</td>
<td>b</td>
<td>¬g</td>
<td>0.24</td>
</tr>
<tr>
<td>¬t</td>
<td>¬b</td>
<td>g</td>
<td>0.06</td>
</tr>
<tr>
<td>¬t</td>
<td>¬b</td>
<td>¬g</td>
<td>0.06</td>
</tr>
</tbody>
</table>

- \( P(t | g) = 0.8 \)
- \( P(t | ¬g) = 0.4 \)
- \( P(b | g) = 0.4 \)
- \( P(b | ¬g) = 0.8 \)

Probabilistic Models

- Models describe how (a portion of) the world works

- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - “Essentially, all models are wrong; but some are useful.”
    – George E. P. Box

- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information
Bayes’ Nets: Big Picture

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

- Bayes’ nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - For about 10 min, we’ll be vague about how these interactions are specified

Example Bayes’ Net: Insurance
Graphical Model Notation

- **Nodes:** variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)

- **Arcs:** interactions
  - Similar to CSP constraints
  - Indicate “direct influence” between variables

- For now: imagine that arrows mean direct causation
Example: Coin Flips

- N independent coin flips

- No interactions between variables: absolute independence

Example: Traffic

- Variables:
  - R: It rains
  - T: There is traffic

- Model 1: independence

- Model 2: rain causes traffic

- Why is an agent using model 2 better?
Example: Traffic II

- Let's build a causal graphical model

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

Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!
Bayes’ Net Semantics

- Let’s formalize the semantics of a Bayes’ net
- A set of nodes, one per variable $X$
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over $X$, one for each combination of parents’ values
    $$P(X|a_1 \ldots a_n)$$
  - CPT: conditional probability table
  - Description of a noisy “causal” process

$A$ Bayes net = Topology (graph) + Local Conditional Probabilities

Probabilities in BNs

- Bayes’ nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:
    $$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$$
  - Example:
    $$P(\text{cavity, catch, } \neg \text{toothache})$$

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies
Example: Coin Flips

\[ P(h, h, t, h) = \]

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

Example: Traffic

\[ P(r, \neg t) = \]

23

24
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 $P(X | e)$?
  - 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

- Let’s formalize the semantics of a Bayes’ net
- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over X, one for each combination of parents’ values
  \[ P(X|a_1 \ldots a_n) \]
  - CPT: conditional probability table
  - Description of a noisy “causal” process

\[ A \text{ Bayes net} = \text{Topology (graph)} + \text{Local Conditional Probabilities} \]

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: Alarm Network

\[ \prod_i P(X_i | \text{Parents}(X_i)) = P(B) \cdot P(E) \cdot P(A | B, E) \cdot P(J | A) \cdot P(M | A) \]

Example: Alarm Network

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>0.001</td>
</tr>
<tr>
<td>¬b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>0.002</td>
</tr>
<tr>
<td>¬e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

| A  | J   | P(J | A) |
|----|-----|-------|
| a  | j   | 0.9   |
| a  | ¬j  | 0.1   |
| ¬a | j   | 0.05  |
| ¬a | ¬j  | 0.95  |

| A  | M   | P(M | A) |
|----|-----|-------|
| a  | m   | 0.7   |
| a  | ¬m  | 0.3   |
| ¬a | m   | 0.01  |
| ¬a | ¬m  | 0.99  |
Size of a Bayes’ Net

- How big is a joint distribution over $N$ Boolean variables?
  $2^N$

- How big is an $N$-node net if nodes have up to $k$ parents?
  $O(N \times 2^{k+1})$

- 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 (coming)

Bayes’ Nets So Far

- We now know:
  - What is a Bayes’ net?
  - What joint distribution does a Bayes’ net encode?

- Next: properties of that joint distribution (independence)
  - 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

- Tomorrow: how to compute posteriors quickly (inference)
Conditional Independence

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

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

- (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!
Topography 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.
- Full conditioning can encode any distribution.

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:

- 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?
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?
  \[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} \]
  \[ = P(z|y) \quad \text{Yes!} \]

- Evidence along the chain “blocks” the influence

Common Cause

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

- Are X and Z independent given Y?
  \[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} \]
  \[ = P(z|y) \quad \text{Yes!} \]

- Observing the cause blocks influence between effects.
Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: the ballgame and the rain cause traffic, but they are not correlated
    - Still need to prove they must be (try it!)
  - Are X and Z independent given Y?
    - No: seeing traffic puts the rain and the ballgame in competition as explanation?
    - This is backwards from the other cases
      - Observing the effect enables influence between effects.

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 (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

[Demo]
Example

\begin{align*}
A & \perp W & \text{Yes} \\
A & \perp W|R
\end{align*}

Example

\begin{align*}
L & \perp T'|T & \text{Yes} \\
L & \perp B & \text{Yes} \\
L & \perp B|T \\
L & \perp B|T' \\
L & \perp B|T', R & \text{Yes}
\end{align*}
Example

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

- **Questions:**
  \[
  T \perp D \\
  T \perp D | R \quad \text{Yes} \\
  T \perp D | R, S
  \]

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

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

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

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

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

Example: Reverse Traffic

- Reverse causality?

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

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

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

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

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

\[
\begin{align*}
X_1 & \quad & X_2 \\
\begin{array}{c|c}
h & 0.5 \\
t & 0.5 \\
\end{array} & \quad & \begin{array}{c|c}
h & 0.5 \\
t & 0.5 \\
\end{array}
\]

- Adding arcs isn’t wrong, it’s just inefficient

Changing Bayes’ Net Structure

- The same joint distribution can be encoded in many different Bayes’ nets
  - Causal structure tends to be the simplest

- Analysis question: given some edges, what other edges do you need to add?
  - One answer: fully connect the graph
  - Better answer: don’t make any false conditional independence assumptions
Example: Alternate Alarm

If we reverse the edges, we make different conditional independence assumptions.

To capture the same joint distribution, we have to add more edges to the graph.

Summary

- Bayes nets compactly encode joint distributions.
- Guaranteed independencies of distributions can be deduced from BN graph structure.
- D-separation gives precise conditional independence assumptions.
- A Bayes’ net’s joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution.