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:
  \[ P(X_i | x_{e1}, \ldots, x_{ek}) \]
  - 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

\[
\begin{array}{c|c|c|c}
T & B & G & P(T, B, G) \\
\hline
T & b & g & 0.16 \\
T & b & \neg g & 0.16 \\
T & \neg b & g & 0.24 \\
T & \neg b & \neg g & 0.04 \\
\neg T & b & g & 0.04 \\
\neg T & b & \neg g & 0.24 \\
\neg T & \neg b & g & 0.06 \\
\neg T & \neg b & \neg g & 0.06 \\
\end{array}
\]

Example: Independence

- N fair, independent coin flips:

\[
P(X_1) = \begin{cases} H & 0.5 \\ T & 0.5 \end{cases}, \quad P(X_2) = \begin{cases} H & 0.5 \\ T & 0.5 \end{cases}, \quad \ldots, \quad P(X_n) = \begin{cases} H & 0.5 \\ T & 0.5 \end{cases}
\]

\[
P(X_1, X_2, \ldots, X_n) = P(X_1) \cdot P(X_2) \cdots P(X_n)
\]

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: \( X \indep Y \)
- 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?

\[
P_1(T, W) \]

\[
P_2(T, W) \]

\[
P(T) \]

\[
P(W) \]

\[
P_1(T, W) = \begin{cases} \text{warm} & 0.5 \\ \text{cold} & 0.5 \end{cases}, \quad P_2(T, W) = \begin{cases} \text{warm} & 0.3 \\ \text{cold} & 0.2 \end{cases}
\]

\[
P(T) = \begin{cases} \text{warm} & 0.5 \\ \text{cold} & 0.5 \end{cases}, \quad P(W) = \begin{cases} \text{sun} & 0.6 \\ \text{rain} & 0.4 \end{cases}
\]
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:
  - $P(\text{catch} | \text{toothache}, \text{cavity}) = P(\text{catch} | \text{cavity})$
- The same independence holds if I don't have a cavity:
  - $P(\text{catch} | \text{toothache}, \neg \text{cavity}) = P(\text{catch} | \neg \text{cavity})$
- Catch is conditionally independent of Toothache given Cavity:
  - $P(\text{Catch} | \text{Toothache, Cavity}) = P(\text{Catch} | \text{Cavity})$
- Equivalent statements:
  - $P(\text{Toothache} | \text{Catch, Cavity}) = P(\text{Toothache} | \text{Cavity})$
  - $P(\text{Toothache, Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity})$
  - One can be derived from the other easily.

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

Conditional independence is our most basic and robust form of knowledge about uncertain environments:

- $\forall x, y, z : P(x, y | z) = P(x | z) P(y | z)$
- What about this domain:
  - Traffic
  - Umbrella
  - Rain
- What about fire, smoke, alarm?

The Chain Rule

- Can always factor any joint distribution as an incremental product of conditional distributions
  - $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2) \ldots$
  - $P(X_1, X_2, \ldots X_n) = \prod P(X_i | X_1, \ldots X_{i-1})$
- 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:
  - $P(\text{Traffic, Rain, Umbrella}) = P(\text{Rain}) P(\text{Traffic} | \text{Rain}) P(\text{Umbrella} | \text{Rain, Traffic})$
- With assumption of conditional independence:
  - $P(\text{Traffic, Rain, Umbrella}) = P(\text{Rain}) P(\text{Traffic} | \text{Rain}) P(\text{Umbrella} | \text{Rain})$
- 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
  - $P(\text{t} | g) = 0.8$
  - $P(\text{t} | \neg g) = 0.4$
  - $P(\text{b} | g) = 0.4$
  - $P(\text{b} | \neg g) = 0.8$

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

Example Bayes’ Net: Car

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

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?

No interactions between variables: absolute independence
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:
    \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]
  - 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

\[
P(X_1) = \begin{array}{c|c}
h & 0.5 \\
\hline
\neg h & 0.5
\end{array} \quad ~ \quad P(X_2) = \begin{array}{c|c}
h & 0.5 \\
\hline
\neg h & 0.5
\end{array} \quad ~ \quad \ldots \quad ~ \quad \ldots \quad ~ \quad P(X_n) = \begin{array}{c|c}
h & 0.5 \\
\hline
\neg h & 0.5
\end{array}
\]

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

Example: Traffic

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

Only distributions whose variables are absolutely independent can be represented by a Bayes’ net with no arcs.
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 Bayes net = Topology (graph) + 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

\[
P(b, e, \neg a, j, m) = \prod P(X_i|\text{Parents}(X_i)) = P(B) \cdot P(E) \cdot P(A|B, E) \cdot P(J|A) \cdot P(M|A)
\]
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
    \[ P(x, y) = P(x)P(y) \quad \quad \Rightarrow \quad \quad X \perp\!\!\!\!\!\perp Y \]
  - \( X \) and \( Y \) are conditionally independent given \( Z \)
    \[ P(x, y | z) = P(x | z)P(y | z) \quad \quad \Rightarrow \quad \quad X \perp\!\!\!\!\!\perp Y | Z \]
- (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!

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

\[ Y: \text{Project due} \]
\[ X: \text{Newsgroup busy} \]
\[ Z: \text{Lab full} \]

- 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:
  - 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

\[
A \perp W \quad \text{Yes}
\]
\[
A \perp W | R
\]

Example

\[
L \perp T' | T \quad \text{Yes}
\]
\[
L \perp B \quad \text{Yes}
\]
\[
L \perp B | T
\]
\[
L \perp B | T'
\]
\[
L \perp B | T, R \quad \text{Yes}
\]

Example

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

- Questions:
  \( T \perp D \quad \text{Yes} \)
  \( T \perp D | R \)
  \( T \perp D | R, S \)

Example: Traffic

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

Example: Reverse Traffic

- Reverse causality?
**Example: Coins**

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

|       | $P(X_1)$ | $P(X_2)$ | $P(X_1)$ | $P(X_2|X_1)$ |
|-------|----------|----------|----------|--------------|
| h     | 0.5      | h        | 0.9      | h/h, 0.9     |
| l     | 0.5      | l        | 0.5      | l/l, 0.5     |

- 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