Probability recap

- Conditional probability: \( P(x|y) = \frac{P(x,y)}{P(y)} \)
- Product rule: \( P(x,y) = P(x|y)P(y) \)
- Chain rule: \( P(X_1,X_2,\ldots,X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2,X_1)\ldots = \prod_{i=1}^{n} P(X_i|X_{i-1},\ldots,X_1) \)
- \( X, Y \) independent iff: \( \forall x,y : P(x,y) = P(x)P(y) \)
- \( X \) and \( Y \) are conditionally independent given \( Z \) iff:
  \( \forall x,y,z : P(x,y|z) = P(x|z)P(y|z) \quad X \perp Y | Z \)

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
  - “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

Bayesian 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. For \( n \) variables with domain size \( d \), joint table has \( d^n \) entries — exponential in \( n \).
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
- Bayesian 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

Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)
- Arrows: interactions
  - Similar to CSP constraints
  - Indicate “direct influence” between variables
  - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don’t!)
Example: Coin Flips
- \(N\) independent coin flips
- \(X_1, X_2, \ldots, X_n\)
- 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
  - \(P(X|\alpha_1 \ldots \alpha_m)\)
- 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:
    \[ P(+\text{cavity}, +\text{catch}, -\text{toothache}) \]
  - 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

\[ X_1 \quad X_2 \quad \ldots \quad X_n \]

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

\[
\begin{array}{cc}
\text{h} & 0.5 \\
\text{t} & 0.5 \\
\end{array}
\quad \begin{array}{cc}
\text{h} & 0.5 \\
\text{t} & 0.5 \\
\end{array}
\quad \begin{array}{cc}
\text{h} & 0.5 \\
\text{t} & 0.5 \\
\end{array}
\]

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

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

Example: Traffic

\[ P(R) \]

\[
P(+r, -t) =
\]

\[
P(T|R) \]

\[
P(+, +, +) =
\]

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>

Example Bayes’ Net: Insurance

Example Bayes’ Net: Car

Build your own Bayes nets!

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)

Representing Joint Probability Distributions

- Table representation:
  number of parameters: \(d^n - 1\)
- Chain rule representation:
  \(P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_1 \ldots x_{i-1})\)
  number of parameters: \(d^{n-1} + d(d-1) + d^2(d-1) + \cdots + d^n(d-1) = d^{n-1}\)
  Size of CPT = (number of different joint instantiations of the preceding variables)
  times (number of values current variable can take on minus 1)
- Both can represent any distribution over the n random variables.
  Makes sense same number of parameters needs to be stored.
- Chain rule applies to all orderings of the variables, so for a given distribution we can represent it in \(n! = n\) factorial = \(n(n-1)(n-2) \ldots 2.1\) different ways with the chain rule

Bayes’ Nets

- Representation
  - Informal first introduction of Bayes’ nets through causality “intuition”
  - More formal introduction of Bayes’ nets
- Conditional Independences
- Probabilistic Inference
- Learning Bayes’ Nets from Data

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

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

\[
P(T)
\begin{array}{c|c|c|c}
\text{T} & \text{r} & \text{t} & \text{r} \rightarrow \text{t} \\
\hline
\text{r} & 1/4 & 3/4 & 1/16 \\
\text{t} & 3/4 & 1/4 & 3/16 \\
\end{array}
\]

\[
P(T|R)
\begin{array}{c|c|c|c|c}
& \text{r} & \text{t} & \text{r} \leftarrow \text{t} & \text{r} \rightarrow \text{t} \\
\hline
\text{r} & 1/4 & 3/4 & 1/16 & 3/16 \\
\text{t} & 3/4 & 1/4 & 3/16 & 1/16 \\
\end{array}
\]
Example: Reverse Traffic

- Reverse causality?

\[
\begin{array}{c|c|c|c}
T & P(T) & P(T, R) \\
\hline
r & 8/16 & 9/16 \\
\hline
-t & 7/16 & 7/16 \\
\end{array}
\]

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

Example: Coins

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

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

Bayes’ Nets

- Representation
  - Informal first introduction of Bayes’ nets through causality “intuition”
  - More formal introduction of Bayes’ nets

- Conditional Independences

- Probabilistic Inference

- Learning Bayes’ Nets from Data

Bayes Nets: Assumptions

- To go from chain rule to Bayes’ net representation, we made the following assumption about the distribution:
  \[ P(x_i | x_1 \cdots x_{i-1}) = P(x_i | \text{parents}(X_i)) \]

- Turns out that probability distributions that satisfy the above ("chain-rule \rightarrow Bayes net") conditional independence assumptions
  - often can be guaranteed to have many more conditional independences
  - These guaranteed additional conditional independences can be read off directly from the graph

- Important for modeling: understand assumptions made when choosing a Bayes net graph

Example

- Conditional independence assumptions directly from simplifications in chain rule:

- Additional implied conditional independence assumptions?

Independence in a BN

- Given a Bayes net graph
  - Important question:
    Are two nodes guaranteed to be independent given certain evidence?
  - Equivalent question:
    Are two nodes independent given the evidence in all distributions that can be encoded with the Bayes net graph?
  - Before proceeding: How about opposite question: Are two nodes guaranteed to be dependent given certain evidence?

- No! For any BN graph you can choose all CPT’s such that all variables are independent by having \( P(X | \text{Pa}(X)) = \text{pa}(X) \) not depend on the value of the parents. Simple way of doing so: pick all entries in all CPTs equal to 0.5 (assuming binary variables)
Independence in a BN

- Given a Bayes net graph
  Are two nodes guaranteed to be independent given certain evidence?
- If no, can prove with a counter example
  i.e., pick a distribution that can be encoded with the BN graph, i.e., pick a set of CPT’s, and show that the independence assumption is violated
- If yes,
  - For now we are able to prove using algebra (tedious in general)
  - Next we will study an efficient graph-based method to prove yes: “D-separation”

D-separation: Outline

- Study independence properties for triples
- Any complex example can be analyzed by considering relevant triples

Causal Chains

- This configuration is a “causal chain”

\[
P(x, y, z) = P(x)P(y|x)P(z|y)
\]
- Is it guaranteed that X is independent of Z? No!
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
  - Example: \( P(y|x) = 1 \) if \( y=x \), 0 otherwise
    \( P(z|y) = 1 \) if \( z=y \), 0 otherwise
    Then we have \( P(x|z) = 1 \) if \( x=z \), 0 otherwise
    hence X and Z are not independent in this example

Common Cause

- Another basic configuration: two effects of the same cause
- Is it guaranteed that X and Z are independent?
  - No!
  - Counterexample:
    Choose \( P(X|Y) = 1 \) if \( x=y \), 0 otherwise,
    \( P(Z|Y) = 1 \) if \( z=y \), 0 otherwise.
    Then \( P(x|z) = 1 \) if \( x=z \) and 0 otherwise, hence X and Z are not independent in this example and hence it is not guaranteed that if a distribution can be encoded with the Bayes’ net structure on the right that X and Z are independent in that distribution
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 an effect activates influence between possible causes.

D-Separation

- Given query $X_i \perp\!\!\!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
- Shade all evidence nodes
- For all (undirected!) paths between and
  - Check whether path is active
    - If active return: 
      - not guaranteed that $X_i \perp\!\!\!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
  - (If reaching this point all paths have been checked and shown inactive)
    - Return: guaranteed tat $X_i \perp\!\!\!\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$

Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I’m sad
- Questions:
  - $T \perp D$
  - $T \perp D \mid R$
  - $T \perp D \mid R, S$
All Conditional Independences

- Given a Bayes net structure, can run d-separation to build a complete list of conditional independences that are guaranteed to be true, all of the form

  \[ X_i \perp \perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\} \]

Possible to have same full list of conditional independencies for different BN graphs?

- Yes!
- Examples:

  - If two Bayes’ Net graphs have the same full list of conditional independencies then they are able to encode the same set of 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
- Full conditioning can encode any distribution

Bayes Nets Representation Summary

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

Bayes’ Nets

- Representation
- Conditional Independences
- Probabilistic Inference
  - Enumeration (exact, exponential complexity)
  - Variable elimination (exact, worst-case exponential complexity, often better)
  - Probabilistic inference is NP-complete
  - Sampling (approximate)
- Learning Bayes’ Nets from Data