CSEP 573: Artificial Intelligence

Graphical Models

slides adapted from
Stuart Russel, Dan Klein, Pieter Abbeel from ai.berkeley.edu
And Hanna Hajishirzi, Dan Weld, Luke Zettlemoyer
Reminder: elementary probability

- **Basic laws:** $0 \leq P(\omega) \leq 1 \quad \sum_{\omega \in \Omega} P(\omega) = 1$

- **Events:** subsets of $\Omega$: $P(A) = \sum_{\omega \in A} P(\omega)$

- **Random variable** $X(\omega)$ has a value in each $\omega$
  - Distribution $P(X)$ gives probability for each possible value $x$
  - Joint distribution $P(X,Y)$ gives total probability for each combination $x,y$

- **Summing out/marginalization:** $P(X=x) = \sum_y P(X=x, Y=y)$

- **Conditional probability:** $P(X|Y) = P(X,Y)/P(Y)$

- **Product rule:** $P(X|Y)P(Y) = P(X,Y) = P(Y|X)P(X)$
  - Generalize to chain rule: $P(X_1,..,X_n) = \prod_i P(X_i | X_1,..,X_{i-1})$
Bayes’ Nets: Big Picture

Encoding Complex Distributions

In 12 Easy Steps!
Bayes nets: a technique for describing complex joint distributions (models) using simple, conditional distributions
- A subset of the general class of graphical models
- Also called belief networks

Use local causality/conditional independence:
- the world is composed of many variables,
- each interacting locally with a few others
Bayes Nets

Part I: Representation

Part II: Independence

Part III: Exact inference

Part IV: Approximate Inference
Graphical Model Notation

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

- **Arcs:** interactions
  - Indicate “direct influence” between variables
  - Formally: encode conditional independence (more on this later)
Example Bayes’ Net: Coin Flips

- N independent coin flips

- No interactions between variables: absolute independence
Conditional Independence: Traffic

- What about this domain:
  - Traffic
  - Umbrella
  - Raining
Example Bayes’ Net: Traffic

- Variables:
  - T: There is traffic
  - U: I’m holding my umbrella
  - R: It rains
Conditional Independence: Fire

- What about this domain:
  - Fire
  - Smoke
  - Alarm
Example Bayes’ Net: Smoke alarm

- **Variables:**
  - F: There is fire
  - S: There is smoke
  - A: Alarm sounds
Example Bayes’ Net: Car Insurance

- **Age**
- **YearsLicensed**
- **DrivingSkill**
- **DrivingRecord**
- **DrivingBehavior**
- **GoodStudent**
- **RiskAversion**
- **SocioEcon**
- **ExtraCar**
- **MakeModel**
- **VehicleYear**
- **Mileage**
- **SafetyFeatures**
- **Garaged**
- **AntiTheft**
- **Ruggedness**
- **CarValue**
- **Airbag**
- **Theft**
- **Cushioning**
- **Accident**
- **OwnCarDamage**
- **OwnCarCost**
- **MedicalCost**
- **LiabilityCost**
- **OtherCost**
- **PropertyCost**

Why do conditional independence?-- Ghostbusters

- A ghost is in the grid somewhere

- Sensor readings tell how close a square is to the ghost
  - On the ghost: usually red
  - 1 or 2 away: mostly orange
  - 3 or 4 away: typically yellow
  - 5+ away: often green

- Click on squares until confident of location, then “bust”
Video of Demo Ghostbusters with Probability

$P(\text{ghost is in this position given all of the evidence that we have seen so far})$
Ghostbusters model

- **Variables and ranges:**
  - $G$ (ghost location) in $\{(1,1),\ldots,(3,3)\}$
  - $C_{x,y}$ (color measured at square $x,y$) in \{red, orange, yellow, green\}

- **Ghostbuster physics:**
  - *Uniform prior distribution* over ghost location: $P(G)$
  - *Sensor model*: $P(C_{x,y} \mid G)$ (depends only on distance to $G$)
    - E.g. $P(C_{1,1} = \text{yellow} \mid G = (1,1)) = 0.1$
Ghostbusters model, contd.

- $P(G, C_{1,1}, \ldots, C_{3,3})$ has ...
  - $9 \times 4^9 = 2,359,296$ entries!
  - $|G| = 9$, $|C_{i,j}| = 4$; Grid squares times size of each

- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?

- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ depends only on distance to $G$
    - So $P(C_{1,1} = \text{yellow} \mid G = (2,3)) = P(C_{1,1} = \text{yellow} \mid G = (2,3), C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is conditionally independent of $C_{1,2}$ given $G$
Apply the chain rule to decompose the joint probability model:

\[ P(G, C_{1,1}, \ldots, C_{3,3}) = P(G) P(C_{1,1} | G) P(C_{1,2} | G, C_{1,1}) P(C_{1,3} | G, C_{1,1}, C_{1,2}) \ldots P(C_{3,3} | G, C_{1,1}, \ldots, C_{3,2}) \]

Now simplify using conditional independence:

\[ P(G, C_{1,1}, \ldots, C_{3,3}) = P(G) P(C_{1,1} | G) P(C_{1,2} | G) P(C_{1,3} | G) \ldots P(C_{3,3} | G) \]

I.e., conditional independence properties of ghostbuster physics simplify the probability model from \textit{exponential} to \textit{quadratic} in the number of squares

\[ |P(C_{i,i} | G)| = 4 \times 9 \text{ rather than } |P(C_{3,3} | G, C_{1,1}, \ldots, C_{3,2})| = 4 \times 9 \times 4^8 \]

In total: \(9 + 9 \times (4 \times 9) = 333\) entries, before was \(9 \times 4^9 = 2,359,296\) entries

This is called a \textit{Naïve Bayes} model:

- One discrete query variable (often called the \textit{class} or \textit{category} variable)
- All other variables are (potentially) evidence variables
- Evidence variables are all conditionally independent given the query variable
Ghostbusters Full Joint
Ghostbusters Naïve Bayes

\[ C_{1,1} \quad C_{1,2} \quad \ldots \quad C_{i,j} \quad \ldots \quad C_{3,2} \quad C_{3,3} \]
Bayes Net Syntax and Semantics
Bayes’ Net Syntax

- A set of nodes, one per variable $X_i$
- A directed, acyclic graph
- A conditional distribution for each node given its parent variables in the graph
  - CPT (conditional probability table)
    - each row is a distribution for child given values of its parents

Bayes net = Topology (graph) + Local Conditional Probabilities
Example: Alarm Network

<table>
<thead>
<tr>
<th>$P(B)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.001</td>
</tr>
<tr>
<td>false</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(E)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.002</td>
</tr>
<tr>
<td>false</td>
<td>0.998</td>
</tr>
</tbody>
</table>

| $B$ | $E$ | $P(A|B,E)$ |   |
|-----|-----|-----------|---|
| true| true| 0.95      | 0.05 |
| true| false| 0.94     | 0.06 |
| false| true| 0.29    | 0.71 |
| false| false| 0.001 | 0.999 |

Factor size of each CPT:

$$d \prod d_i$$

Parent range sizes: $d_1, ..., d_k$
Child range size: $d$

Each table row must sum to 1
General formula for sparse BNs

- Suppose
  - $n$ variables
  - Maximum range size is $d$
  - Maximum number of parents is $k$
- Full joint distribution has size $O(d^n)$
- Bayes net has size $O(n \cdot d^k)$
  - Linear scaling with $n$ as long as causal structure is local
- Often $O(n \cdot d^k) \ll O(d^n)$
Bayes net global semantics

- Bayes nets encode joint distributions as product of conditional distributions on each variable:

\[ P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \text{Parents}(X_i)) \]
P(b, ¬e, a, ¬j, ¬m) =
P(b) P(¬e) P(a | b, ¬e) P(¬j | a) P(¬m | a)
= 0.001 x 0.998 x 0.94 x 0.1 x 0.3
= 0.000028
Example: Your turn

\[ P(B, e, a, j, m) = P(B) P(e) P(a|B,e) P(j|a) P(m|a) \]

\[ = 0.001 \times 0.002 \times 0.95 \times 0.29 \times 0.9 \times 0.7 \]

\[ = 0.00000120 \times 0.000365 \]

\[ = 0.000000120 \times 0.000365 \]

\[ = 0.000000120 \times 0.000365 \]
Question

- Which of the following does a Bayes’ net model explicitly?
  - The joint probability distribution?
  - The conditional probability distribution?
- Is one of the following more expressive than the other?
  - The joint probability distribution
  - The conditional probability distribution
- Why do we use Bayes’ nets?
Independence and conditional independence are important forms of probabilistic knowledge.

Bayes nets encode joint distributions efficiently by taking advantage of conditional independence:
- Global joint probability = product of local conditionals

Next: more on independence
Then: how to answer queries, i.e., compute conditional probabilities of queries given evidence.
Bayes Nets

Part I: Representation

Part II: Independence

Part III: Exact inference

- Enumeration (always exponential complexity)
- Variable elimination (worst-case exponential complexity, often better)
- Inference is NP-hard in general

Part IV: Approximate Inference
Conditional independence in BNs

- Compare the Bayes net global semantics
  \[ P(X_1, \ldots, X_n) = \prod_i P(X_i | \text{Parents}(X_i)) \]
  with the chain rule identity
  \[ P(X_1, \ldots, X_n) = \prod_i P(X_i | X_1, \ldots, X_{i-1}) \]

- Assume (without loss of generality) that \( X_1, \ldots, X_n \) sorted in topological order according to the graph (i.e., parents before children), so \( \text{Parents}(X_i) \subseteq X_1, \ldots, X_{i-1} \)

- So the Bayes net asserts conditional independences \( P(X_i | X_1, \ldots, X_{i-1}) = P(X_i | \text{Parents}(X_i)) \)
  - To ensure these are valid, choose parents for node \( X_i \) that “shield” it from other predecessors
- Every variable is conditionally independent of its non-descendants given its parents
- Conditional independence semantics \(\iff\) global semantics
A variable’s Markov blanket consists of parents, children, children’s other parents

*Every variable is conditionally independent of all other variables given its Markov blanket*
Reminder: Conditional Independence

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

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

- (Conditional) independence is a property of a distribution

- Example: \( \text{Alarm} \perp \text{Fire}|\text{Smoke} \)
Example

- Conditional independence assumptions directly from simplifications in chain rule:
  \[ P(x, y, z, w) = P(x)P(y|x)P(z|x, y)P(w|x, y, z) \]
  \[ P(x, y, z, w) = P(x)P(y|x)P(z|y)P(w|z) \]
- Additional implied conditional independence assumptions?
Example

- Conditional independence assumptions directly from simplifications in chain rule:
  \[ X \indep Z | Y \]
  \[ W \indep \{X, Y\} | Z \]
- Additional implied conditional independence assumptions?
  \[ W \indep X | Y \]
Independence in a Bayes’ Net

- Important question about a Bayes’ Net:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general)
  - If no, can prove with a counter-example
  - Example:

```
X → Y → Z
```

- Question: are X and Z necessarily independent?
  - Answer: no. Example: low pressure causes rain, which causes traffic.
  - X can influence Z, Z can influence X (via Y)
  - Addendum: they could be independent: how?
D-separation: Outline
D-separation: Outline

- Study independence properties for triples
  - Why triples?

- Analyze complex cases in terms of member triples

- D-separation: a condition / algorithm for answering such queries
This configuration is a “causal chain”

Guaranteed X 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:

Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic

In numbers:

\[ P(+y | +x) = 1, P(-y | -x) = 1, \]
\[ P(+z | +y) = 1, P(-z | -y) = 1 \]
Causal Chains

- This configuration is a “causal chain”

- Guaranteed X independent of Z given Y?

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

Yes!

- Evidence along the chain “blocks” the influence

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

- This configuration is a “common cause”

Guaranteed $X$ 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:

- Project due causes both forums busy and lab full

In numbers:

$$P( +x | +y ) = 1, P( -x | -y ) = 1,$$
$$P( +z | +y ) = 1, P( -z | -y ) = 1$$
Common Cause

- This configuration is a “common cause”

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

- Guaranteed 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.

Y: Project due  
X: Forums busy  
Z: Lab full  
Project Due!
Common Effect

- Last configuration: two causes of one effect (v-structures)

- Are X and Y independent?
  - Yes: the ballgame and the rain cause traffic, but they are not correlated

- Proof:
  \[ P(x, y) = \sum P(x, y, z) \]
Common Effect

- Last configuration: two causes of one effect (v-structures)
  - X: Raining
  - Y: Ballgame
  - Z: Traffic

- Are X and Y independent?
  - Yes: the ballgame and the rain cause traffic, but they are not correlated
  - (Proved previously)

- Are X and Y independent given Z?
  - 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.
The General Case

Conditional Independence in 3 Easy Steps!
The General Case

- General question: in a given BN, are two variables independent (given evidence)?
- Solution: analyze the graph
- Any complex example can be broken into repetitions of the three canonical cases
Reachability

- Recipe: shade evidence nodes, look for paths in the resulting graph

- Attempt 1: if two nodes are **not** connected* they are conditionally independent
  - *There does not exist an undirected path between them, excluding those blocked by a shaded node.

- 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”
Question: Are X and Y conditionally independent given evidence variables \{Z\}?
- Yes, if X and Y “d-separated” by Z
- Consider all (undirected) paths from X to Y
- No active paths = independence!

A path is active if each triple is active:
- **Causal chain**: A -> B -> C where B is unobserved (either direction)
- **Common cause**: A <- B -> C where B is unobserved
- **Common effect**: (aka \(v\)-structure)
  A -> B <- C where B or one of its descendants is observed

All it takes to block a path is a single inactive segment
D-Separation

- **Query:** $X_i \perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$  

- **Check all (undirected!) paths between $X_i$ and $X_j$**
  - If one or more active, then independence not guaranteed
    $$X_i \not\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$$
  - Otherwise (i.e. if all paths are inactive), then independence is guaranteed
    $$X_i \perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$$
def d-separated(first, second):
    for path in paths(first, second):
        path_active = True
        for triple in path:
            if not active(triple):
                path_active = False
                break
        if path_active:
            return False
    return True
Example: which assumptions apply?

\[ R \perp B \]
\[ R \perp B \mid T \]
\[ R \perp B \mid T' \]

Yes
Example: which assumptions apply?

\[
L \perp T' | T \quad \text{Yes}
\]
\[
L \perp B \quad \text{Yes}
\]
\[
L \perp B | T
\]
\[
L \perp B | T'
\]
\[
L \perp B | T, R \quad \text{Yes}
\]
Example: which assumptions apply?

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

- **Questions:**

  \[
  T \perp D \\
  T \perp D|R \\
  T \perp D|R, S
  \]

  Yes