Probabilistic Reasoning;
Network-based reasoning

COMPSCI 276, Spring 2013
Set 1: Introduction and Background
Rina Dechter

(Reading: Pearl chapter 1-2, Darwiche chapters 1,3)
Class Description

- Instructor: Rina Dechter

- Days: Tuesday & Thursday
- Time: 11:00 - 12:20 pm

- Class page:
Outline

- Why uncertainty?
- Basics of probability theory and modeling
Why Uncertainty?

- AI goal: to have a declarative, model-based, framework that allow computer system to reason.
- People reason with partial information
- Sources of uncertainty:
  - Limitation in observing the world: e.g., a physician see symptoms and not exactly what goes in the body when he performs diagnosis. Observations are noisy (test results are inaccurate)
  - Limitation in modeling the world,
  - maybe the world is not deterministic.
Example of common sense reasoning

- Explosive noise at UCI
- Parking in Cambridge
- The missing garage door
- Years to finish an undergrad degree in college
Shooting at UCI

what is the likelihood that there was a criminal activity if S1 called?
What is the probability that someone will call the police?
Why uncertainty

- **Summary of exceptions**
  - Birds fly, smoke means fire (cannot enumerate all exceptions.

- **Why is it difficult?**
  - Exception combines in intricate ways
  - e.g., we cannot tell from formulas how exceptions to rules interact:

\[
\begin{align*}
A &\rightarrow C \\
B &\rightarrow C \\
\text{-------------} \\
A \text{ and } B &\rightarrow C
\end{align*}
\]
# The problem

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>All men are mortal</td>
<td>T</td>
</tr>
<tr>
<td>All penguins are birds</td>
<td>T</td>
</tr>
<tr>
<td>Socrates is a man</td>
<td>p1</td>
</tr>
<tr>
<td>Men are kind</td>
<td>p2</td>
</tr>
<tr>
<td>Birds fly</td>
<td></td>
</tr>
<tr>
<td>T looks like a penguin</td>
<td></td>
</tr>
<tr>
<td>Turn key → car starts</td>
<td>P_n</td>
</tr>
</tbody>
</table>

**Q:** Does T fly?  
**P(Q)?** Logic?....but how we handle exceptions  
Probability: astronomical
Managing Uncertainty

- Knowledge obtained from people is almost always loaded with uncertainty
- Most rules have exceptions which one cannot afford to enumerate
- Antecedent conditions are ambiguously defined or hard to satisfy precisely
- First-generation expert systems combined uncertainties according to simple and uniform principle
- Lead to unpredictable and counterintuitive results
- Early days: logicist, new-calculist, neo-probabilist
Extensional vs Intensional Approaches

- **Extensional** (e.g., Mycin, Shortliffe, 1976) certainty factors attached to rules and combine in different ways.

  \[ P(A|B) = m \]

- **Intensional**, semantic-based, probabilities are attached to set of worlds.

  \[ P(A|B) = m \]
Certainty combination in Mycin

If A then C (x)
If B then C (y)
If C then D (z)

1. Parallel Combination:
   \[ CF(C) = x+y-xy, \text{ if } x,y>0 \]
   \[ CF(C) = (x+y)/(1-\min(x,y)), \text{ if } x,y \text{ have different sign} \]
   \[ CF(C) = x+y+xy, \text{ if } x,y<0 \]

2. Series combination...
3. Conjunction, negation

**Computational desire**: locality, detachment, modularity
The limits of modularity

Deductive reasoning: modularity and detachment

P → Q
P
-----
Q

K and P
-----
Q

K
-----
Q

Plausible Reasoning: violation of locality

Wet → rain
Wet
--------------
rain

wet → rain
Sprinkler and wet
----------------------------------
rain?
Violation of detachment

Deductive reasoning

\[
P \rightarrow Q \\
K \rightarrow P \\
K \\
------ \\
Q
\]

Plausible reasoning

\[
\text{Wet} \rightarrow \text{rain} \\
\text{Sprinkler} \rightarrow \text{wet} \\
\text{Sprinkler} \\
------ \\
\text{rain?}
\]
Burglery Example

A → B
A more credible
------------------
B more credible

IF Alarm → Burglery
A more credible (after radio)
But B is less credible

Issue: Rule from effect to causes
All frameworks for reasoning with uncertainty today are “intentional” model-based. All are based on the probability theory implying calculus and semantics.
Outline

- Why uncertainty?
- Basics of probability theory and modeling
Degrees of Belief

- Assign a **degree of belief** or **probability** in \([0, 1]\) to each world \(\omega\) and denote it by \(\Pr(\omega)\).
- The belief in, or probability of, a sentence \(\alpha\):

\[
\Pr(\alpha) \overset{\text{def}}{=} \sum_{\omega \models \alpha} \Pr(\omega).
\]

<table>
<thead>
<tr>
<th>world</th>
<th>Earthquake</th>
<th>Burglary</th>
<th>Alarm</th>
<th>(\Pr(.))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\omega_1)</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>.0190</td>
</tr>
<tr>
<td>(\omega_2)</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>.0010</td>
</tr>
<tr>
<td>(\omega_3)</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>.0560</td>
</tr>
<tr>
<td>(\omega_4)</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>.0240</td>
</tr>
<tr>
<td>(\omega_5)</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>.1620</td>
</tr>
<tr>
<td>(\omega_6)</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>.0180</td>
</tr>
<tr>
<td>(\omega_7)</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>.0072</td>
</tr>
<tr>
<td>(\omega_8)</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>.7128</td>
</tr>
</tbody>
</table>
Properties of Beliefs

- A bound on the belief in any sentence:

$$0 \leq Pr(\alpha) \leq 1 \quad \text{for any sentence } \alpha.$$ 

- A baseline for inconsistent sentences:

$$Pr(\alpha) = 0 \quad \text{when } \alpha \text{ is inconsistent.}$$ 

- A baseline for valid sentences:

$$Pr(\alpha) = 1 \quad \text{when } \alpha \text{ is valid.}$$
Properties of Beliefs

- The belief in a sentence given the belief in its negation:
  \[ \Pr(\alpha) + \Pr(\neg\alpha) = 1. \]

---

Example

\[
\begin{align*}
\Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\
\Pr(\neg\text{Burglary}) &= \Pr(\omega_3) + \Pr(\omega_4) + \Pr(\omega_7) + \Pr(\omega_8) = .8
\end{align*}
\]
Properties of Beliefs

- The belief in a disjunction:

  \[ \Pr(\alpha \lor \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha \land \beta). \]

- Example:

  \[
  \begin{align*}
  \Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
  \Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\
  \Pr(\text{Earthquake} \land \text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) = .02 \\
  \Pr(\text{Earthquake} \lor \text{Burglary}) &= .1 + .2 - .02 = .28
  \end{align*}
  \]
Properties of Beliefs

- The belief in a disjunction:

\[ \Pr(\alpha \lor \beta) = \Pr(\alpha) + \Pr(\beta) \quad \text{when } \alpha \text{ and } \beta \text{ are mutually exclusive.} \]
Quantify uncertainty about a variable $X$ using the notion of entropy:

$$\text{ENT}(X) \overset{\text{def}}{=} - \sum_x \Pr(x) \log_2 \Pr(x),$$

where $0 \log 0 = 0$ by convention.

<table>
<thead>
<tr>
<th></th>
<th>Earthquake</th>
<th>Burglary</th>
<th>Alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>.1</td>
<td>.2</td>
<td>.2442</td>
</tr>
<tr>
<td>false</td>
<td>.9</td>
<td>.8</td>
<td>.7558</td>
</tr>
<tr>
<td>ENT(.)</td>
<td>.469</td>
<td>.722</td>
<td>.802</td>
</tr>
</tbody>
</table>
Entropy

- The entropy for a binary variable $X$ and varying $p = \Pr(X)$.
- Entropy is non-negative.
- When $p = 0$ or $p = 1$, the entropy of $X$ is zero and at a minimum, indicating no uncertainty about the value of $X$.
- When $p = \frac{1}{2}$, we have $\Pr(X) = \Pr(\neg X)$ and the entropy is at a maximum (indicating complete uncertainty).
Bayes Conditioning

Alpha and beta are events

Closed form for Bayes conditioning:

$$\Pr(\alpha|\beta) = \frac{\Pr(\alpha \wedge \beta)}{\Pr(\beta)}.$$  

Defined only when $\Pr(\beta) \neq 0.$
### Degrees of Belief

<table>
<thead>
<tr>
<th>world</th>
<th>Earthquake</th>
<th>Burglary</th>
<th>Alarm</th>
<th>(\Pr(.))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\omega_1)</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>.0190</td>
</tr>
<tr>
<td>(\omega_2)</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>.0010</td>
</tr>
<tr>
<td>(\omega_3)</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>.0560</td>
</tr>
<tr>
<td>(\omega_4)</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>.0240</td>
</tr>
<tr>
<td>(\omega_5)</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>.1620</td>
</tr>
<tr>
<td>(\omega_6)</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>.0180</td>
</tr>
<tr>
<td>(\omega_7)</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>.0072</td>
</tr>
<tr>
<td>(\omega_8)</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>.7128</td>
</tr>
</tbody>
</table>

\[
\Pr(\text{Earthquake}) = \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) = .2 \\
\Pr(\neg \text{Burglary}) = .8 \\
\Pr(\text{Alarm}) = .2442
\]
**Belief Change**

_Burglary is independent of Earthquake_

<table>
<thead>
<tr>
<th>Conditioning on evidence Earthquake:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pr(\text{Burglary})$ = .2</td>
</tr>
<tr>
<td>$\Pr(\text{Burglary}</td>
</tr>
<tr>
<td>$\Pr(\text{Alarm})$ = .2442</td>
</tr>
<tr>
<td>$\Pr(\text{Alarm}</td>
</tr>
</tbody>
</table>

The belief in Burglary is not changed, but the belief in Alarm increases.
Belief Change

Earthquake is independent of burglary

Conditioning on evidence Burglary:

\[
\begin{align*}
\Pr(\text{Alarm}) &= .2442 \\
\Pr(\text{Alarm}|\text{Burglary}) &\approx .905 \\
\Pr(\text{Earthquake}) &= .1 \\
\Pr(\text{Earthquake}|\text{Burglary}) &= .1
\end{align*}
\]

The belief in Alarm increases in this case, but the belief in Earthquake stays the same.
The belief in Burglary increases when accepting the evidence Alarm. How would such a belief change further upon obtaining more evidence?

- Confirming that an Earthquake took place:
  \[
  \Pr(\text{Burglary}|\text{Alarm}) \approx 0.741 \\
  \Pr(\text{Burglary}|\text{Alarm} \land \text{Earthquake}) \approx 0.253 \\
  \]
  We now have an explanation of Alarm.

- Confirming that there was no Earthquake:
  \[
  \Pr(\text{Burglary}|\text{Alarm}) \approx 0.741 \\
  \Pr(\text{Burglary}|\text{Alarm} \land \neg\text{Earthquake}) \approx 0.957 \\
  \]
  New evidence will further establish burglary as an explanation.
Conditional Independence

Pr finds α conditionally independent of β given γ iff

$$Pr(\alpha | \beta \land \gamma) = Pr(\alpha | \gamma) \quad \text{or} \quad Pr(\beta \land \gamma) = 0.$$  

Another definition

$$Pr(\alpha \land \beta | \gamma) = Pr(\alpha | \gamma)Pr(\beta | \gamma) \quad \text{or} \quad Pr(\gamma) = 0.$$
Pr finds $\mathbf{X}$ independent of $\mathbf{Y}$ given $\mathbf{Z}$, denoted $l_{Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, means that $Pr$ finds $x$ independent of $y$ given $z$ for all instantiations $x$, $y$ and $z$.

Example

$\mathbf{X} = \{A, B\}$, $\mathbf{Y} = \{C\}$ and $\mathbf{Z} = \{D, E\}$, where $A, B, C, D$ and $E$ are all propositional variables. The statement $l_{Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is then a compact notation for a number of statements about independence:

- $A \land B$ is independent of $C$ given $D \land E$;
- $A \land \neg B$ is independent of $C$ given $D \land E$;
- ...;
- $\neg A \land \neg B$ is independent of $\neg C$ given $\neg D \land \neg E$;

That is, $l_{Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is a compact notation for $4 \times 2 \times 4 = 32$ independence statements of the above form.
Conditional Entropy

To quantify the average uncertainty about the value of $X$ after observing the value of $Y$.

**Conditional entropy of a variable $X$ given another variable $Y$**

$$\text{ENT}(X|Y) \overset{\text{def}}{=} \sum_y \Pr(y)\text{ENT}(X|y),$$

where

$$\text{ENT}(X|y) \overset{\text{def}}{=} -\sum_x \Pr(x|y) \log_2 \Pr(x|y).$$

* Entropy never increases after conditioning:
  $$\text{ENT}(X|Y) \leq \text{ENT}(X).$$
* Observing the value of $Y$ reduces our uncertainty about $X$.
* For a particular value $y$, we may have $\text{ENT}(X|y) > \text{ENT}(X)$. 
Conditional Entropy

<table>
<thead>
<tr>
<th></th>
<th>Burglary</th>
<th>Burglary</th>
<th>Burglary</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>.2</td>
<td>.741</td>
<td>.025</td>
</tr>
<tr>
<td>false</td>
<td>.8</td>
<td>.259</td>
<td>.975</td>
</tr>
<tr>
<td>ENT(.)</td>
<td>.722</td>
<td>.825</td>
<td>.169</td>
</tr>
</tbody>
</table>

The conditional entropy of Burglary given Alarm is then:

\[
\text{ENT}(\text{Burglary} | \text{Alarm}) = \text{ENT}(\text{Burglary} | \text{Alarm} = \text{true}) \Pr(\text{Alarm} = \text{true}) + \\
\quad \text{ENT}(\text{Burglary} | \text{Alarm} = \text{false}) \Pr(\text{Alarm} = \text{false})
\]

\[= .329,\]

indicating a decrease in the uncertainty about variable Burglary.
Further Properties of Beliefs

**Chain rule**

\[
\Pr(\alpha_1 \land \alpha_2 \land \ldots \land \alpha_n) = \Pr(\alpha_1|\alpha_2 \land \ldots \land \alpha_n) \Pr(\alpha_2|\alpha_3 \land \ldots \land \alpha_n) \ldots \Pr(\alpha_n).
\]

**Case analysis (law of total probability)**

\[
\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha \land \beta_i),
\]

where the events \(\beta_1, \ldots, \beta_n\) are mutually exclusive and exhaustive.
Further Properties of Beliefs

Another version of case analysis

\[ \Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha|\beta_i)\Pr(\beta_i), \]

where the events \( \beta_1, \ldots, \beta_n \) are mutually exclusive and exhaustive.

Two simple and useful forms of case analysis are these:

\[ \Pr(\alpha) = \Pr(\alpha \land \beta) + \Pr(\alpha \land \neg\beta) \]
\[ \Pr(\alpha) = \Pr(\alpha|\beta)\Pr(\beta) + \Pr(\alpha|\neg\beta)\Pr(\neg\beta). \]

The main value of case analysis is that, in many situations, computing our beliefs in the cases is easier than computing our beliefs in \( \alpha \). We shall see many examples of this phenomena in later chapters.
Bayes rule

\[ \Pr(\alpha|\beta) = \frac{\Pr(\beta|\alpha) \Pr(\alpha)}{\Pr(\beta)} . \]

- Classical usage: \( \alpha \) is perceived to be a cause of \( \beta \).
- Example: \( \alpha \) is a disease and \( \beta \) is a symptom—
- Assess our belief in the cause given the effect.
- Belief in an effect given its cause, \( \Pr(\beta|\alpha) \), is usually more readily available than the belief in a cause given one of its effects, \( \Pr(\alpha|\beta) \).
Difficult: Complexity in model construction and inference

- In Alarm example:
  - 31 numbers needed,
  - Quite unnatural to assess: e.g.

\[ P(B = y, E = y, A = y, J = y, M = y) \]

- Computing \( P(B=y|M=y) \) takes 29 additions.

- In general,
  - \( P(X_1, X_2, \ldots, X_n) \) needs at least \( 2^n - 1 \) numbers to specify the joint probability. Exponential model size.
  - Knowledge acquisition difficult (complex, unnatural),
  - Exponential storage and inference.
Overcome the problem of exponential size by exploiting conditional independence

- The chain rule of probabilities:

\[
P(X_1, X_2) = P(X_1)P(X_2|X_1)
\]
\[
P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)
\]
\[\ldots\]
\[
P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)\ldots P(X_n|X_1, \ldots, X_{n-1})
\]
\[
= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}).
\]

- No gains yet. The number of parameters required by the factors is:

\[
2^{n-1} + 2^{n-1} + \ldots + 1 = 2^n - 1.
\]
Conditional Independence

- About $P(X_i|X_1, \ldots, X_{i-1})$:
  - Domain knowledge usually allows one to identify a subset $pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ such that
    - Given $pa(X_i)$, $X_i$ is independent of all variables in $\{X_1, \ldots, X_{i-1}\} \setminus pa(X_i)$, i.e.
    $P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|pa(X_i))$
  - Then
    $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i))$
    - Joint distribution factorized.
    - The number of parameters might have been substantially reduced.
Example

$P(B,E,A,J,M)=$?
Example continued

\[ P(B, E, A, J, M) = P(B)P(E|B)P(A|B, E)P(J|B, E, A)P(M|B, E, A, J) = P(B)P(E)P(A|B, E)P(J|A)P(M|A) \text{(Factorization)} \]

- \(pa(B) = \{\}, \ pa(E) = \{\},pa(A) = \{B, E\}, \ pa(J) = \{A\},pa(M) = \{A\}.\)

- Conditional probabilities tables (CPT)

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>.01</td>
</tr>
<tr>
<td>N</td>
<td>.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>.02</td>
</tr>
<tr>
<td>N</td>
<td>.98</td>
</tr>
</tbody>
</table>

| A   | B   | E   | P(A|B, E) |
|-----|-----|-----|----------|
| Y   | Y   | Y   | .95      |
| N   | Y   | Y   | .05      |
| Y   | Y   | N   | .94      |
| N   | Y   | N   | .06      |
| Y   | N   | Y   | .29      |
| N   | Y   | N   | .71      |
| Y   | N   | N   | .001     |
| N   | N   | N   | .999     |
Example continued

- Model size reduced from $31$ to $1+1+4+2+2=10$
- Model construction easier
  - Fewer parameters to assess.
  - Parameters more natural to assess: e.g.
    \[
    P(B = Y), P(E = Y), P(A = Y|B = Y, E = Y),
    \]
    \[
    P(J = Y|A = Y), P(M = Y|A = Y)
    \]
- Inference easier. Will see this later.
From Factorizations to Bayesian Networks

Graphically represent the conditional independency relationships:

- construct a directed graph by drawing an arc from $X_j$ to $X_i$ iff $X_j \in pa(X_i)$

\[
pa(B) = \{\}, \quad pa(E) = \{\}, \quad pa(A) = \{B, E\}, \quad pa(J) = \{A\}, \quad pa(M) = \{A\}.
\]

- Also attach the conditional probability (table) $P(X_i|pa(X_i))$ to node $X_i$.

- What results in is a **Bayesian network**. Also known as belief network, probabilistic network.
A **Bayesian network** is:

- An directed acyclic graph (DAG), where
- Each node represents a random variable
- And is associated with the conditional probability of the node given its parents.
Bayesian Networks: Representation

Bayesian Network: \( (G, \Theta) \)

- **P(S)**
- **P(C|S)**
- **P(B|S)**
- **P(X|C,S)**
- **P(D|C,B)**

\[ P(S, C, B, X, D) = P(S) \cdot P(C|S) \cdot P(B|S) \cdot P(X|C,S) \cdot P(D|C,B) \]

Conditional Independencies → Efficient Representation

<table>
<thead>
<tr>
<th>C</th>
<th>B</th>
<th>D=0</th>
<th>D=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>0.1</td>
</tr>
</tbody>
</table>