Task 6: Coping with Incomplete Knowledge: Overview

1. Approaches to incomplete knowledge
2. Modeling uncertainty with probabilities
3. Bayes Nets: representation and algorithms for dealing with probabilities
4. Utilities: from probabilities to Actions

Text: Chapters 13-16 of Russell & Norvig
Introductory Material: Sections 13.1 and 14.7

Randomness

- You flip a coin. It either comes up H or T.
  (truly random.)

- The Weather pattern. Will it rain tomorrow at 2pm?
  is it random ?
  or maybe we do not have a good enough theory ?

- The traffic jam at 4:45pm on South Bridge.

Aspects of Input

- What is the distance to the nearby wall?
  (measurement uncertainty)

- What is written here ?
  (input ambiguity)

- Understanding a speaker when there is background noise.
  (noise in measurements)

Information not Available

- Does the car across the street have 4 wheels?
  (default assumptions)
- A person arrives at the doctor's describing some symptoms.
  What is the diagnosis?
  (no complete theory, not enough evidence)
  What tests might help get a good diagnosis?
- You have just looked at the rear mirror of the car and now looking ahead intending to switch lanes.
  Is there a car behind in the other lane? (dynamics)

The Qualification Problem

- I need to get to class at 9am and have a plan to leave home half an hour early and drive to Forest Hill. Would like to conclude that the plan will get me there in time.
- But, the road may be blocked due to an accident,
- or a heavy fall of snow,
- or the road may be flooded due to an unexpected torrential rain,
- or my car may break down,
- or . . .

We don't want to (and sometimes cannot) list all possible qualifications.

Logical Approaches

- Formalise a theory of the world including actions and their results.
- Include a mechanism to derive logical conclusions.
- Include a mechanism to derive conclusions that do not follow logically but normally hold, e.g. "by default".

Default Logic

- Use first order logic as the base language and add "default rules" for inference:
- "if you see a car across the street, and it looks ok,
  and you can see at least 2 wheels
  and there is no other conflicting evidence
  then you can conclude that it has 4 wheels"
- "if x is the spouse of y, and x lives in town t,
  and there is no other conflicting evidence
  then you can conclude that y lives in t.
- These rules are not always true!

Default Logic

Pattern of inference is not monotonic. For example,

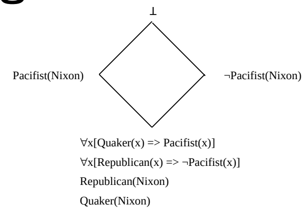
- Seeing a car . . . you conclude that it has 4 wheels, but as you cross the street you see that the car is on a jack
  ⇒ you *retract* your previous conclusion.
- You intend to phone a friend around 11:30am and thus plan to call them at work. But then your roommate says the friend called to say they caught a flu. So,
  ⇒ you revise your previous conclusion
  (on friend's location) and call them at home.

- Inference pattern of first order logic *on its own* is monotonic! Given more evidence we can get more conclusions and never retract conclusions.
- Default logic, and other logical approaches are known as *non-monotonic reasoning* systems.

Default Logic: Conflicting Conclusions

- May have a problem of conflicting defaults, for example:
- The “Nixon Diamond”
  (1) y is a Quaker $\Rightarrow$ conclude that y is a Pacifist
  (2) y is a Republican $\Rightarrow$ conclude that y is not a Pacifist in t
- Richard Nixon was a Republican and a Quaker

Resolving Inconsistent Defaults



- There are two ways to handle inconsistency.
  - Truth-Maintenance, or editing inconsistent statements out of the logic.
  - Probabilistic modeling

Assumption-Based Truth Maintenance Systems

- One comparatively efficient way of deciding which items to withdraw is to associate each proposition with the set of axioms or assumptions that it is consistent with.
- That way we can work out that a plausible way to restore consistency is to drop the assumption that Nixon is a Quaker.
- The sets of consistent propositions form a partial ordering or lattice like a version space.
- So we can exploit efficiency of the Version Space algorithm.

Default Logic: Many Conclusions?

- But what is the right conclusion . . .
  (1) if a person sneezes ⇒ conclude they have a cold
  (2) if a person allergic to cats sneezes ⇒ conclude there is a cat around
  Now we see a person sneezing. Should we conclude both possible outcomes? none? how do we choose?

Modeling Uncertainty with Probabilities

1. Model "causal" information:
   (1) if a person has a cold ⇒ they sneeze with probability 75%
   (2) if a person has an allergy to cats and there is a cat around ⇒ they sneeze with probability 90%
   (3) allergy and colds are otherwise independent
2. Now given an observation (sneeze) compute the probability of the person having a cold or of a cat in the vicinity.
3. How can we do this?

Summary

- We need to deal with uncertainty in many forms
- Logical approaches are possible;
  appeal to "jumping to conclusions" if no counter evidence exist.
- Can use probabilities to model uncertainty.
- This is the main topic of the module.

Probabilities and Bayesian Inference

1. Basic Properties of Probability
2. Bayes' Rule and Inference
3. Product Probability Spaces

Text: Sections 13.2 to 13.6 of Russell & Norvig

Probabilities

- Can be used to model "objectively" the situation in the world.
- A person with a cold sneezes (say at least once a day) with probability 75%.
- The objective interpretation is that this is a truly random event. Every person with a cold may or may not sneeze and does so with probability 75%.
- Or . . .

- Probabilities can model our "subjective" belief. For example: if we know that a person has a cold then we believe they will sneeze with probability 75%.
  The probabilities relate to an agent's state of knowledge.
  They change with new evidence.

We will use the subjective interpretation, though probability theory itself is independent of this choice.

Probabilities: Some Terminology

- **Elementary Event:** Outcome of some experiment e.g. coin comes out Head (H) when tossed
- **Sample Space:** SET of possible elementary events e.g. $\{H, T\}$. If sample space, S, is **finite**, we can denote the number of elementary events in S by $n(S)$.
- Example: For one throw of a die, the sample space is given by:

  $$S = \{1, 2, 3, 4, 5, 6\}$$

  and so

  $$n(S) = 6$$

- **Event:** *subset* of sample space i.e. if event is denoted by E then

  $$E \subseteq S$$

  and also

  $$n(E) \leq n(S)$$

- Example: Let
  E_o be event "the number is odd" when a die is rolled then $E_o = \{1, 3, 5\}$ and $n(E_o) = 3$

  E_e be event "the number is less than 3"when a die is rolled then $E_e = \{1, 2\}$ and $n(E_e) = 2$

Probabilities: Running Example

- We throw 2 dice (each with 6 sides and uniform construction).
- Each *elementary event* is a pair of number (a, b).
- The *sample space* i.e. set of elementary events is

$$\{(1,1),(1,2),\ldots,(6,6)\}$$

- *Events* i.e. subsets of the sample space include:
  E_1: outcomes where the first die has value 1.
  E_2: outcomes where the sum of the two numbers is even.
  E_3: outcomes where the sum of the two numbers is at least 11.

- Events A and B are *Mutually Exclusive* if $A \cap B = \emptyset$ (the empty set).
- **All** pairs of elementary events are mutually exclusive.
- In Example:
  Events E_1 and E_3 are mutually exclusive
  Events E_1 and E_2 are not
  Events E_2 and E_3 are not

Classical Definition of Probability

If the sample space S consists of a finite (but non-zero) number of equally likely outcomes, then the probability of an event E, written $\mathtt{Pr}\{E\}$ is defined as

$$\mathtt{Pr}\{E\} = \frac{n(E)}{n(S)}$$

This definition satisfies the **axioms of probability**:

1. $\mathtt{Pr}\{E\} \geq 0$ for any event E since $n(E) \geq 0$ and $n(S) > 0$
2. $\mathtt{Pr}\{S\} = \frac{n(S)}{n(S)} = 1$
3. If A and B are mutually exclusive: $\mathtt{Pr}\{A \cup B\} = \mathtt{Pr}\{A\} + \mathtt{Pr}\{B\}$

This is because the intersection of A and B is empty i.e. $A \cap B = \emptyset$. So, $n(A \cup B) = n(A) + n(B)$, and therefore

$$\begin{aligned}\mathtt{Pr}\{A \cup B\} &= \frac{n(A \cup B)}{n(S)} \\ &= \frac{n(A) + n(B)}{n(S)} \\ &= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} \\ &= \mathtt{Pr}\{A\} + \mathtt{Pr}\{B\}\end{aligned}$$

Note: $\mathtt{Pr}\{A \cap B\}$ is also denoted by $\mathtt{Pr}\{A, B\}$ and $\mathtt{Pr}\{A \text{ and } B\}$

Probability Distributions

- It is sufficient to specify the probability of elementary events.
- Other probabilities can be computed from these.
- In example: Each die gives a result 1,2,3,4,5,6 with probability $\frac{1}{6}$ *independently* of the other die. So each elementary event has probability $\frac{1}{36}$
- $\mathtt{Pr}\{E_1\} = \mathtt{Pr}\{(1,1)\} + \ldots \mathtt{Pr}\{(1,6)\} = 6/36 = 1/6$
- $\mathtt{Pr}\{E_2\} = \mathtt{Pr}\{(1,1)\} + \mathtt{Pr}\{(1,3)\} + \mathtt{Pr}\{(1,5)\} +$
  $\mathtt{Pr}\{(2,2)\} + \mathtt{Pr}\{(2,4)\} + \mathtt{Pr}\{(2,6)\} + \ldots = 1/2$
- $\mathtt{Pr}\{E_3\} = \mathtt{Pr}\{(5,6)\} + \mathtt{Pr}\{(6,5)\} + \mathtt{Pr}\{(6,6)\} = 3/36 = 1/12$

Properties of Probabilities

- We know that $\mathtt{Pr}\{A\} = \frac{n(A)}{n(S)}$
- Since event A is a subset of sample space S,

$$0 \leq n(A) \leq n(S) \quad \text{i.e.} \quad 0 \leq \frac{n(A)}{n(S)} \leq 1$$

- So, we have $0 \leq \mathtt{Pr}\{A\} \leq 1$
- If $\mathtt{Pr}\{A\} = 0$ then event cannot occur e.g. mutually exclusive events A and B, since then $\mathtt{Pr}\{A \cap B\} = \mathtt{Pr}\{\emptyset\} = 0$
- If $\mathtt{Pr}\{A\} = 1$ then event is certain to occur

Properties of Probabilities

- Let $\overline{A}$ denote the event "A does not occur"

$$\begin{aligned}\mathtt{Pr}\{\overline{A}\} &= \frac{n(\overline{A})}{n(S)} \\ &= \frac{n(S) - n(A)}{n(S)} \\ &= 1 - \frac{n(A)}{n(S)} = 1 - \mathtt{Pr}\{A\}\end{aligned}$$

- So, $\mathtt{Pr}\{\overline{A}\} = 1 - \mathtt{Pr}\{A\}$

Conditional Probability

- If A and B are two events and $\mathtt{Pr}\{B\} \neq 0$, then the **probability of A given B has already occurred** is written $\mathtt{Pr}\{A|B\}$ and defined as

$$\mathtt{Pr}\{A|B\} = \frac{\mathtt{Pr}\{A \cap B\}}{\mathtt{Pr}\{B\}}$$

- Where does this formula come from?
  Since we know that event B has already occurred, we know that

the sample space for **event** A **given event** B is B.

$$\begin{aligned}\texttt{Pr}\{A|B\} &= \frac{n(A\cap B)}{n(B)} \\ &= \frac{n(A\cap B)/n(S)}{n(B)/n(S)} \\ &= \frac{\texttt{Pr}\{A\cap B\}}{\texttt{Pr}\{B\}}\end{aligned}$$

Conditional Probability: Example

- $\texttt{Pr}\{E_1 \cap E_2\} = \texttt{Pr}\{(1,1)\} + \texttt{Pr}\{(1,3)\} + \texttt{Pr}\{(1,5)\} = 1/12$
  $\texttt{Pr}\{E_2 \cap E_3\} = \texttt{Pr}\{(6,6)\} = 1/36$
  $\texttt{Pr}\{E_1 \cap E_3\} = \texttt{Pr}\{\phi\} = 0$

- $\texttt{Pr}\{E_1|E_2\} = \frac{\texttt{Pr}\{E_1\cap E_2\}}{\texttt{Pr}\{E_2\}} = \frac{1/12}{1/2} = 1/6$
  $\texttt{Pr}\{E_2|E_3\} = \frac{\texttt{Pr}\{E_2\cap E_3\}}{\texttt{Pr}\{E_3\}} = \frac{1/36}{1/12} = 1/3$
  $\texttt{Pr}\{E_1|E_3\} = \frac{\texttt{Pr}\{E_1\cap E_3\}}{\texttt{Pr}\{E_3\}} = \frac{0}{1/12} = 0$

More on Conditional Probability

- Conditional probability rule is often written and used in the following form:

$$\texttt{Pr}\{A\cap B\} = \texttt{Pr}\{A|B\}\texttt{Pr}\{B\}$$

Since $\texttt{Pr}\{A\cap B\} = \texttt{Pr}\{B\cap A\}$, we also have:

$$\texttt{Pr}\{A\cap B\} = \texttt{Pr}\{B|A\}\texttt{Pr}\{A\}$$

and hence that $\texttt{Pr}\{A|B\}\texttt{Pr}\{B\} = \texttt{Pr}\{B|A\}\texttt{Pr}\{A\}$

- if A and B are **mutually exclusive** events then, as $\texttt{Pr}\{A\cap B\} = 0$ and $\texttt{Pr}\{B\} \neq 0$ (from def. of conditional probability), it follows that $\texttt{Pr}\{A|B\} = 0$.

- Next, we look at an important result: **Bayes' Theorem**

Bayes' Theorem

- We know that $\texttt{Pr}\{A \cap B\} = \texttt{Pr}\{A|B\}\texttt{Pr}\{B\} = \texttt{Pr}\{B|A\}\texttt{Pr}\{A\}$
- Reorganizing we get:

$$\texttt{Pr}\{A|B\} = \frac{\texttt{Pr}\{A\}\texttt{Pr}\{B|A\}}{\texttt{Pr}\{B\}}$$

- Dice Example: $\texttt{Pr}\{E_2|E_1\} = \frac{\texttt{Pr}\{E_2\}\texttt{Pr}\{E_1|E_2\}}{\texttt{Pr}\{E_1\}} = \frac{(1/2)(1/6)}{1/6} = 1/2$
- This will be the basis of our Bayesian inference and learning procedures!

Independent Events

If the occurrence or non-occurrence of an event A does *not* influence in any way the probability of an event B, then the event B is **(statistically) independent** of event A, and we have:

$$\texttt{Pr}\{A|B\} = \texttt{Pr}\{A\}$$

Since

$$\begin{aligned} & \texttt{Pr}\{A|B\}\texttt{Pr}\{B\} = \texttt{Pr}\{B|A\}\texttt{Pr}\{A\} \\ \Rightarrow\ & \texttt{Pr}\{A\}\texttt{Pr}\{B\} = \texttt{Pr}\{B|A\}\texttt{Pr}\{A\} \\ \Rightarrow\ & \texttt{Pr}\{B|A\} = \texttt{Pr}\{B\} \text{ also holds} \end{aligned}$$

But, we also know that

$$\texttt{Pr}\{A \cap B\} = \texttt{Pr}\{A|B\}\texttt{Pr}\{B\}$$

So, A and B independent also means that

$$\texttt{Pr}\{A \cap B\} = \texttt{Pr}\{A\}\texttt{Pr}\{B\}$$

- In Dice Example:
  Events E_1 and E_2 are statistically independent
  Events E_2 and E_3 are not
  Events E_1 and E_3 are not
- Independence can be used to simplify computations!

Independence: Example

- Sometimes we know from the probabilistic experiment that certain events are "physically" *independent*. In this case they are also statistically independent.
- Event E_4: outcomes where the second die has value 4.
- Since each throw of the die is independent (the first throw does not change the probabilities of outcomes of the second throw), E_1 and E_4 are independent. They are also statistically independent.
- This implies: $\texttt{Pr}\{E_1 \cap E_4\} = \texttt{Pr}\{E_1\}\texttt{Pr}\{E_4\} = 1/36$
  Can be verified using equations as above.

Probability Distributions

- A **random variable** associates a value with each possible outcome of an experiment e.g. $X = H$ or $X = T$ for outcomes of tossing a coin.
- An elementary event is an assignment of values to **all** variables.
- A probability distribution describes the probability of every elementary event and can be represented using a **probability table**.
- A **Joint Probability Distribution** contains information about probabilities of **all** variables and the connections between them.

Joint Distribution: Example

- Two variables A and B each of which can take values $(0, 1)$ (i.e. boolean variables).
- Table has one column for each variable
- Table has $2 \times 2 = 4$ rows
-

A	B	$\texttt{Pr}\{\}$
0	0	v_{00}
0	1	v_{01}
1	0	v_{10}
1	1	v_{11}

  - $\texttt{Pr}\{A = 0, B = 0\} = v_{00}$
  - $\texttt{Pr}\{A = 1, B = 0\} = v_{10}$
  - Particular probabilities denoted by $\texttt{Pr}\{\}$
  - The table is denoted by $\mathbf{Pr}(A, B)$

Joint Probability Distribution: Another Example

- We have 4 variables describing the situation about a person with an allergy to cats:
  Cold: person has a cold
  Cat: there is a cat in the vicinity of the person
  Allergy: the persons is showing an allergic reaction
  Sneeze: person sneezed
- Each of these is Boolean: taking values 0 or 1.
- The joint distribution can be described in a table of size $2^4 = 16$.

The Joint Distribution

Cold	*Cat*	*Allergy*	*Sneeze*	$\texttt{Pr}\{\}$
0	0	0	0	0.84645
0	0	0	1	0.0
0	0	1	0	0.004455
0	0	1	1	0.040095
0	1	0	0	0.0018
0	1	0	1	0.0
0	1	1	0	0.00072
0	1	1	1	0.00648
1	0	0	0	0.009405
1	0	0	1	0.084645
1	0	1	0	0.000225
1	0	1	1	0.004725
1	1	0	0	0.00002
1	1	0	1	0.00018
1	1	1	0	0.00004
1	1	1	1	0.00076

Marginal Distribution

- This is an induced probability distribution for one variable or more variables obtained from a joint distribution.
- To compute the marginal distribution of one or more variables, we **ignore** the other variables.

Example:

A	B	$\texttt{Pr}\{\}$
0	0	v_{00}
0	1	v_{01}
1	0	v_{10}
1	1	v_{11}

$\mathbf{Pr}(A, B)$

A	$\texttt{Pr}\{\}$
0	$v_{00} + v_{01}$
1	$v_{10} + v_{11}$

$\mathbf{Pr}(A)$

B	$\texttt{Pr}\{\}$
0	$v_{00} + v_{10}$
1	$v_{01} + v_{11}$

$\mathbf{Pr}(B)$

- How do we "ignore variables"?
- Example: We ignore B when computing marginal distribution for A by **summing it out**. This is described generically by:

$$\mathbf{Pr}(A) = \sum_{v \in \{0,1\}} \texttt{Pr}\{A, B = v\}$$

- Similarly, for marginal distribution of B, we have:

$$\mathbf{Pr}(B) = \sum_{v \in \{0,1\}} \texttt{Pr}\{A = v, B\}$$

- And for particular values:

$$\texttt{Pr}\{A = 0\} = \sum_{v=0}^{1} \texttt{Pr}\{A = 0 \text{ and } B = v\}$$

Dice Example:

$$\texttt{Pr}\{X_1 = 1\} = \sum_{v=1}^{6} \texttt{Pr}\{X_1 = 1 \text{ and } X_2 = v\} = 1/6$$

Allergy Example:

$$\begin{aligned} \texttt{Pr}\{Cold = 1\} &= \Sigma \ldots = 0.1 \\ \texttt{Pr}\{Cold = 0\} &= \Sigma \ldots = 0.9 \\ \texttt{Pr}\{Sneeze = 1\} &= \Sigma \ldots = 0.136885 \end{aligned}$$

Marginal Distribution: Another Example

C	A	B	$\texttt{Pr}\{\}$
0	0	0	v_{000}
0	0	1	v_{001}
0	1	0	v_{010}
0	1	1	v_{011}
1	0	0	v_{100}
1	0	1	v_{101}
1	1	0	v_{110}
1	1	1	v_{111}

$\mathbf{Pr}(A, B|C)$

Sum out B:

C	A	$\texttt{Pr}\{\}$
0	0	$v_{000} + v_{001}$
0	1	$v_{010} + v_{011}$
1	0	$v_{100} + v_{101}$
1	1	$v_{110} + v_{111}$

$\mathbf{Pr}(A|C) = \sum_v \texttt{Pr}\{A, B = v|C\}$

Inference using the Joint

- Compute probabilities of events:
$\texttt{Pr}\{Cold = 1 \text{ and } Sneeze = 1\} = \sum \ldots = 0.0902875$
- Causal Inference:
$\texttt{Pr}\{Sneeze = 1 | Cold = 1\} = \frac{0.0902875}{0.1} = 0.902875$
- Diagnostic Inference:
$\texttt{Pr}\{Cold = 1 | Sneeze = 1\} = \frac{0.0902875}{0.136885} = 0.66$
- Inter-Causal Inference:
$\texttt{Pr}\{Cold = 1 | Sneeze = 1 \text{ and } Allergy = 1\}$

Normalization

- $\texttt{Pr}\{Cold = 1 | Sneeze = 1\} = \frac{\texttt{Pr}\{Sneeze=1|Cold=1\}\texttt{Pr}\{Cold=1\}}{\texttt{Pr}\{Sneeze=1\}} = \frac{A}{\alpha}$
- $\texttt{Pr}\{Cold = 0 | Sneeze = 1\} = \frac{\texttt{Pr}\{Sneeze=1|Cold=0\}\texttt{Pr}\{Cold=0\}}{\texttt{Pr}\{Sneeze=1\}} = \frac{B}{\alpha}$
- But $\frac{A}{\alpha} + \frac{B}{\alpha} = 1$ so $\alpha = A + B$ and
- $\texttt{Pr}\{Cold = 1 | Sneeze = 1\} = \frac{A}{A+B}$
- Will often use normalization to avoid computing the denominator.

Conditional Independence

- Events A and B are statistically independent given event C iff

$$\texttt{Pr}\{A \cap B | C\} = \texttt{Pr}\{A|C\}\texttt{Pr}\{B|C\}$$

- This is equivalent to the condition $\underline{\texttt{Pr}\{A|B \text{ and } C\} = \texttt{Pr}\{A|C\}}$
- As in standard independence this can simplify the computations.
- Event E_5: outcomes where at least one die has value 6.
- $\texttt{Pr}\{E_2|E_3\} = 1/3$
- $\texttt{Pr}\{E_2|E_3 \text{ and } E_5\} = 1/3$

Conditional Independence: Derivation

$$\begin{aligned}
\texttt{Pr}\{A \cap B|C\} &= \frac{\texttt{Pr}\{A \cap B \cap C\}}{\texttt{Pr}\{C\}} \\
&= \frac{\texttt{Pr}\{A|B \cap C\}\texttt{Pr}\{B \cap C\}}{\texttt{Pr}\{C\}} \\
&= \frac{\texttt{Pr}\{A|B \cap C\}\texttt{Pr}\{B|C\}\texttt{Pr}\{C\}}{\texttt{Pr}\{C\}} \\
\text{So, } \texttt{Pr}\{A|C\}\texttt{Pr}\{B|C\} &= \texttt{Pr}\{A|B \cap C\}\texttt{Pr}\{B|C\} \\
\Rightarrow \texttt{Pr}\{A|B \cap C\} &= \texttt{Pr}\{A|C\}
\end{aligned}$$

Independence in Joint Probability Distributions

- In joint probability distributions we can express a more general form of independence.
- X_1 and X_2 are independent iff $\mathbf{Pr}(X_1|X_2) = \mathbf{Pr}(X_1)$
- This means that for all v_1 and v_2
  $\texttt{Pr}\{X_1 = v_1|X_2 = v_2\} = \texttt{Pr}\{X_1 = v_1\}$
- And similarly for conditional independence
  X_1 and X_2 are independent given X_3 iff
  $\mathbf{Pr}(X_1|X_2, X_3) = \mathbf{Pr}(X_1|X_3)$
- This means that for all v_1, v_2, v_3

$$\texttt{Pr}\{X_1 = v_1|X_2 = v_2, X_3 = v_3\} = \texttt{Pr}\{X_1 = v_1|X_3 = v_3\}$$

Bayesian Networks

1. Representing Distributions with Bayesian Networks
2. How to Construct the Network
3. Inference using Networks

Text: Sections 14.1 to 14.5 of Russell & Norvig

Product Probability Spaces

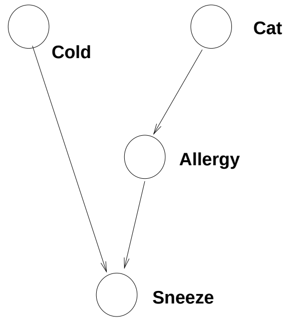
- We have n variables, $X_1, \ldots, X_n$
- Each variable X_i ranges over a finite set of values $v_{i,1}, \ldots, v_{i,k_i}$.
- We can write a big table with n columns and $\prod_i k_i$ rows describing the probability of every elementary event.
- Table grows exponentially with n.
  Not feasible unless n is very small.

Bayesian Networks

- Allow us to represent distributions more compactly.
- Take advantage of the structure available in a domain.
- Basic idea: represent dependence and independence *explicitly*.
- If $\mathbf{Pr}(X_1 \cap X_2) = \mathbf{Pr}(X_1)\mathbf{Pr}(X_2)$ then we can use two 1-dimensional tables instead of a 2-dimensional table.
- If each has 6 values, this means 12 entries instead of 36!

Example

- Edges represent "direct influence"
- Assume that *Cat* and *Cold* do not depend on other variables.
- *Allergy* depends only on *Cat*
- *Sneeze* depends on *Cold* and *Allergy*
- *Sneeze* depends on *Cat* BUT only through *Allergy*
- For each node we associate a *conditional probability table*



The network structure expresses *independence* of variables.

- $\mathbf{Pr}(Cat|Cold) = \mathbf{Pr}(Cat)$
- $\mathbf{Pr}(Allergy|Cat, Cold) = \mathbf{Pr}(Allergy|Cat)$
- $\mathbf{Pr}(Sneeze|Allergy, Cat, Cold) = \mathbf{Pr}(Sneeze|Allergy, Cold)$

More generally, the joint distribution can be expressed as the product of the distributions in the network. Example:

$$\mathbf{Pr}(Cat, Cold, Allergy, Sneeze) = \\ \mathbf{Pr}(Cat)\mathbf{Pr}(Cold)\mathbf{Pr}(Allergy|Cat)\mathbf{Pr}(Sneeze|Allergy, Cold)$$

$\mathtt{Pr}\{Cat = 0\}$	$\mathtt{Pr}\{Cat = 1\}$
0.99	0.01

$\mathtt{Pr}\{Cold = 0\}$	$\mathtt{Pr}\{Cold = 1\}$
0.90	0.10

Cat	$\mathtt{Pr}\{Allergy = 0\}$	$\mathtt{Pr}\{Allergy = 1\}$
0	0.95	0.05
1	0.20	0.80

$Cold$	$Allergy$	$\mathtt{Pr}\{Sneeze = 0\}$	$\mathtt{Pr}\{Sneeze = 1\}$
0	0	1.00	0.00
0	1	0.10	0.90
1	0	0.10	0.90
1	1	0.05	0.95

Example: Joint Distribution in Terms of Conditional Probability Distributions

We can use the conditional probability rule (in its various forms) and Bayes' Theorem to express **Joint Distributions** in terms of conditional probability distributions (stored with nodes in Bayesian network as Conditional Probability Tables (CPTs)).

Example: Express $\mathbf{Pr}(Cat, Cold, Allergy, Sneeze)$ in terms of **known** conditional probability distributions (tables):

$$\mathbf{Pr}(Cat), \mathbf{Pr}(Cold), \mathbf{Pr}(Allergy|Cat), \mathbf{Pr}(Sneeze|Allergy, Cold)$$

$\mathbf{Pr}(Cat, Cold, Allergy, Sneeze)$

$= \mathbf{Pr}(Sneeze|Allergy, Cat, Cold)\mathbf{Pr}(Allergy, Cat, Cold)$

by conditional probability rule applied to distributions

$= \mathbf{Pr}(Sneeze|Allergy, Cold)\mathbf{Pr}(Allergy|Cat, Cold)\mathbf{Pr}(Cat, Cold)$

by conditional probability rule applied to distributions

$= \mathbf{Pr}(Sneeze|Allergy, Cold)\mathbf{Pr}(Allergy|Cat)\mathbf{Pr}(Cat|Cold)\mathbf{Pr}(Cold)$

by conditional probability rule applied to distributions

$= \mathbf{Pr}(Sneeze|Allergy, Cold)\mathbf{Pr}(Allergy|Cat)\mathbf{Pr}(Cat)\mathbf{Pr}(Cold)$

Another Useful Result

$$\underline{\mathbf{Pr}(A, B|C) = \mathbf{Pr}(A|B, C)\mathbf{Pr}(B|C)} \quad \dagger$$

Proof:

$$\begin{aligned}
\text{RHS} = \mathbf{Pr}(A|B, C)\mathbf{Pr}(B|C) &= \frac{\mathbf{Pr}(A, B, C)}{\mathbf{Pr}(B, C)} \cdot \frac{\mathbf{Pr}(B, C)}{\mathbf{Pr}(C)} \\
&= \frac{\mathbf{Pr}(A, B, C)}{\mathbf{Pr}(C)} \\
&= \frac{\mathbf{Pr}(A, B|C)\mathbf{Pr}(C)}{\mathbf{Pr}(C)} \\
&= \mathbf{Pr}(A, B|C) = \text{LHS}
\end{aligned}$$

Also: $\mathtt{Pr}\{A = a, B = b|C = c\} = \mathtt{Pr}\{A = a|B = b, C = c\}\mathtt{Pr}\{B = b|C = c\}$

Bayesian Networks

- We have n variables, $X_1, \ldots, X_n$
- The graph has n nodes, each corresponding to one variable.
- The graph is *acyclic*; there is no directed path from X_i to itself.
- For each node associate a conditional probability table (CPT) describing $\mathbf{Pr}(X_i|$ parents$(X_i))$.
- The joint probability distribution can be expressed as the product of the distributions in the network:
  $\mathbf{Pr}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{Pr}(X_i|$ parents$(X_i))$

- $\texttt{Pr}\{Cat{=}0, Cold{=}1, Allergy{=}0, Sneeze{=}1\} =$
  $\texttt{Pr}\{Cat{=}0\}\texttt{Pr}\{Cold{=}1\}\texttt{Pr}\{Allergy{=}0|Cat{=}0\}$
  $\texttt{Pr}\{Sneeze{=}1|Allergy{=}0, Cold{=}1\} =$
  $0.99 \cdot 0.1 \cdot 0.95 \cdot 0.9 = 0.084645$
- So to represent a distribution we need to represent the network and CPTs.
- If for all nodes the number of parents is small then all CPTs are small and we have a compact representation.

How to Construct a Network?

- Can represent any distribution using a network.
  By repeated application of $\texttt{Pr}\{A, B\} = \texttt{Pr}\{A|B\}\texttt{Pr}\{B\}$
- Choose ordering of variables $X_1, \ldots, X_n$.
- For $i{=}1$ to n
  Add X_i to network with $\mathbf{Pr}(X_i|X_1, \ldots, X_{i-1})$.
- The joint distribution is:
  $\mathbf{Pr}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{Pr}(X_i|X_1, \ldots, X_{i-1})$.
- But this is no improvement as X_n is connected to all predecessors (so we need a huge table for it).

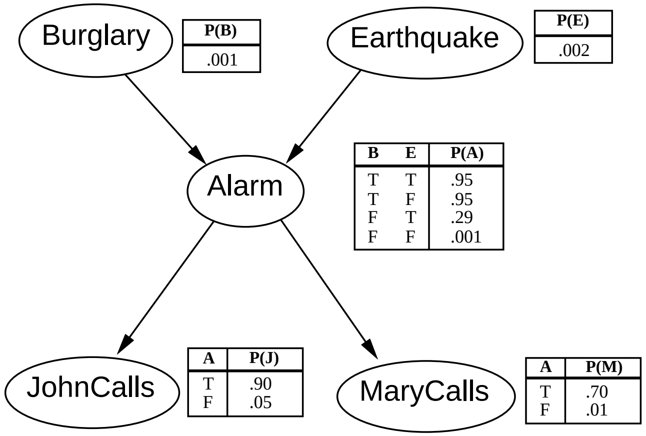
- Instead, at each stage choose a subset such that
  parents$(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$.
- Choice of parents must satisfy
  $\mathbf{Pr}(X_i|X_1, \ldots, X_{i-1}) = \mathbf{Pr}(X_i|$ parents$(X_i))$
  so that X_i is independent of other predecessors given its parents.
- If parents(X_i) is small then representation is compact.
  For example if for all X_i, $|$ parents$(X_i)| \leq 3$ then
  instead of 2^n entries we have $n2^3 = 8n$ entries !
- How should we order the variables?
- Causal links tend to produce small representations.
- Choose "root causes" first. Then continue with causal structure as much as possible.

Example

You are at work, neighbour John calls to say your home alarm is ringing, but neighbour Mary doesn't call. Sometimes alarm set off by minor earthquakes. Is there a burglar?

Variables and ordering:
$Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:

The Network



P(B)
.001

P(E)
.002

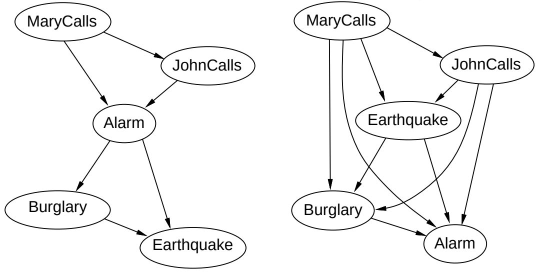
B	E	P(A)
T	T	.95
T	F	.95
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

What if we choose another ordering?

Variable ordering (1):
$MaryCalls$, $JohnCalls$, $Alarm$, $Burglar$, $Earthquake$
Variable ordering (2):
$MaryCalls$, $JohnCalls$, $Earthquake$ $Burglar$, $Alarm$,



How to Construct a Network

- May work with domain experts to decide on structure and then find values of probabilities.
- Choosing a "bad order" can have adverse effect:
  Large probability tables.
  "Unnatural" dependencies, that are in turn hard to estimate.
- Luckily experts are often good at identifying causal structure, and prefer giving probability judgements for causal rules.
- *Learning* the CPTs and even the structure is an active research area.

Computing with Bayes Nets

- We have seen how to reconstruct the joint distribution from the network.

- Can we compute other probabilities efficiently?
  $\mathtt{Pr}\{Cold = 1 \text{ and } Sneeze = 1\} = ?$
  $\mathtt{Pr}\{Sneeze = 1 | Cold = 1\} = ?$
  $\mathtt{Pr}\{Cold = 1 | Sneeze = 1\} = ?$
  $\mathtt{Pr}\{Cold = 1 | Sneeze = 1 \text{ and } Allergy = 1\} = ?$

Computing a Marginal Distribution

We want to compute $\mathbf{Pr}(Cold, Sneeze)$.
By conditional probability rule:

$$\mathbf{Pr}(Cold, Sneeze) = \mathbf{Pr}(Sneeze|Cold)\mathbf{Pr}(Cold)$$

Recall that we have the following CPTs attached to nodes of our network:

$$\mathbf{Pr}(Cat) \qquad \mathbf{Pr}(Cold)$$

$$\mathbf{Pr}(Allergy|Cat) \qquad \mathbf{Pr}(Sneeze|Allergy, Cold)$$

We need to compute $\mathbf{Pr}(Sneeze|Cold)$:

$$\mathbf{Pr}(Sneeze|Cold)$$

$$= \sum_{v\in\{0,1\}} \mathtt{Pr}\{Sneeze, Allergy = v|Cold\}$$

by def. of marginal distribution (we sum Allergy out)

$$= \sum_{v\in\{0,1\}} \mathtt{Pr}\{Sneeze|Allergy = v, Cold\}\mathtt{Pr}\{Allergy = v|Cold\}$$

by result (†)

$$= \sum_{v\in\{0,1\}} \mathtt{Pr}\{Sneeze|Allergy = v, Cold\}\mathtt{Pr}\{Allergy = v\}$$

Since Allergy is independent of Cold

We can read out $\mathtt{Pr}\{Sneeze|Allergy = v, Cold\}$ for all v using our CPT for $\mathbf{Pr}(Sneeze|Allergy, Cold)$.
We now need to compute the marginal distribution $\mathbf{Pr}(Allergy)$.

$$\mathtt{Pr}\{Allergy = v\}$$

$$= \sum_{v_2\in\{0,1\}} \mathtt{Pr}\{Allergy = v, Cat = v_2\}$$

by def. of marginal distribution (we sum Cat out)

$$\sum_{v_2\in\{0,1\}} \mathtt{Pr}\{Allergy = v|Cat = v_2\}\mathtt{Pr}\{Cat = v_2\}$$

by def. of conditional probability rule

So, we have now expressed **all** computations in terms of the CPTs in the network. We can write the computation fully as:

$$\mathbf{Pr}(Cold, Sneeze)$$

$$= (\sum_{v\in\{0,1\}} (\mathtt{Pr}\{Sneeze|Allergy = v, Cold\}$$

$$\sum_{v_2\in\{0,1\}} \mathtt{Pr}\{Allergy = v|Cat = v_2\}\mathtt{Pr}\{Cat = v_2\}))$$

$$\mathbf{Pr}(Cold)$$

In general, "sum out" variables that do not appear in the question. Try to maintain small tables along the way.

$\mathtt{Pr}\{Allergy = 0\}$	$\mathtt{Pr}\{Allergy = 1\}$
0.9425	0.0575

$\mathbf{Pr}(Allergy)$

Cold	$\mathtt{Pr}\{Sneeze = 0\}$	$\mathtt{Pr}\{Sneeze = 1\}$
0	0.94825	0.05175
1	0.097125	0.902875

$\mathbf{Pr}(Sneeze|Cold)$

Cold	*Sneeze*	$\mathtt{Pr}\{\}$
0	0	0.853425
0	1	0.046575
1	0	0.0097125
1	1	0.0902875

$\mathbf{Pr}(Sneeze$ and $Cold)$

Causal Inference

- To compute $\mathtt{Pr}\{Sneeze = 1|Cold = 1\}$
- First compute $\mathbf{Pr}(Allergy)$ as in previous example.
- Then compute
  $\mathtt{Pr}\{Sneeze = 1|Cold = 1\} =$
  $\sum_{v\in\{0,1\}} \mathtt{Pr}\{Sneeze{=}1|Allergy{=}v, Cold{=}1\}\mathtt{Pr}\{Allergy{=}v\}$

Diagnostic Inference

- Use Bayes' Rule to compute $\mathtt{Pr}\{Cold = 1|Sneeze = 1\}$
- $A_1 = \mathtt{Pr}\{Cold{=}1|Sneeze{=}1\} = \frac{\mathtt{Pr}\{Sneeze=1|Cold=1\}\mathtt{Pr}\{Cold=1\}}{\mathtt{Pr}\{Sneeze=1\}} = \frac{N_1}{\mathtt{Pr}\{Sneeze=1\}}$
- $A_0 = \mathtt{Pr}\{Cold{=}0|Sneeze{=}1\} = \frac{\mathtt{Pr}\{Sneeze=1|Cold=0\}\mathtt{Pr}\{Cold=0\}}{\mathtt{Pr}\{Sneeze=1\}} = \frac{N_0}{\mathtt{Pr}\{Sneeze=1\}}$
- But $A_0 + A_1 = 1$ so
- $(\mathtt{Pr}\{Cold{=}0|Sneeze{=}1\}, \mathtt{Pr}\{Cold{=}1|Sneeze{=}1\}) = Normalise(N_0, N_1) = (\frac{N_0}{N_0+N_1}, \frac{N_1}{N_0+N_1})$

- From the CPT we have:
  $\mathtt{Pr}\{Cold = 0\} = 0.90$ and $\mathtt{Pr}\{Cold = 1\} = 0.10$
- $\mathtt{Pr}\{Sneeze{=}1|Cold{=}0\} = 0.05175$
  $\mathtt{Pr}\{Sneeze{=}1|Cold{=}1\} = 0.902875$
  computed as in previous example using $\mathbf{Pr}(Allergy)$
- $N_0 = 0.05175 \cdot 0.90 = 0.046575$
- $N_1 = 0.902875 \cdot 0.10 = 0.0902875$
- $\mathtt{Pr}\{Cold{=}1|Sneeze{=}1\} = \frac{N_1}{N_0+N_1} = 0.66$

Inference in Burglary Example

- Use B, E, A, J, M to denote variables
- Compute $\mathtt{Pr}\{A{=}1|E{=}1, J{=}1\}$
- $\mathtt{Pr}\{A{=}1|E{=}1, J{=}1\} = \frac{\mathtt{Pr}\{J=1|A=1,E=1\}\mathtt{Pr}\{A=1|E=1\}}{\mathtt{Pr}\{J=1|E=1\}}$
- By computing and normalizing the following: $\mathtt{Pr}\{J{=}1|A{=}b, E{=}1\}\mathtt{Pr}\{A{=}b|E{=}1\}$ for $b \in \{0, 1\}$
- $\mathtt{Pr}\{A = 1|E{=}1\} = 0.001 \cdot 0.95 + 0.999 \cdot 0.29 = 0.29066$
- $\mathtt{Pr}\{A = 0|E{=}1\} = 0.70934$
- $\mathtt{Pr}\{J{=}1|A{=}1, E{=}1\}\mathtt{Pr}\{A{=}1|E{=}1\} = 0.90 \cdot 0.29066 = 0.261994$
- $\mathtt{Pr}\{J{=}1|A{=}0, E{=}1\}\mathtt{Pr}\{A{=}0|E{=}1\} = 0.05 \cdot 0.70934 = 0.035467$
- $\mathtt{Pr}\{A{=}1|E{=}1, J{=}1\} = \frac{0.261994}{0.261994+0.035467} = 0.88$

Inference in Burglary Example: Derivation

$$
\begin{aligned}
&\mathtt{Pr}\{A = 1|E = 1, J = 1\}\\
&= \frac{\mathtt{Pr}\{A = 1, E = 1, J = 1\}}{\mathtt{Pr}\{J = 1, E = 1\}}\\
&= \frac{\mathtt{Pr}\{J = 1|A = 1, E = 1\}\mathtt{Pr}\{A = 1, E = 1\}}{\mathtt{Pr}\{J = 1, E = 1\}}\\
&= \frac{\mathtt{Pr}\{J = 1|A = 1, E = 1\}\mathtt{Pr}\{A = 1|E = 1\}\mathtt{Pr}\{E = 1\}}{\mathtt{Pr}\{J = 1|E = 1\}\mathtt{Pr}\{E = 1\}}\\
&= \frac{\mathtt{Pr}\{J = 1|A = 1, E = 1\}\mathtt{Pr}\{A = 1|E = 1\}}{\mathtt{Pr}\{J = 1|E = 1\}}
\end{aligned}
$$

$$
\begin{aligned}
&\mathtt{Pr}\{A = 1|E = 1\}\\
&= \frac{\mathtt{Pr}\{A = 1, E = 1\}}{\mathtt{Pr}\{E = 1\}}
\end{aligned}
$$

by def. of marginal distribution:

$$
\begin{aligned}
&= \frac{\Sigma_{v\in\{0,1\}}\mathtt{Pr}\{A = 1, E = 1, B = v\}}{\mathtt{Pr}\{E = 1\}}\\
&= \frac{\Sigma_{v\in\{0,1\}}\mathtt{Pr}\{A = 1|E = 1, B = v\}\mathtt{Pr}\{E = 1, B = v\}}{\mathtt{Pr}\{E = 1\}}
\end{aligned}
$$

since E (earthquake) and B (burglary) are independent:

$$
= \frac{\Sigma_{v\in\{0,1\}}\mathtt{Pr}\{A = 1|E = 1, B = v\}\mathtt{Pr}\{E = 1\}\mathtt{Pr}\{B = v\}}{\mathtt{Pr}\{E = 1\}}
$$

$$= \Sigma_{v\in\{0,1\}}\mathtt{Pr}\{A = 1|E = 1, B = v\}\mathtt{Pr}\{B = v\}$$

by expanding summation:

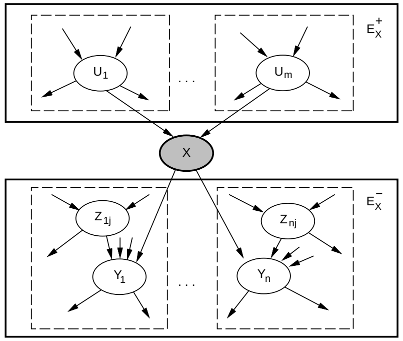
$$= \mathtt{Pr}\{A = 1|E = 1, B = 0\}\mathtt{Pr}\{B = 0\} + \mathtt{Pr}\{A = 1|E = 1, B = 1\}\mathtt{Pr}\{B = 1\}$$

by probabilities from CPTs in network

$$= (0.29)(0.999) + (0.95)(0.01) = 0.29066$$

Poly-Tree Networks

Poly-Tree is a singly connected network. There are no cycles even ignoring the direction of edges.



Inference

- The problem of inference in Bayes Networks is NP-Hard.
- Can always compute via the joint but this may not be efficient.
- Sneeze and Burglary examples were Poly Trees.
- Efficient Algorithms are known for graphs with poly-tree structure.
- Otherwise, try to turn graph into a tree, or use simulation methods to approximate the probability.
- Simulation is not guaranteed to give good answers but works well in some cases.
- Generates samples from the prior Joint Distribution.
- cf. Estimation of conditional probability.

Inference by Simulation

- To compute $\mathtt{Pr}\{X = v_1 | Y = v_2\} = \frac{\mathtt{Pr}\{X=v_1, Y=v_2\}}{\mathtt{Pr}\{Y=v_2\}}$
- Repeat many times:

```
    choose random values for all nodes using CPTs
        by choosing first for nodes with no parents
        and then for nodes whose parents have been chosen
    if $Y = v_2$ was chosen then:
        counter = counter +1
        if $X = v_1$ was chosen then Xcounter = Xcounter+1
Return Xcounter/Counter
```

- May require many rounds if $Y = v_2$ occurs with low probability.

Simulation by Likelihood Weighting

- Improves the simulation by forcing evidence values ($Y = v_2$) to be chosen.
- Samples only the non-evidence variables, weighting each event by the *likelihood* that the event accords to the evidence, and incrementing the two counters in proportion to the likelihood of the event.
- *Likelihood* of an event is the product of the conditional probabilities that each evidence variable takes the given value given the values that have been chosen randomly for its parents.
- Unlikely events contribute less to the counts.

Computing Likelihood Weighting of an Event

- Initialise weight $w = 1$
- In sampling if a node X_i corresponds to an evidence variable choose the given value v_i.
- Update the weight
  $w \leftarrow w \cdot \texttt{Pr}\{X_i = v_i|$ values chosen for parents$\}$
- This ensures that an event which assigns low *a priori* probability to the actual values of the evidence variables is given less weight in the simulation.

Example

- Compute $\texttt{Pr}\{Allergy = 1|Cold = 1, Sneeze = 1\}$
- Set up general counter C_G and counter C_A for $Allergy = 1$ (both initialised to zero).
- Here we illustrate just one sample
- Choose $Cold = 1$ so $w \leftarrow 0.1$
- Choose Cat randomly. In the following we assume 0 was chosen.
- Choose $Allergy$ using $\texttt{Pr}\{Allergy|Cat = 0\}$.
  In the following we assume 0 was chosen.
- Choose $Sneeze = 1$ and update
  $w \leftarrow w \cdot \texttt{Pr}\{Sneeze{=}1|Cold{=}1, Allergy{=}0\} = 0.1 \cdot 0.9 = 0.09$

- The round is over.
  We forced the evidence to succeed.
  But $Allergy = 1$ did not succeed.
- $C_G \leftarrow C_G + 1 * 0.09$
  $C_A \leftarrow C_A + 0 * 0.09$ (no change)
- After Many rounds return C_A/C_G
- Faster than simple simulation but may still need a long time to get good answers

Making Decisions

1. From Probabilities to Decisions
2. Utility as a Basis for Decisions
3. Dynamic Bayesian Networks

Text: Russell & Norvig,
Sections 13.1, 16.1 to 16.5-16.7, 15.5, 17.5

What is the Conclusion?

- Imagine we have a model with two nodes $D \rightarrow S$
- And the CPTs, $\mathbf{Pr}(D)$ and $\mathbf{Pr}(S|D)$.
- D captures possible diseases, and S possible symptoms.
- NB This assumes only one disease or symptom at a time or alternatively that D, S capture possible combinations.
- We observe a value $S = s$ and can compute $\mathbf{Pr}(D|S = s)$.
- But what should we do with these probabilities?

Maximum A Posteriori Hypothesis

- Given $\mathbf{Pr}(D|S = s)$
- Choose d that maximises the posterior probability.
- $d = \text{argmax}_{v_d} \mathbf{Pr}(D = v_d|S = s)$
- Use treatment for d.
- Is that reasonable?

MAP and Utilities

- Imagine a similar setting with $C \rightarrow A$
- A denotes our alarm being triggered (values 0,1) and C the possible causes: burglary (b), strong wind (w), system fault (f).
- $\mathbf{Pr}(A = 1|C) = (0.95, 0.7, 0.07)$ [ordered by | (b,w,f)]
- $\texttt{Pr}\{C\} = (0.05, 0.10, 0.01)$
- We observe $A = 1$. By Bayes Rule and $\texttt{Pr}\{A = 1\}$: $\texttt{Pr}\{C|A = 1\} = (0.403, 0.595, 0.00059)$
- So the MAP is $C = w$
- But should we really ignore the Alarm? It depends on how much we may loose (by ignoring or by being distracted)

Utilities

- Every state (elementary event), say R, has some utility associated with it, denoted $U(R)$
- This captures the desirability of this state.
- Using actions we may change probabilities of states.
- In last example we must include a node for consequences of our actions (catch thief, waste time).
- Even with these new nodes MAP will not help us incorporate the utilities.
- So more machinery is needed.

Maximum Expected Utility (MEU)

- The *Expected Utility* of an action Act given the evidence is $EU(Act|E) = \sum_{R \in \mathsf{Result}(Act)} \mathtt{Pr}\{R|E, Act\} U(R)$
- Maximum Expected Utility: a rational agent should choose action Act that maximises $EU(Act|E)$.
- Utility is not necessarily "personal benefit", "monetary situation of agent" etc. but may include "well being of world"
- In a sort of circular manner we may say that a rational agent's utility is implicit in the actions it takes.
- Are human rational according to this definition?
  See discussion in R&N

Example Revisited

- Add an action G (go home to check for burglary)
- And Binary variables T (time wasted) and S (things stolen)
- And probabilities for T, S
  $\mathtt{Pr}\{T = 1|G\} = 1$
  $\mathtt{Pr}\{T = 1|\neg G\} = 0$
  $\mathtt{Pr}\{S = 1|\neg G \text{ and } C = b\} = 1$
  $\mathtt{Pr}\{S = 1|G \text{ or } C \neq b\} = 0$
- And $U(C, A, T, S) = (-1) \cdot T + (-10) \cdot S$
- $EU(G|A = 1) = (-1) \cdot 1 + (-10) \cdot 0 = -1$
  $EU(\neg G|A = 1) = (-1) \cdot 0 + (-10) \cdot \mathtt{Pr}\{S{=}1|A{=}1, \neg G\} = -4.03$

Computing $EU(G|A = 1)$

- The *Expected Utility* of an action Act given the evidence is

$$EU(Act|E) = \sum_{R \in \mathsf{Result}(Act)} \mathtt{Pr}\{R|E, Act\} U(R)$$

- The utilities are for T and S are:

$$U(T = x) = (-1) \cdot x \text{ and } U(S = x) = (-10) \cdot x$$

- Compute $EU(G|A = 1)$ as follows:

$$\begin{aligned}
& EU(G|A = 1) \\
= & \sum_{R \in \mathsf{Result}(G)} \mathtt{Pr}\{R|A = 1, G\} U(R) \\
= & \mathtt{Pr}\{T = 1|A = 1, G\} \cdot U(T = 1) + \mathtt{Pr}\{T = 0|A = 1, G\} \cdot U(T = 0) \\
& + \mathtt{Pr}\{S = 1|A = 1, G\} \cdot U(S = 1) + \mathtt{Pr}\{S = 0|A = 1, G\} \cdot U(S = 0) \\
= & \mathtt{Pr}\{T = 1|G\} \cdot (-1) + (0) \cdot (0) + (0) \cdot (-11) + \mathtt{Pr}\{S = 0|A = 1, G\} \cdot (0) \\
& \text{—as time wasted is independent of alarm being triggered given G} \\
= & (1) \cdot (-1) = -1
\end{aligned}$$

Computing $EU(\neg G|A=1)$

$$
\begin{aligned}
&= \sum_{R \in \text{Result}(\neg G)} \texttt{Pr}\{R|A=1,\neg G\}U(R) \\
&= \texttt{Pr}\{T=1|A=1,\neg G\}\cdot U(T=1) + \texttt{Pr}\{T=0|A=1,\neg G\}\cdot U(T=0)) \\
&\quad +\texttt{Pr}\{S=1|A=1,\neg G\}\cdot U(S=1) + \texttt{Pr}\{S=0|A=1,\neg G\}\cdot U(S=0) \\
&= (0)\cdot(-1) + \texttt{Pr}\{T=0|A=1,\neg G\}\cdot(0) \\
&\quad +\texttt{Pr}\{S=1|A=1,\neg G\}\cdot(-10) + \texttt{Pr}\{S=0|A=1,\neg G\}\cdot(0) \\
&= \texttt{Pr}\{S=1|A=1,\neg G\}\cdot(-10) \\
&= (0.403)\cdot(-10) = -4.03
\end{aligned}
$$

Computing $\texttt{Pr}\{S=1|A=1,\neg G\}$

$$
\begin{aligned}
&\texttt{Pr}\{S=1|A=1,\neg G\} \\
&= \sum_v \texttt{Pr}\{S=1|A=1,\neg G, C=v\}\cdot \texttt{Pr}\{C=v|A=1,\neg G\} \\
&= \sum_v \texttt{Pr}\{S=1|\neg G, C=v\}\cdot \texttt{Pr}\{C=v|A=1\}
\end{aligned}
$$

Continues on next slide

$$
\begin{aligned}
&= \texttt{Pr}\{S=1|\neg G, C=b\}\cdot \texttt{Pr}\{C=b|A=1\} + \\
&\quad \texttt{Pr}\{S=1|\neg G, C=w\}\cdot \texttt{Pr}\{C=w|A=1\} + \\
&\quad \texttt{Pr}\{S=1|\neg G, C=f\}\cdot \texttt{Pr}\{C=f|A=1\} \\
&= (1)\cdot(0.403) + (0)\cdot(0.595) + (0)\cdot(0.00059)
\end{aligned}
$$

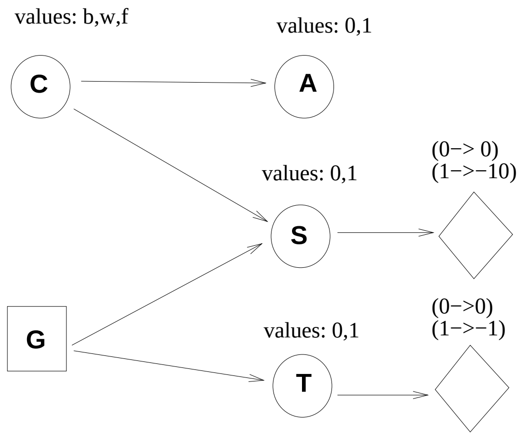
(probs. from slide 4-4)

$$= 0.403$$

Decision Networks

- Augment Bayesian networks to capture these ideas explicitly.
- **Chance Nodes** represent random variables as before
- **Decision Nodes** represent actions that can be chosen by agent.
- **Utility Nodes** represent utility as a function of variables.
  Can represent using tables likes CPTs (and sum over tables).
- For inference or evaluating what actions to take:
  (1) fix values for decision nodes
  (2) compute $\mathbf{Pr}$(unobserved|observed and decision nodes)
  (3) compute EU(Actions)

Example Revisited



Value of Information

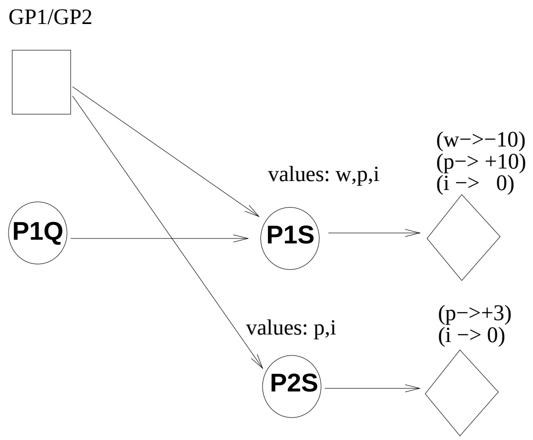
- The best action now yields: $EU(\alpha|E) = \max_{Act} \sum_{R \in \mathsf{Result}(Act)} \mathtt{Pr}\{R|E, Act\} U(R)$
- Imagine that at any state we can choose to perform a "sense" operation and find out the value of one of the variables X_i.
- Suppose we sense the value of X_i to be v.
- The best action now yields: $EU(\alpha_v|E, X_i = v) =$
  $\max_{Act} \sum_{R \in \mathsf{Result}(Act)} \mathtt{Pr}\{R|E, Act, X_i = v\} U(R)$
- But we don't know in advance which v is the case.

- The expected value of the sense operation is: $EVS(X_i|E) = \sum_v \mathtt{Pr}\{X_i = v|E\} EU(\alpha_v|E, X_i = v)$
- And subtracting the original utility we get the *Value of Perfect Information*: $VPI(X_i|E) = EVS(X_i|E) - EU(\alpha|E)$
- Sense operations may have a cost as well which must be weighted against the VPI.

Example

- You have access to two printers P1 a colour printer shared by everyone in the building and P2 a b/w private one.
- P1 may have a long queue in which case you have to wait. P2 prints immediately but with less good quality.
- 3 chance nodes $P1Q$ (binary; for its queue) and $P1S, P2S$ for status of printers (waiting (w), printed (p), idle(i))
- One decision node sending job to printer: $GP1$ or $GP2$.
- $U(P1Q, P1S, P2S) =$
  $(P1S{=}w) \cdot (-10) + (P1S{=}p) \cdot 10 + (P2S{=}p) \cdot 3$

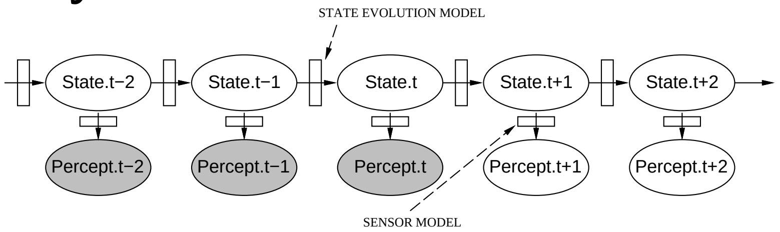
The Network



Example Continued

- $\texttt{Pr}\{P1Q = 1\} = 0.7$
- Other links are deterministic ($\texttt{Pr}\{\}$ is 0 or 1).
- The best action with no evidence is $GP2$:
  $EU(GP1) = 0.7 \cdot (-10) + 0.3 \cdot 10 = -4$
  $EU(GP2) = 1 \cdot 3 = 3$
- But things improve if we sense $P1Q$:
  (if $P1Q = 0$ the best action is $GP1$ otherwise it is $GP2$)
  $EVS(P1Q) =$
  $\texttt{Pr}\{P1Q{=}0\}EU(GP1|P1Q{=}0) + \texttt{Pr}\{P1Q{=}1\}EU(GP2|P1Q{=}1) =$
  $0.3 \cdot (0 + 10 + 0) + 0.7 \cdot (0 + 0 + 3) = 5.1$
- $VPI(P1Q) = 5.1 - 3 = 2.1$

Dynamic Belief Networks



- Divide variables into State and Percept (the observable part)
- State evolution model describes change with time
- Sensor model describes behaviour of observations
- Both are the same in every “slice”
- Network calculates probability distribution for state at time t

Dynamic Belief Networks and HMMs

- Every Hidden Markov Model (HMM) can be represented as a DBN with a single state and percept variable.
- Every DBN can be represented as an HMM by combining all the n state variables into a single variable with n-tuple values.
- The DBN is much more efficient than the equivalent HMM, because for sensible temporal probability models, each state variable has few parents in the preceding slice.
- The relation between DBNs and HMMs is roughly analogous to that between ordinary Bayes Nets and fully tabulated joint distributions.

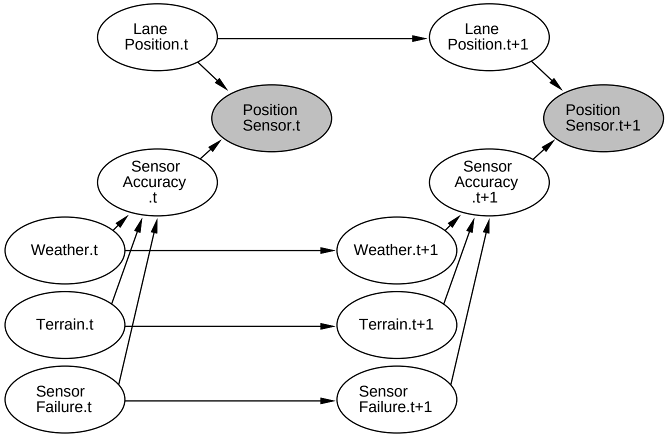
Dynamic Belief Networks

- Work with 2 slices at any one time and change network with time
- $\mathbf{Pr}(P_t|S_t) = \mathbf{Pr}(P|S)$
- $\mathbf{Pr}(S_{t+1}|S_t) = \mathbf{Pr}(S_{new}|S_{old})$
- Start with a prior distribution over S_0, denoted $Bel^*(S_0)$
- Set $i = 0$; Network always has slices $i, i+1$
- **Estimation:** Observe P_i and update the belief function

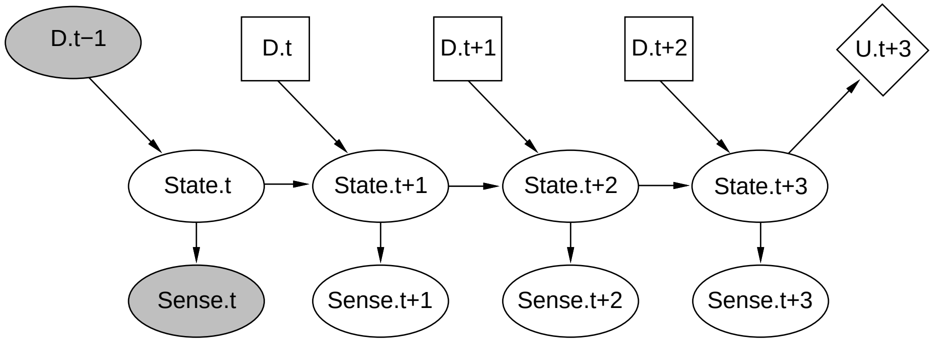
$$Bel(S_i) = \mathbf{Pr}(S_i|P_i) = \frac{\mathbf{Pr}(P|S)Bel^*(S_i)}{\mathbf{Pr}(P_i)} = Normalise(\mathbf{Pr}(P|S)Bel^*(S_i))$$

- **Prediction:** Compute $Bel^*(S_{i+1}) = \mathbf{Pr}(S_{new}|S_{old})Bel(S_i)$
- **Rollup:** Remove slice i; add slice $i+2$; set $i \leftarrow i+1$

A dynamic belief network for monitoring lane position of vehicle



Dynamic Decision Networks



Dynamic Decision Networks

- Add decision and utility nodes.
- Can deal with fixed finite horizon by putting utility on last step. (Other variations exist.)
- Similar to a planning problem augmented with probabilities.
- Cannot devise a plan ahead of time as we do not know what the percept will be.
- But can choose best next action by propagating with all information we have at the moment.

Uncertainty in User Models

The Lumiere Project by Horvitz et al. (1998)

- A Bayesian help system.
- The technology underpinning Microsoft's Office Assistant
- Model relationships among user's goals and needs, program's state, actions recently taken, and words in user query.
- Use a dynamic network to model changes over time.
- Estimate user's needs and whether they might want to get advice.
- Can decide to offer help or respond to queries.

Space Shuttle Monitoring and Control

The Vista project by Horvitz and Barry (1995)

- Human flight controllers work in controlling a space shuttle.
- Large amount of raw information.
- Vista provides decision support to flight controllers by managing the information displayed to them.
- Bayesian networks used to model sub-systems (e.g. engine), negative utilities of various anomalous conditions, user's beliefs, user's actions, the effect of display on user's decisions.
- The system presents processed information and likely diagnoses as well as suggesting possible actions.

- System operating in Johnson Space Center, Houston, Texas

Medical Diagnosis

Pathfinder system by Heckerman et al (1992).

- Diagnosis system for lymph node tissue.
- Bayesian network models diseases and symptoms.
- System offers decision support by suggesting likely diagnoses and identifying tests that will help diagnosis.
- Commercialized as IntelliPath; deals with 18 tissue types.

Medical Diagnosis

TRAUMAID system by Webber et al (1998).

- Diagnosis system for gunshot and knife wounds.
- Bayesian network models location of trauma.
- System offers decision support by suggesting likely diagnoses and identifying tests that will help diagnosis.
- Uses Utilities in a Decision Network to decide whether it is worth intervening with such decision support!
- Its diagnoses and treatments were evaluated as slightly better than actual human treatments for real cases.

Travel Arrangements

COMIC System by Moore et al. (2005).

- User-specific Utilities determine use of contrastive words like “but” in Machine-Human dialog.

User: I want to travel from Edinburgh to Brussels, arriving by 5 pm.

Student: There s a direct flight on BMI with a good price. It arrives at four ten p.m. and costs one hundred and twelve pounds. The cheapest flight is on Ryanair. It arrives at twelve forty five p.m. and costs just fifty pounds, but it requires a connection in Dublin.

BusClass: You can fly business class on British Airways, arriving at four twenty p.m., but you d need to connect in Manchester. There s a direct flight on BMI, arriving at four ten p.m., but there s no availability in business class.

Summary

- Must deal with uncertainty in various forms.
- Logical approaches are possible.
- Probabilities can be used to model uncertainty.
- Bayesian inference provides a sound way of updating beliefs given new evidence.
- Graphical network structures can be used to make this more efficient computationally (and hence feasible).
- Probabilities + utilities lead to decision theory, a prescriptive treatment of “rational decision making”.
- We can model this with decision networks.

- And Deal with change over time with dynamic decision networks.
- Interesting Applications.
- A lot more needed (both modelling and computationally).
- An active area of research.