

# COMP9414: Artificial Intelligence

## Reasoning Under Uncertainty

Wayne Wobcke

Room J17-433

wobcke@cse.unsw.edu.au

Based on slides by Maurice Pagnucco

## Overview

- Problems with Logical Approach
- What Do the Numbers Mean?
- Review of Probability Theory
- Conditional Probability and Bayes' Rule
- Bayesian Belief Networks
  - ▶ Semantics of Bayesian Networks
  - ▶ Inference in Bayesian Networks
- Conclusion

## Reasoning Under Uncertainty

- One drawback of the logical approach to reasoning is that an agent can rarely ascertain the truth of all propositions in the environment
- In fact, propositions (and their logical structure) may be inappropriate for modelling some domains – especially those involving **uncertainty**
- Rational decisions for a decision theoretic agent depend on importance of goals and the likelihood that they can be achieved
- References:
  - ▶ Ivan Bratko, [Prolog Programming for Artificial Intelligence](#), Addison-Wesley, 2001. (Chapter 15.6)
  - ▶ Stuart J. Russell and Peter Norvig, [Artificial Intelligence: A Modern Approach](#), Third Edition, Pearson Education, 2010. (Chapters 13, 14)

## Problems with Logical Approach

- Consider trying to formalise a medical diagnosis system:

$$\forall p(\text{Symptom}(p, \text{AbdominalPain}) \rightarrow \text{Disease}(p, \text{Appendicitis}))$$

- This rule is not correct since patients with abdominal pain may be suffering from other diseases

$$\forall p(\text{Symptom}(p, \text{AbdominalPain}) \rightarrow$$

$$\text{Disease}(p, \text{Appendicitis}) \vee \text{Disease}(p, \text{Ulcer}) \vee \text{Disease}(p, \text{Indig}) \dots)$$

- We could try to write a causal rule:

$$\forall p(\text{Disease}(p, \text{Ulcer}) \rightarrow \text{Symptom}(p, \text{AbdominalPain}))$$

## Sources of Uncertainty

- Difficulties arise with the logical approach due to:
  - incompleteness** agent may not have complete theory for domain
  - ignorance** agent may not have enough information about domain
  - noise** information agent does have may be unreliable
  - non-determinism** environment itself may be inherently unpredictable
- Probability gives us a way of summarising this uncertainty
  - ▶ e.g. may believe that there is a probability of 0.75 that patient suffers from appendicitis if they have abdominal pains

## Sample Space and Events

- Flip a coin three times
- The possible outcomes are:
 

TTT	TTH	THT	THH
HTT	HTH	HHT	HHH
- Set of all possible outcomes:
 
$$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$
- Any subset of the sample space is known as an **event**
- Any singleton subset of the sample space is known as a **simple event**

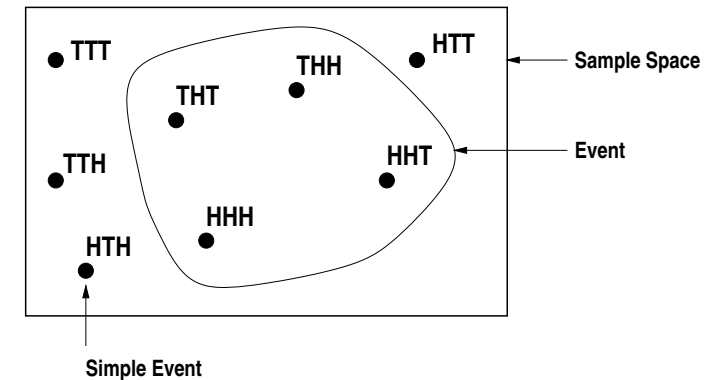
## What Do the Numbers Mean?

**Statistical/Frequentist View** Long-range frequency of a set of “events”  
 e.g. probability of the event of “heads” appearing on the toss of a coin  
 — long-range frequency of heads that appear on coin toss

**Objective View** Probabilities are real aspects of the world

**Personal/Subjective/Bayesian View** Measure of belief in proposition based on agent’s knowledge, e.g. probability of heads is measure of your belief that coin will land heads based on your belief about the coin; other agents may assign a different probability based on their beliefs (subjective)

## Sample Space and Events



## Prior Probability

- $P(A)$  **prior** or **unconditional probability** that proposition  $A$  is true
- For example,  $P(\textit{Appendicitis}) = 0.3$
- In the absence of any other information, agent believes there is a probability of 0.3 (30%) of the event of the patient suffering from appendicitis
- As soon as we get new information we must reason with **conditional probabilities**

## Axioms of Probability

1.  $0 \leq P(A) \leq 1$ 
  - All probabilities are between 0 and 1
2.  $P(\textit{True}) = 1$        $P(\textit{False}) = 0$ 
  - Valid propositions have probability 1
  - Unsatisfiable propositions have probability 0
3.  $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ 
  - Can determine probabilities of all other propositions
  - For example,  $P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$   
 $P(\textit{True}) = P(A) + P(\neg A) - P(\textit{False})$   
 $1 = P(A) + P(\neg A) - 0$   
Therefore  $P(\neg A) = 1 - P(A)$

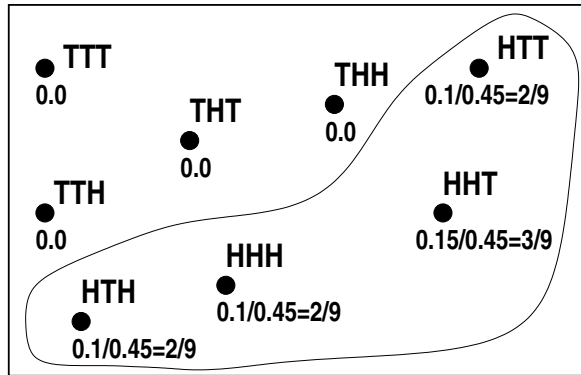
## Random Variables

- Propositions are **random variables** that can take on several values
  - $P(\textit{Weather} = \textit{Sunny}) = 0.8$
  - $P(\textit{Weather} = \textit{Rain}) = 0.1$
  - $P(\textit{Weather} = \textit{Cloudy}) = 0.09$
  - $P(\textit{Weather} = \textit{Snow}) = 0.01$
- Every random variable  $X$  has a **domain** of possible values  $\langle x_1, x_2, \dots, x_n \rangle$
- Probabilities of all possible values  $\mathbf{P}(\textit{Weather}) = \langle 0.8, 0.1, 0.09, 0.01 \rangle$  is a **probability distribution**
- $\mathbf{P}(\textit{Weather}, \textit{Appendicitis})$  is a combination of random variables represented by cross product (can also use logical connectives  $P(A \wedge B)$  to represent compound events)

## Conditional Probability

- When new information is gained we can no longer use prior probabilities
- **Conditional** or **posterior** probability  
 $P(A|B)$  is the probability of  $A$  given that all we know is  $B$ 
  - ▶ e.g.  $P(\textit{Appendicitis}|\textit{AbdominalPain}) = 0.75$
- **Product Rule:**  $P(A \wedge B) = P(A|B).P(B)$
- Therefore  $P(A|B) = \frac{P(A \wedge B)}{P(B)}$  provided  $P(B) > 0$
- $\mathbf{P}(X|Y) = P(X = x_i|Y = y_j)$  for all  $i, j$   
 $\mathbf{P}(X, Y) = \mathbf{P}(X|Y).\mathbf{P}(Y)$  — a set of equations

## Normalisation



- Conditional probability distribution given that first coin is H

## Joint Probability Distribution

- Simple events are mutually exclusive and jointly exhaustive
- Probability of complex event is sum of probabilities of compatible simple events

$$P(\text{Appendicitis}) = 0.04 + 0.06 = 0.10$$

$$P(\text{Appendicitis} \vee \text{AbdominalPain}) = 0.04 + 0.06 + 0.01 = 0.11$$

$$P(\text{Appendicitis} | \text{AbdominalPain}) = \frac{P(\text{Appendicitis} \wedge \text{AbdominalPain})}{P(\text{AbdominalPain})} = \frac{0.04}{0.04 + 0.01} = 0.8$$

- Problem:** With many random variables the number of probabilities is vast

## Joint Probability Distribution

- Complete specification of probabilities to all propositions in the domain
- Suppose we have random variables  $X_1, X_2, \dots, X_n$
- An **atomic (simple) event** is an assignment of particular values to all variables
- Joint probability distribution  $\mathbf{P}(X_1, X_2, \dots, X_n)$  assigns probabilities to all possible atomic events
- For example, a simple medical domain with two Boolean random variables:

	AbdominalPain	$\neg$ AbdominalPain
Appendicitis	0.04	0.06
$\neg$ Appendicitis	0.01	0.89

## Bayes' Rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- Modern AI systems abandon joint probabilities and work with conditional probabilities utilising Bayes' Rule

- Deriving Bayes' Rule:

$$P(A \wedge B) = P(A|B)P(B) \quad (\text{Definition})$$

$$P(B \wedge A) = P(B|A)P(A) \quad (\text{Definition})$$

$$\text{So } P(A|B)P(B) = P(B|A)P(A) \text{ since } P(A \wedge B) = P(B \wedge A)$$

$$\text{Hence } P(B|A) = \frac{P(A|B)P(B)}{P(A)} \text{ if } P(A) \neq 0$$

- Note:** If  $P(A) = 0$ ,  $P(B|A)$  is undefined

## Applying Bayes' Rule

- Example (Russell & Norvig, 1995)
- Doctor knows that
  - meningitis causes a stiff neck 50% of the time
  - chance of patient having meningitis is  $\frac{1}{50000}$
  - chance of patient having a stiff neck is  $\frac{1}{20}$
- $P(\text{StiffNeck}|\text{Meningitis}) = 0.5$   
 $P(\text{Meningitis}) = \frac{1}{50000}$   
 $P(\text{StiffNeck}) = \frac{1}{20}$
- $P(\text{Meningitis}|\text{StiffNeck}) = \frac{P(\text{StiffNeck}|\text{Meningitis}) \cdot P(\text{Meningitis})}{P(\text{StiffNeck})} = 0.5 \frac{1}{50000} \frac{1}{\frac{1}{20}} = 0.0002$

## Conditional Independence

- **Observe:** Appendicitis is direct cause of both abdominal pain and nausea
- If we know patient is suffering from appendicitis, then probability of nausea should not depend on the presence of abdominal pain; likewise probability of abdominal pain should not depend on nausea
- We say that nausea and abdominal pain are **conditionally independent** given appendicitis
- An event  $X$  is independent of an event  $Y$  conditional on the background knowledge  $K$  if knowing  $Y$  does not affect the probability of  $X$  given  $K$

$$P(X|K) = P(X|Y, K)$$

## Using Bayes' Rule

- Suppose we have two conditional probabilities for appendicitis
 
$$P(\text{Appendicitis}|\text{AbdominalPain}) = 0.8$$

$$P(\text{Appendicitis}|\text{Nausea}) = 0.1$$
- $P(\text{Appendicitis}|\text{AbdominalPain} \wedge \text{Nausea}) = \frac{P(\text{AbdominalPain} \wedge \text{Nausea}|\text{Appendicitis}) \cdot P(\text{Appendicitis})}{P(\text{AbdominalPain} \wedge \text{Nausea})}$
- Need to know  $P(\text{AbdominalPain} \wedge \text{Nausea}|\text{Appendicitis})$   
With more symptoms that is a daunting task

## Bayesian Belief Networks

- A **Bayesian belief network** (also **Bayesian Network**, **probabilistic network**, **causal network**, **knowledge map**) is a directed acyclic graph (DAG) where:
  - ▶ Each node corresponds to a random variable
  - ▶ Directed links connect pairs of nodes – a directed link from node  $X$  to node  $Y$  means that  $X$  has a **direct influence** on  $Y$
  - ▶ Each node has a conditional probability table quantifying effect of parents on node
- Independence assumption of Bayesian networks:
  - Each random variable is (conditionally) independent of its nondescendants given its parents

## Bayesian Belief Networks

- Example (Pearl, 1988)
- You have a new burglar alarm at home that is quite reliable at detecting burglars but may also respond at times to an earthquake. You also have two neighbours, John and Mary, who promise to call you at work when they hear the alarm. John always calls when he hears the alarm but sometimes confuses the telephone ringing with the alarm and calls then, also Mary likes loud music and sometimes misses the alarm. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

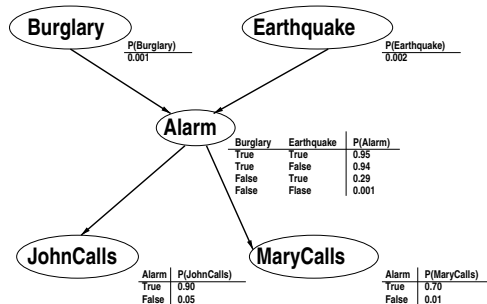
## Conditional Probability Table

- Row contains conditional probability of each node value for a conditioning case (i.e. possible combination of values for parent node)

Burglary	Earthquake	$P(Alarm Burglary \wedge Earthquake)$	
		True	False
True	True	0.950	0.050
True	False	0.940	0.060
False	True	0.290	0.710
False	False	0.001	0.999

## Bayesian Belief Networks

- Example (Pearl, 1988)



- Probabilities summarise potentially infinite set of possible circumstances

## Semantics of Bayesian Networks

- Bayesian network provides a complete description of the domain
- Joint probability distribution can be determined from the belief network
  - ▶  $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i|Parents(X_i))$
- For example,  $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J|A) \cdot P(M|A) \cdot P(A|\neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628$
- Bayesian network is a complete and non-redundant representation of domain (and can be far more compact than joint probability distribution)

## Semantics of Bayesian Networks

- Factorisation of joint probability distribution
- Chain Rule:** Use conditional probabilities to decompose conjunctions  

$$P(X_1 \wedge X_2 \wedge \dots \wedge X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1 \wedge X_2) \cdot \dots \cdot P(X_n|X_1 \wedge X_2 \wedge \dots \wedge X_{n-1})$$
- Now, order the variables  $X_1, X_2, \dots, X_n$  in a belief network so that a variable comes after its parents – let  $\pi_{X_i}$  be the tuple of parents of variable  $X_i$  (this is a complex random variable)  
 Using the chain rule we have  $P(X_1 \wedge X_2 \wedge \dots \wedge X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1 \wedge X_2) \cdot \dots \cdot P(X_n|X_1 \wedge X_2 \wedge \dots \wedge X_{n-1})$

## Semantics of Bayesian Networks

- Each  $P(X_i|X_1 \wedge X_2 \wedge \dots \wedge X_{i-1})$  has the property that it is not conditioned on a descendant of  $X_i$  (given ordering of variables in belief network)
- Therefore, by conditional independence we have  $P(X_i|X_1 \wedge X_2 \wedge \dots \wedge X_{i-1}) = P(X_i|\pi_{X_i})$
- That is, rewriting the chain rule  $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i|\pi_{X_i})$

## Calculation using Bayesian Networks

- Fact 1:** Consider random variable  $X$  with parents  $Y_1, Y_2, \dots, Y_n$ :

$$P(X|Y_1 \wedge \dots \wedge Y_n \wedge Z) = P(X|Y_1 \wedge \dots \wedge Y_n)$$

if  $Z$  doesn't involve a descendant of  $X$  (including  $X$  itself)

- Fact 2:** If  $Y_1, \dots, Y_n$  are pairwise disjoint and exhaust all possibilities:

$$P(X) = \sum P(X \wedge Y_i) = \sum P(X|Y_i) \cdot P(Y_i)$$

$$P(X|Z) = \sum P(X \wedge Y_i|Z)$$

- e.g.  $P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\sum P(J \wedge B \wedge e \wedge a \wedge m)}{\sum P(j \wedge B \wedge e \wedge a \wedge m)}$  where  $j$  ranges over  $J, \neg J$ ,  $e$  over  $E, \neg E$ ,  $a$  over  $A, \neg A$  and  $m$  over  $M, \neg M$

## Calculation using Bayesian Networks

- $P(J \wedge B \wedge E \wedge A \wedge M) = P(J|A) \cdot P(B) \cdot P(E) \cdot P(A|B \wedge E) \cdot P(M|A) = 0.90 \times 0.001 \times 0.002 \times 0.95 \times 0.70 = 0.00000197$
- $P(J \wedge B \wedge \neg E \wedge A \wedge M) = 0.00591016$
- $P(J \wedge B \wedge E \wedge \neg A \wedge M) = 5 \times 10^{-11}$
- $P(J \wedge B \wedge \neg E \wedge \neg A \wedge M) = 2.99 \times 10^{-8}$
- $P(J \wedge B \wedge E \wedge A \wedge \neg M) = 0.000000513$
- $P(J \wedge B \wedge \neg E \wedge A \wedge \neg M) = 0.000253292$
- $P(J \wedge B \wedge E \wedge \neg A \wedge \neg M) = 4.95 \times 10^{-9}$
- $P(J \wedge B \wedge \neg E \wedge \neg A \wedge \neg M) = 2.96406 \times 10^{-6}$

## Calculation using Bayesian Networks

- $P(\neg J \wedge B \wedge E \wedge A \wedge M) = 0.000000133$
- $P(\neg J \wedge B \wedge \neg E \wedge A \wedge M) = 6.56684 \times 10^{-5}$
- $P(\neg J \wedge B \wedge E \wedge \neg A \wedge M) = 9.5 \times 10^{-10}$
- $P(\neg J \wedge B \wedge \neg E \wedge \neg A \wedge M) = 5.6886 \times 10^{-7}$
- $P(\neg J \wedge B \wedge E \wedge A \wedge \neg M) = 0.000000057$
- $P(\neg J \wedge B \wedge \neg E \wedge A \wedge \neg M) = 2.81436 \times 10^{-5}$
- $P(\neg J \wedge B \wedge E \wedge \neg A \wedge \neg M) = 9.405 \times 10^{-8}$
- $P(\neg J \wedge B \wedge \neg E \wedge \neg A \wedge \neg M) = 5.63171 \times 10^{-5}$

## Inference in Bayesian Networks

**Diagnostic Inference** From effects to causes

$$P(\text{Burglary}|\text{JohnCalls}) = 0.016$$

**Causal Inference** From causes to effects

$$P(\text{JohnCalls}|\text{Burglary}) = 0.85; P(\text{MaryCalls}|\text{Burglary}) = 0.67$$

**Intercausal Inference** Explaining away

$P(\text{Burglary}|\text{Alarm}) = 0.3736$  but adding evidence,  $P(\text{Burglary}|\text{Alarm} \wedge \text{Earthquake}) = 0.003$ ; despite the fact that burglaries and earthquakes are independent, the presence of one makes the other **much** less likely

**Mixed Inference** Combinations of the patterns above

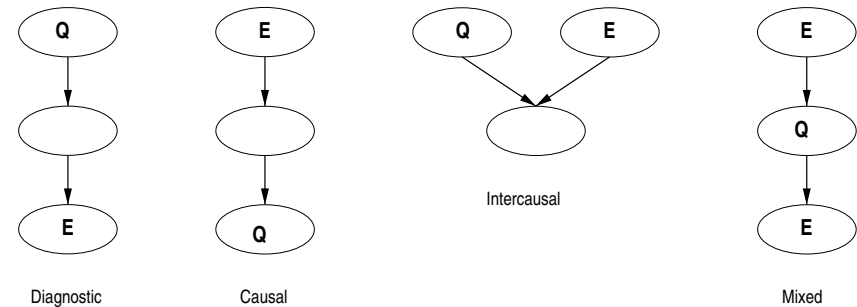
Diagnostic + Causal:  $P(\text{Alarm}|\text{JohnCalls} \wedge \neg \text{Earthquake})$

Intercausal + Diagnostic:  $P(\text{Burglary}|\text{JohnCalls} \wedge \neg \text{Earthquake})$

## Calculation using Bayesian Networks

- Therefore,  $P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\sum P(J \wedge B \wedge e \wedge a \wedge m)}{\sum P(j \wedge B \wedge e \wedge a \wedge m)} = \frac{0.00849017}{0.001}$
- $P(J|B) = 0.849017$
- Can often simplify calculation without using full joint probabilities but not always

## Inference in Bayesian Networks



- $Q$  = query;  $E$  = evidence

## Example — Causal Inference

- $P(\text{JohnCalls}|\text{Burglary})$
- $$P(J|B) = P(J|A \wedge B) \cdot P(A|B) + P(J|\neg A \wedge B) \cdot P(\neg A|B)$$

$$= P(J|A) \cdot P(A|B) + P(J|\neg A) \cdot P(\neg A|B)$$

$$= P(J|A) \cdot P(A|B) + P(J|\neg A) \cdot (1 - P(A|B))$$
- Now  $P(A|B) = P(A|B \wedge E) \cdot P(E|B) + P(A|B \wedge \neg E) \cdot P(\neg E|B)$ 

$$= P(A|B \wedge E) \cdot P(E) + P(A|B \wedge \neg E) \cdot P(\neg E)$$

$$= 0.95 \times 0.002 + 0.94 \times 0.998 = 0.94002$$
- Therefore  $P(J|B) = 0.90 \times 0.94002 + 0.05 \times 0.05998 = 0.849017$
- **Fact 3:**  $P(X|Z) = P(X|Y \wedge Z) \cdot P(Y|Z) + P(X|\neg Y \wedge Z) \cdot P(\neg Y|Z)$ , since  $X \wedge Z \equiv (X \wedge Y \wedge Z) \vee (X \wedge \neg Y \wedge Z)$  (conditional version of Fact 2)

## Conclusion

- Due to noise or uncertainty it may be advantageous to reason with probabilities
- Dealing with joint probabilities can become difficult due to the large number of values involved
- Use of Bayes' Rule and conditional probabilities may be a way around this
- Bayesian belief networks allow compact representation of probabilities and efficient reasoning with probabilities
- They work by exploiting the notion of conditional independence
- Elegant recursive algorithms can be given to automate the process of inference in Bayesian networks
- This is currently one of the "hot" topics in AI

## Example — Diagnostic Inference

- $P(\text{Earthquake}|\text{Alarm})$
- $$P(E|A) = \frac{P(A|E) \cdot P(E)}{P(A)}$$

$$= \frac{P(A|B \wedge E) \cdot P(B) \cdot P(E) + P(A|\neg B \wedge E) \cdot P(\neg B) \cdot P(E)}{P(A)}$$

$$= \frac{0.95 \times 0.001 \times 0.002 + 0.29 \times 0.999 \times 0.002}{0.002516442} = \frac{5.8132 \times 10^{-4}}{0.002516442}$$
- Now  $P(A) = P(A|B \wedge E) \cdot P(B) \cdot P(E) + P(A|\neg B \wedge E) \cdot P(\neg B) \cdot P(E) + P(A|B \wedge \neg E) \cdot P(B) \cdot P(\neg E) + P(A|\neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E)$   
And  $P(A|B \wedge \neg E) \cdot P(B) \cdot P(\neg E) + P(A|\neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E)$   
 $= 0.94 \times 0.001 \times 0.998 + 0.001 \times 0.999 \times 0.998 = 0.001935122$   
So  $P(A) = 5.8132 \times 10^{-4} + 0.001935122 = 0.002516442$
- Therefore  $P(E|A) = \frac{5.8132 \times 10^{-4}}{0.002516442} = 0.2310087$
- **Fact 4:**  $P(X \wedge Y) = P(X) \cdot P(Y)$  if  $X, Y$  are conditionally independent