Reasoning with Uncertainty

LECTURE 4 - part I

Uncertainty • Probability (Syntax and Semantics & Inference)
Conditional Independence and Bayes' Rule
Bayesian Networks (Semantics of Bayesian Networks & Inference in Bayesian Networks)

UNSW
SYDNEY

Reasoning with Uncertainty

- An agent can not always ascertain the truth of all propositions, so may not only have "flat out" beliefs (P or $\neg P$)
- Some environments themselves generate uncertainty for the agent, due to unpredictability or non-determinism, so propositions inadequately model those environments
- Rational decisions for an agent require tradeoffs between the importance of goals and the likelihood of achieving them, and the cost of acting and not achieving them

Uncertainty

In many situations, an AI agent has to choose an action based on incomplete information.

- stochastic environments (e.g. dice rolls in Backgammon)
- partial observability
  - some aspects of environment hidden from agent
  - robots can have noisy sensors, reporting quantities which differ from the “true” values

Problems with Logical Approach

- Consider trying to formalise a medical diagnosis system
- This rule is not correct since patients with abdominal pain may be suffering from other diseases

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

- How about a causal rule?

$$\forall p(Symptom(p, AbdominalPain) \rightarrow Disease(p, Appendicitis) \wedge Disease(p, Ulcer) \wedge Disease(p, Indig)\ldots)$$

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

Planning under Uncertainty

Let action *At* = leave for airport *t* minutes before flight

Will *At* get me there on time? Problems:

A30 gets me there on time ?

A90 gets me there on time?

A120 gets me there on time?

A1440 gets me there on time?

Planning under Uncertainty

Let action *At* = leave for airport *t* minutes before flight

Will *At* get me there on time? Problems:

- partial observability, noisy sensors
- uncertainty in action outcomes (flat tyre, etc.)
- immense complexity of and predicting traffic

Hence a purely logical approach either

1) risks falsehood: “A30 will get me there on time”, or
2) leads to conclusions that are too weak for decision making:
“A30 will get me there on time if there’s no accident on the bridge and it doesn’t rain and my tires remain intact etc etc.”
(A1440 might be safe but I’d have to stay overnight in the airport ...)

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 stochastic
  - **unpredictability** environment may be inherently unpredictable

- Probability gives 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

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 — objective
- Personal/Subjective/Bayesian View
  - Measure of belief in proposition based on agent’s knowledge, e.g. probability of heads is a degree of belief that coin will land heads based on beliefs about the coin or could be just a guess; different agents may assign a different probability — subjective

Sample Space and Events

- Flip a coin three times
- The possible outcomes are

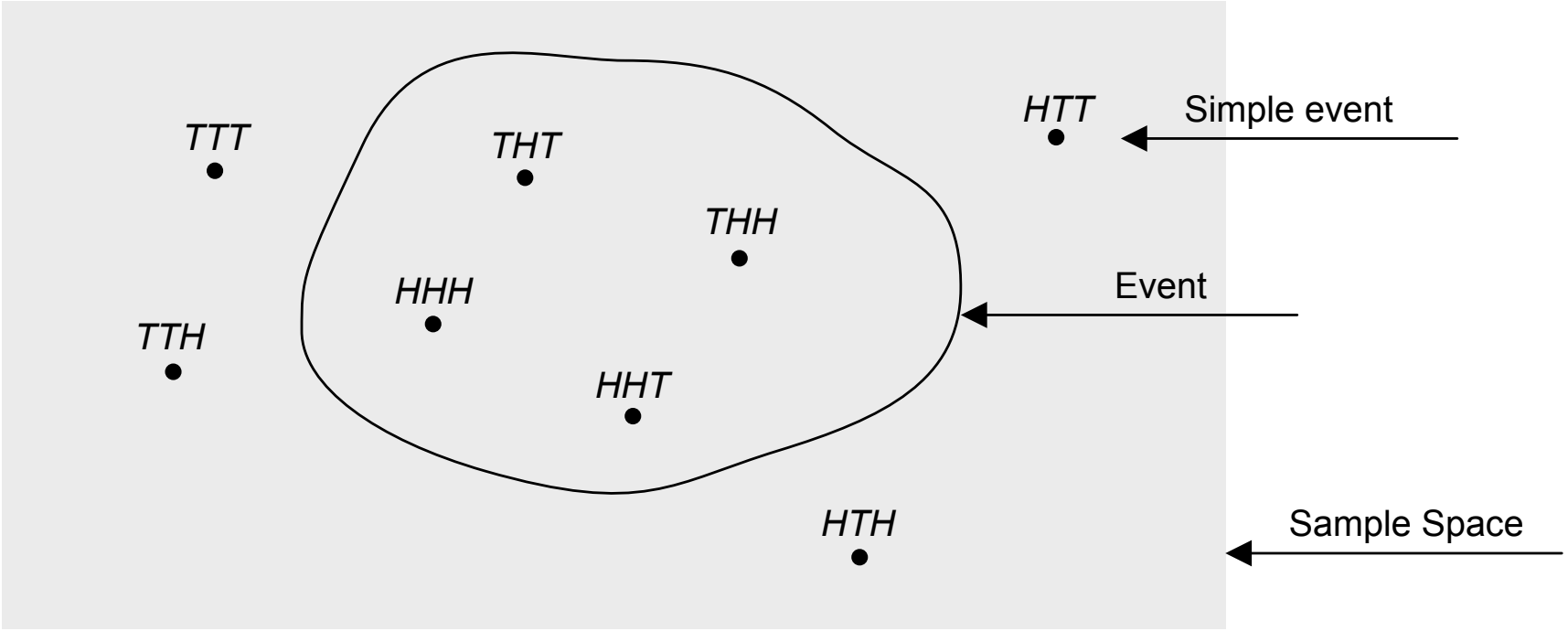
$$TTT \quad TTH \quad THT \quad THH$$
$$HTT \quad HTH \quad HHT \quad HHH$$

- 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

Sample Space and Events



Methods for handling Uncertainty

Default or non-monotonic logic:

- Assume my car does not have a flat tire, etc.
- Assume *A*30 works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?

Probability:

- Given the available evidence, *A*30 will get me there on time with probability 0.04
- Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

Probability

Probabilistic assertions summarise effects of:

- Laziness: failure to enumerate exceptions, qualifications, etc.
- Ignorance: lack of relevant facts, initial conditions, etc.

Subjective or Bayesian probability:

- Probabilities relate propositions to one's own state of knowledge e.g. $P(A30|\text{no reported accidents}) = 0.06$
- These are not claims of a "probabilistic tendency" in the current situation (but might be learned from past experience of similar situations)
- Probabilities of propositions change with new evidence:
- e.g. $P(A30|\text{no reported accidents, 5 a.m.}) = 0.15$
- (Analogous to logical entailment status $KB \models \alpha$, not absolute truth)

Making decisions under uncertainty

Suppose I believe the following:

P(A30 gets me there on timel...) = 0.04
P(A90 gets me there on timel...) = 0.70
P(A120 gets me there on timel...) = 0.95
P(A1440 gets me there on timel...) = 0.9999

Which action to choose?

- Depends on my preferences for missing flight vs. airport cuisine, etc.
- Utility theory is used to represent and infer preferences
- Decision theory = utility theory + probability theory

Probability basics

Begin with a set Ω - the **sample space** (e.g. 6 possible rolls of a die)

$\omega \in \Omega$ is a sample point / atomic event / possible world

A **probability space** or probability model is a sample space with an assignment P(ω) for every $\omega \in \Omega$ s.t.

$$0 \leq P(\omega) \leq 1 \text{ and } \Sigma_{\omega} P(\omega) = 1$$

An **event** A is any subset of Ω

$$P(A) = \Sigma_{\{\omega \in A\}} P(\omega)$$

$$\text{e.g. } P(\text{die roll} < 4) = P(\omega=1) + P(\omega=2) + P(\omega=3) = 1/6 + 1/6 + 1/6 = 1/2$$

Random variables

A random variable is a function from sample points to some range (e.g. the Reals or Booleans)

For example, *Odd(3) = true.*
P induces a probability distribution for any random variable *X*:

$$P(X=x_i) = \Sigma_{\{\omega: X(\omega)= x_i\}} P(\omega)$$

$$e.g.\ P(Odd = true) = P(\omega=1) + P(\omega=3) + P(\omega=3) = 1/6 + 1/6 + 1/6 = 1/2$$

Random Variables

- Propositions are random variables that can take on several values

$$P(Weather = Sunny) = 0.8$$
$$P(Weather = Rain) = 0.1$$
$$P(Weather = Cloudy) = 0.09$$
$$P(Weather = Snow) = 0.01$$

- Every random variable X has a domain of possible values $\langle x_1, x_2, \cdots, x_n \rangle$
- Probabilities of all possible values $P(Weather) = \langle 0.8, 0.1, 0.09, 0.01 \rangle$ is a **probability distribution**
- $P(Weather, Appendicitis)$ is a combination of random variables represented by cross product (can also use logical connectives $P(A \wedge B)$ to represent compound events)

Propositions

Think of a proposition as the event (set of sample points) where the proposition is true
Given Boolean random variables A and B:

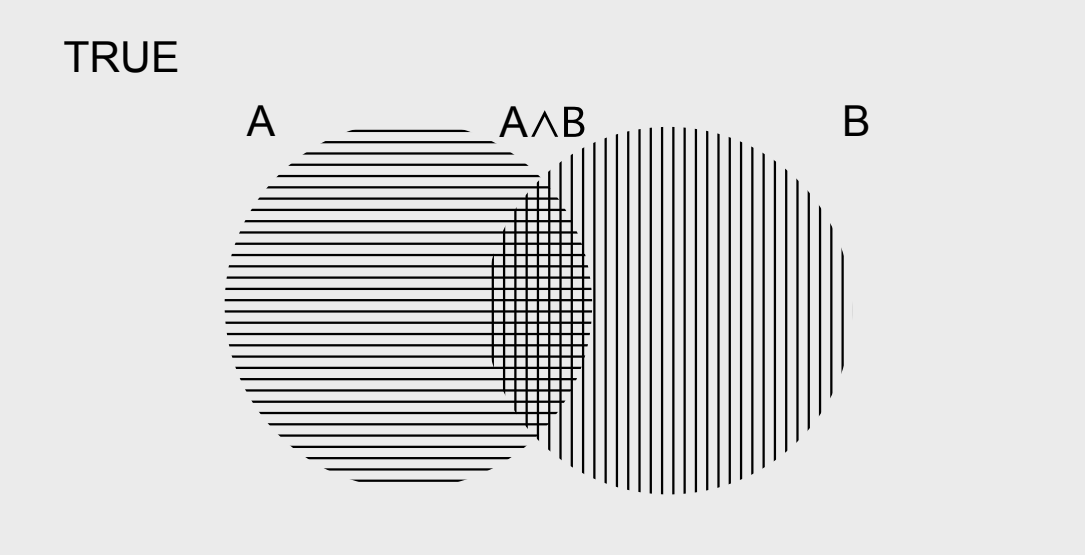
event a = set of sample points where A(ω) = true
event ¬a = set of sample points where A(ω) = false
event a ∧ b = set of sample points where A(ω) = true and B(ω) = true

With Boolean variables for A and B, proposition = disjunction of atomic events in which it is true:

$$e.g,\ P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$$

Why use probability?

The definitions imply that certain logically related events must have related probabilities



de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

Syntax for propositions

Propositional or Boolean random variables
e.g., `Cavity` (do I have a cavity?)
`Cavity = true` is a proposition, also written `Cavity`

Discrete random variables (finite or infinite)
e.g., `Weather` is one of ⟨`sunny`,`rain`,`cloudy`,`snow`⟩
`Weather = rain` is a proposition
Values must be exhaustive and mutually exclusive

Continuous random variables (bounded or unbounded) e.g.
`Temp = 21.6`; also allow, e.g. `Temp < 22.0`

Arbitrary Boolean combinations of basic propositions.

Prior probability

Prior or unconditional probabilities of propositions
e.g. $P(\texttt{Cavity = true}) = 0.1$ and $P(\texttt{Weather = sunny}) = 0.72$

correspond to belief prior to arrival of any (new) evidence.

Probability distribution gives values for all possible assignments:
$P(\texttt{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Prior Probability

- $P(A)$ is the prior or unconditional probability that an event A occurs
- For example, $P(Appendicitis)=0.3$
- In the absence of any other information, agent believes there is a probability of 0.3 (30%) that the patient suffers from appendicitis
- To account for the effect of new information on probabilities, the agent must reason with conditional probabilities

Axioms of Probability

1. $0 \leq P(A) \leq 1$
   All probabilities are between 0 and 1

2. $P(True)=1$ $P(False)=0$
   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

Axioms of Probability

1. $0 \leq P(A) \leq 1$
   All probabilities are between 0 and 1

2. $P(True)=1$ $P(False)=0$
   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, $P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$

$$P(True) = P(A) + P(\neg A) - P(False)$$
$$1 = P(A)+P(\neg A)-0$$

Therefore $\boldsymbol{P(\neg A) = 1 - P(A)}$

Probabilistic Agents

We consider an Agent whose World Model consists not of a set of facts, but rather a set of **probabilities** of certain facts being true, or certain random variables taking particular values.

When the Agent makes an observation, it may **update** its World Model by adjusting these probabilities, based on what it has observed.

Example: Tooth Decay

Assume you live in a community where, at any given time, 20% of people have a cavity in one of their teeth which needs a filling from the dentist.

$$P(\texttt{cavity}) = 0.2$$

We also know about the weather that can be

$$\langle \textit{sunny}, \textit{rain}, \textit{cloudy}, \textit{snow} \rangle$$

Joint probability

Joint probability distribution for a set of r.v.'s gives the probability of every atomic event on those r.v's (i.e., every sample point)

P(Weather, Cavity) is a 4x2 matrix of probability values

Weather =	Sunny	Rain	Cloudy	Snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

What is the probability of being sunny?
What is the probability of having a Cavity?

Joint probability

Joint probability distribution for a set of r.v.'s gives the probability of every atomic event on those r.v's (i.e., every sample point)

P(Weather, Cavity) is a 4x2 matrix of probability values

Weather =	Sunny	Rain	Cloudy	Snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points.

Example: Tooth Decay

Assume you live in a community where, at any given time, 20% of people have a cavity in one of their teeth which needs a filling from the dentist.

$$P(cavity) = 0.2$$

If you have a toothache, suddenly you will think it is much more likely that you have a cavity, perhaps as high as 60%. We say that the **conditional probability** of cavity, given toothache, is 0.6, written as follows:

$$P(cavity|toothache) = 0.6$$

If you go to the dentist, they will use a small hook-shaped instrument called a probe, and check whether this probe can catch on the back of your tooth. If it does catch, this information will increase the probability that you have a cavity.

Conditional Probability

- Need to **update** probabilities based on new information
- Use **conditional** or **posterior probability**
- P(AIB) is the probability of A given that all we know is B

$$P(cavity \mid toothache) = 0.6$$

Definition:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \quad \text{provided } P(B) > 0$$

Product Rule:

$$P(A \wedge B) = P(A \mid B) \bullet P(B)$$

Joint Probability Distribution

- Complete specification of probabilities to all events in domain
- Suppose random variables $X_1, X_2, \cdots, X_n$
- An atomic (simple) event is an assignment of values to all variables
- Joint probability distribution $P(X_1, X_2, \cdots, X_n)$ assigns probabilities to all possible atomic events
- Simple medical domain with two Boolean random variables

	toothache	~toothache
cavity	0.12	0.08
~cavity	0.08	0.72

Joint Probability Distribution

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

	toothache	~toothache
cavity	0.12	0.08
~cavity	0.08	0.72

$P(cavity) = 0.12 + 0.08 = 0.2$

$P(cavity \vee toothache) = 0.12 + 0.08 + 0.08 = 0.26$

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.12}{0.12 + 0.08}$$

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

Joint Probability Distribution

We assume there is some underlying joint probability distribution over the three random variables Toothache, Cavity and Catch, which we can write in the form of a table:

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Note that the sum of the entries in the table is 1.0 .
For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \Sigma_{\{\omega:\, \omega \models \varphi)\}}\ P(\omega)$$

Inference by Enumeration

Start with the joint distribution:

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \Sigma_{\omega:\, \omega \models \varphi} \; P(\omega)$$

Inference by Enumeration

Start with the joint distribution:

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \Sigma_{\omega:\, \omega \models \varphi}\ P(\omega)$$

Inference by Enumeration

Start with the joint distribution:

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \Sigma_{\omega:\, \omega \models \varphi}\ P(\omega)$$

$$P(cavity \vee toothache) =$$
$$0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by Enumeration

Start with the joint distribution:

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Can also compute conditional probabilities:

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

Conditional Probability

If we consider two random variables a and b, with $P(b) \neq 0$, then the conditional probability of a given b is

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)} \quad \text{provided } P(b) > 0 \qquad P(a \wedge b) = P(a \mid b) \cdot P(b) = P(b \mid a) \cdot P(a)$$

When an agent considers a sequence of random variables at successive time steps, they can be chained together using this formula repeatedly:

$$P(X_n, \ldots, X_1) = P(X_n \mid X_{n-1}, \ldots, X_1) P(X_{n-1}, \ldots, X_1)$$

$$= P(X_n \mid X_{n-1}, \ldots, X_1) P(X_{n-1} \mid X_{n-2}, \ldots, X_1)$$

$$= \ldots = \prod_{i=1}^{n} P(X_i \mid X_{i-1}, \ldots, X_1)$$

Independent Variables

Let's consider the joint probability distribution for Cavity and Weather.

Weather =	Sunny	Rain	Cloudy	Snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

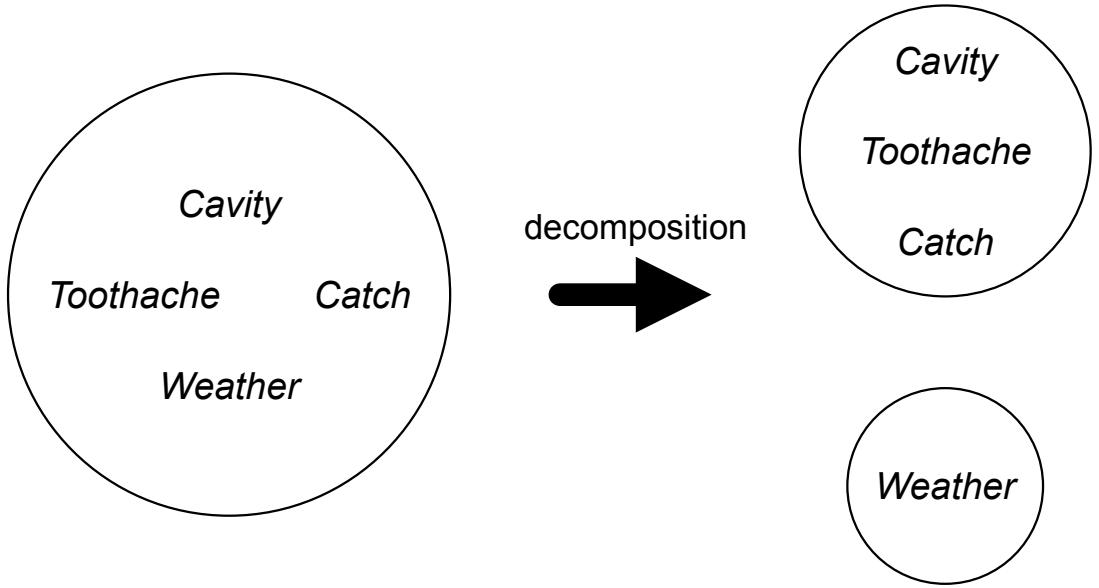
Ratio 1:4 is constant

Note that:

$$P(cavity \mid Weather = sunny) = \frac{0.144}{0.144 + 0.576} = 0.2 = P(cavity)$$

In other words, learning that the Weather is sunny has no effect on the probability of having a cavity (and the same for rain, cloudy and snow). We say that Cavity and Weather are **independent** variables.

Independent Variables



In other words, learning that the Weather is sunny has no effect on the probability of having a cavity (and the same for rain, cloudy and snow). We say that Cavity and Weather are **independent** variables.

Independent Variables

- A and B are independent iff

$$P(A \mid B) = P(A) \text{ or } P(B \mid A) = P(B) \text{ or } P(A, B) = P(A)\,P(B)$$

If variables not independent, would need 32 items in probability table.
Because Weather is independent of the other variables, only need two smaller tables, with a total of 8+4=12 items.

$$P(\text{Toothache, Catch, Cavity, Weather}) = P(\text{Toothache, Catch, Cavity})\,P(\text{Weather})$$

(Note: the number of free parameters is slightly less, because the values in each table must sum to 1).

Conditional independence

The variables Toothache, Cavity and Catch are not independent. But, they do exhibit **conditional independence**.

If you have a cavity, the probability that the probe will catch is 0.9, no matter whether you have a toothache or not.

If you don't have a cavity, the probability that the probe will catch is 0.2, regardless of whether you have a toothache. In other words,

$$P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity})$$

We say that Catch is conditionally independent of Toothache given Cavity.

Conditional independence

If you don't have a cavity, the probability that the probe will catch is 0.2, regardless of whether you have a toothache. In other words,

$$P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity})$$

We say that Catch is conditionally independent of Toothache given Cavity.

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Bayes' Rule

The formula for conditional probability can be manipulated to find a relationship when the two variables are swapped:

$$P(a \wedge b) = P(a \mid b) \cdot P(b) = P(b \mid a) \cdot P(a)$$

$$\rightarrow \text{Bayes' Rule } P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

This is often useful for assessing the probability of an underlying **cause** after an **effect** has been observed:

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)\, P(Cause)}{P(Effect)}$$

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- AI systems abandon joint probabilities and work directly with conditional probabilities using Bayes' Rule
- Deriving Bayes' Rule:

$$P(A \wedge B) = P(A \mid B)\, P(B) \qquad \text{(definition)}$$
$$P(B \wedge A) = P(B \mid A)\, P(A) \qquad \text{(definition)}$$

$$\text{Since } P(A \wedge B) = P(B \wedge A): \; P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$$

$$\text{Hence: } P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- Note: If $P(A) = 0$, the $P(B \mid A)$ 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 1 for 50000 cases
  - chance of patient having a stiff neck is 1 for 20 cases

$$P(StiffNeck \mid Meningitis) = 0.5$$
$$P(Meningitis) = 1/50000$$
$$P(StiffNeck) = 1/20$$

$$P(Meningitis \mid StiffNeck) = \frac{P(StiffNeck \mid Meningitis)\, P(Meningitis)}{P(StiffNeck)} = \frac{0.5 \cdot 1/50000}{1/20}$$

Normalisation

- Avoiding assessment of symptoms

$$P(Meningitis \mid StiffNeck) = \frac{P(StiffNeck \mid Meningitis)\, P(Meningitis)}{P(StiffNeck)}$$

$$P(\neg Meningitis \mid StiffNeck) = \frac{P(StiffNeck \mid \neg Meningitis)\, P(\neg Meningitis)}{P(StiffNeck)}$$

$$P(StiffNeck) = P(StiffNeck \mid Meningitis)\, P(Meningitis) + P(StiffNeck \mid \neg Meningitis)\, P(\neg Meningitis)$$

Since $P(Meningitis \mid StiffNeck) + P(\neg Meningitis \mid StiffNeck) = 1$

$$So\ P(Meningitis \mid StiffNeck) = \frac{P(StiffNeck \mid Meningitis)\, P(Meningitis)}{P(StiffNeck \mid Meningitis)\, P(Meningitis) + P(StiffNeck \mid \neg Meningitis)\, P(\neg Meningitis)}$$

Similarly $P(\neg Meningitis \mid StiffNeck)$

Therefore, from both *P(StiffNeck I ...)* can derive *P(StiffNeck)* and the denominator is normalisation factor

Conditional Independence

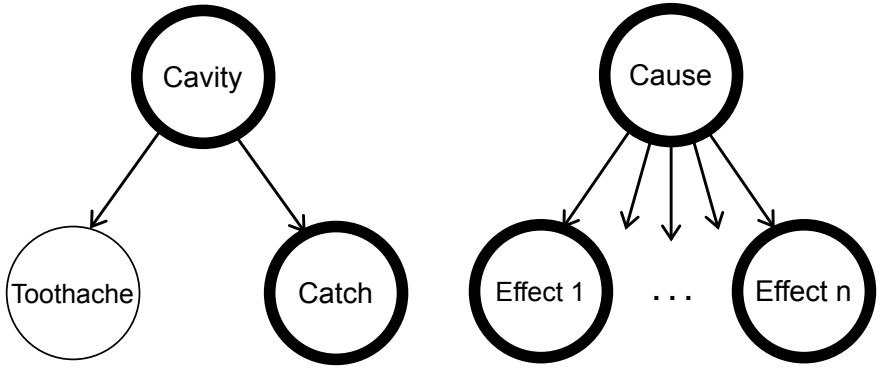
- Appendicitis is direct cause of both abdominal pain and nausea
- If we know a patient is suffering from appendicitis, the probability of nausea should not depend on the presence of abdominal pain; likewise probability of abdominal pain should not depend on nausea
- Nausea and abdominal pain are conditionally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

$$P(X \mid K) = P(X \mid Y, K)$$

Bayes' Rule and Conditional Independence

$$P(cavity, toothache, catch) = P(toothache \mid catch, cavity)\, P(catch \mid cavity)\, P(cavity)$$
$$= P(toothache \mid cavity)\, P(catch \mid cavity)\, P(cavity)$$

- This is an example of a naive Bayes model:



- Total number of parameters is **linear** in n

$$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause) \prod P(Effect_i \mid Cause)$$

Bayesian Networks

- A Bayesian network (also Bayesian Belief 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

Belief Networks - *students and exams*

- Consider a probabilistic model of students and exams. It is reasonable to assume that the random variable Intelligence is independent of Works_hard, given no observations. If you find that a student works hard, it does not tell you anything about their intelligence.

Belief Networks

- The answers to the exam (the variable Answers would depend on whether the student is intelligent and works hard.
- Thus, given Answers, Intelligent would be dependent on Works_hard
  - if you found someone had insightful answers, and did not work hard, your belief that they are intelligent would go up.

Belief Networks

- The grade on the exam (variable Grade should depend on the student's answers, not on the intelligence or whether the student worked hard. Thus *Grade* would be independent of Intelligence given *Answers*. However, if the answers were not observed, *Intelligence* will affect *Grade* (because highly intelligent students would be expected to have different answers than not so intelligent students); thus *Grade* is **dependent** on *Intelligence* given no observations.

Belief Networks

- The notion of conditional independence is used to give a concise representation of many domains.
- A **belief network** is a directed model of conditional dependence among a set of random variables. The conditional independence in a belief network takes in an ordering of the variables, and results in a directed graph.
- To define a belief network on a set of random variables, $\{X_1,\ldots,X_n\}$, first select a total ordering of the variables, say, $X_1,\ldots,X_n$.
- The chain rule is used to decompose a conjunction into conditional probabilities:

Belief Networks

- The chain rule is used to decompose a conjunction into conditional probabilities:

$$P(X_1 = v_1 \wedge X_2 = v_2 \wedge \ldots \wedge X_n = v_n) = \prod_{i=1}^{n} P(X_i = v_i \mid X_1 = v_1 \wedge X_2 = v_2 \wedge \ldots \wedge X_{i-1} = v_{i-1})$$

- in terms of random variables and probability distributions,

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$$

Belief Networks

- Define the **parents** of random variable X_i, written parents(X_i) to be a minimal set of predecessors of X_i in the total ordering such that the other predecessors of Xi are conditionally independent of Xi given parents(Xi)
- Thus X_i **probabilistically depends on** each of its parents, but is independent of its other predecessors. That is, parents(X_i)$\subseteq${X1,…,$_{Xi-1}$} such that

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parent\,(X_i))$$

Belief Networks

- When there are multiple minimal sets of predecessors satisfying this condition, any minimal set may be chosen to be the parents.
  - There can be more than one minimal set only when some of the predecessors are deterministic functions of others.

- Putting the chain rule and the definition of parents together gives:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parent(X_i))$$

Belief Networks

- The probability over all of the variables, $P(X_1,X_2,\ldots,X_n)$ is the **joint probability distribution**.
- A belief network defines a **factorisation** of the joint probability distribution into a product of conditional probabilities.

Belief Networks - Bayesian network

- A belief network or **Bayesian network** consists of
  - a DAG, where each node is labeled by a random variable
  - a domain for each random variable, and
  - a set of conditional probability distributions giving P(X|parents(X)) for each variable X.

- A belief network is acyclic by construction.
- How the chain rule decomposes a conjunction depends on the ordering of the variables.
  - Different orderings can result in different belief networks

Belief Networks - Bayesian network

- Consider the four variables of with the ordering: *Intelligent, Works_hard, Answers, Grade.*
- Consider the variables in order.
  - Intelligent does not have any predecessors in the ordering, so it has no parents, thus parents(Intelligent)={}.
  - Works_hard is independent of Intelligent, and so it too has no parents.
  - Answers depends on both Intelligent and Works_hard,

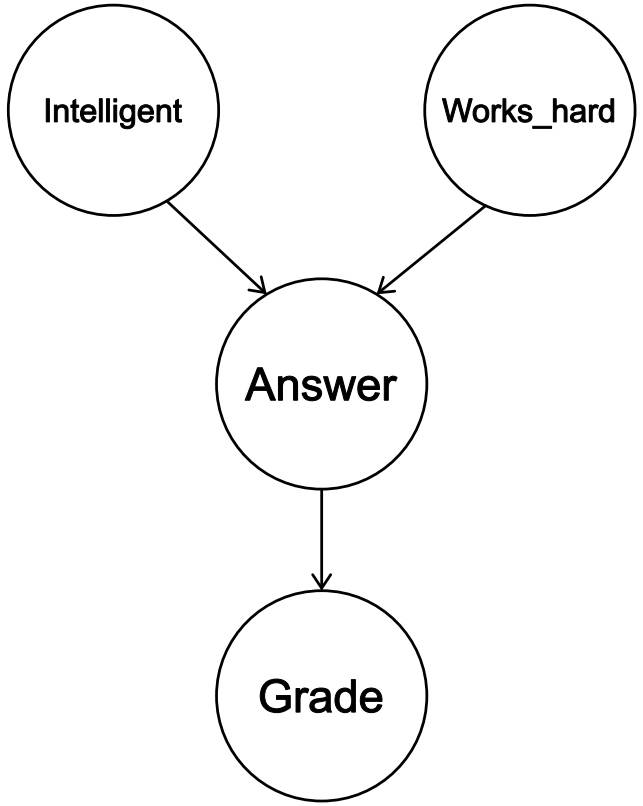
$$parents\,(Answers) = \{Intelligent, Works\_hard\}$$

  - Grade is independent of Intelligent and Works_hard given Answers and so

$$parents\,(Grade) = \{Answers\}\,.$$

Belief Networks - Bayesian network

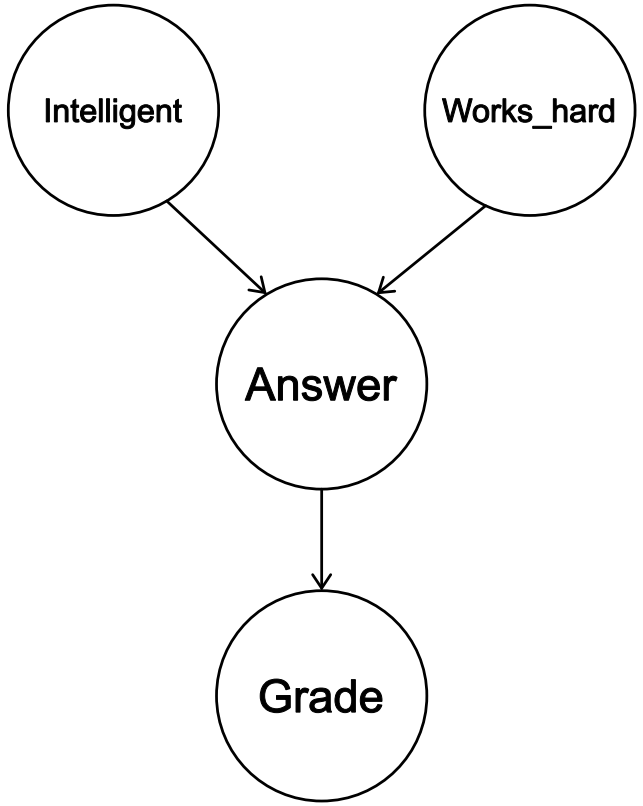
- The corresponding belief network



P (Intelligent, Works_hard, Answers, Grade) =
*P(Intelligent) **
*P(Works_hard) **
*P(Answers | Intelligent, Works_hard) **
P(Grade | Answers)

Belief Networks - Bayesian network

- The corresponding belief network



P (Intelligent, Works_hard, Answers, Grade) =
*P(Intelligent) **
*P(Works_hard) **
*P(Answers | Intelligent, Works_hard) **
P(Grade | Answers)

The domains of the variables are simple,
for example, answers c = {insightful, clear, superficial}

Belief Networks - Bayesian network

- The independence of a belief network, according to the definition of parents, is that each variable is independent of all of the variables that are not descendants of the variable (its non-descendants) given the variable's parents.
- A belief network specifies a joint probability distribution from which arbitrary conditional probabilities can be derived.
  - The most common probabilistic inference task is to compute the **posterior distribution** of a **query variable**, or variables, given some evidence, where the evidence is a conjunction of assignment of values to some of the variables.

Belief Networks - Bayesian network

- Independence and conditional independence relationships among variables can greatly reduce the number of probabilities that need to be specified in order to define the full joint distribution.
- Bayesian networks can represent essentially *any* full joint probability distribution
  - in many cases can do so very concisely.

- The probability over all of the variables, $P(X_1,X_2,\ldots,X_n)$ is the **joint probability distribution**.
- A belief network defines a **factorisation** of the joint probability distribution into a product of conditional probabilities.

Observations and Queries

- Before any observations, the distribution over intelligence is P (Intelligent), which is provided as part of the network.
- We use inference to determine the distribution over grades, P(Grade)

If a grade of A is observed, the posterior distribution of Intelligent is

$$P(Intelligent \mid Grade = A)\,.$$

If it was also observed that Works_hard is false, the posterior distribution of Intelligent is

$$P(Intelligent \mid Grade = A \wedge Works\_hard = false)$$

Although Intelligent and Works_hard are independent given no observations, they are dependent given the grade.
This might explain why some people claim they did not work hard to get a good grade; it increases the probability they are intelligent

Constructing belief networks

- To represent a domain in a belief network, you need to consider:
- What are the relevant variables?
  - What will you observe?
  - What would you like to find out (query)?
  - What other features make the model simpler?
- What values should these variables take?
- What is the relationship between them? This should be expressed in terms of a directed graph, representing how each variable is generated from its predecessors.
- How does the value of each variable depend on its parents? This is expressed in terms of the conditional probabilities.

Bayesian networks

- Independence and conditional independence relationships among variables can greatly reduce the number of probabilities that need to be specified in order to define the full joint distribution.
- Bayesian networks can represent essentially *any* full joint probability distribution
  - in many cases can do so very concisely.

Bayesian Networks - burglar alarm

- 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.

Bayesian Networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
  - Syntax:
    a set of nodes, one per variable
    a directed, acyclic graph (link ≈ “directly influences”)
    a conditional distribution for each node given its parents:

$$P(X_i \mid Parents(X_i))$$

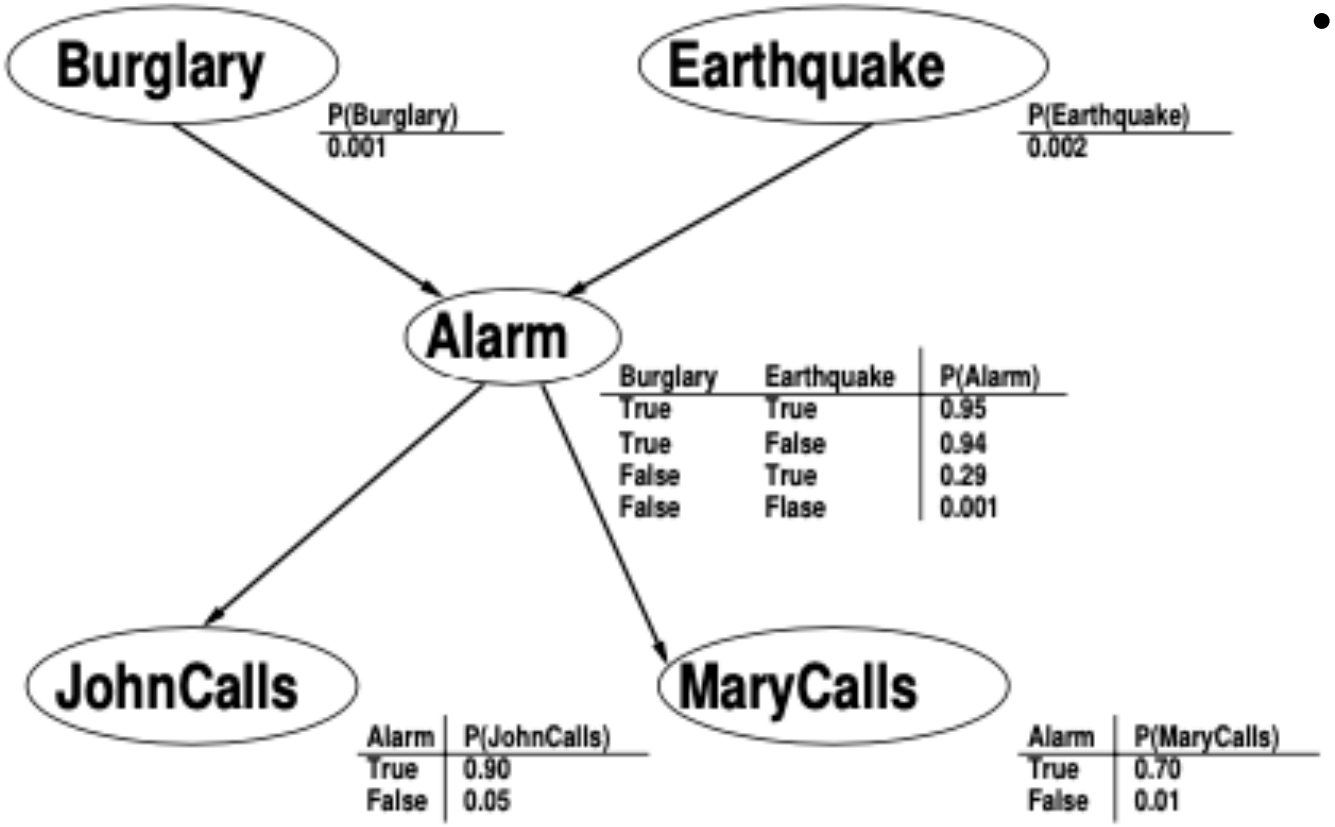
- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example - burglar alarm

- I'm at work, neighbour John calls to say my alarm is ringing, but neighbour Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglar, Earthquake, Alarm, JohnCalls, MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Bayesian Networks



- Example (Pearl, 1988)

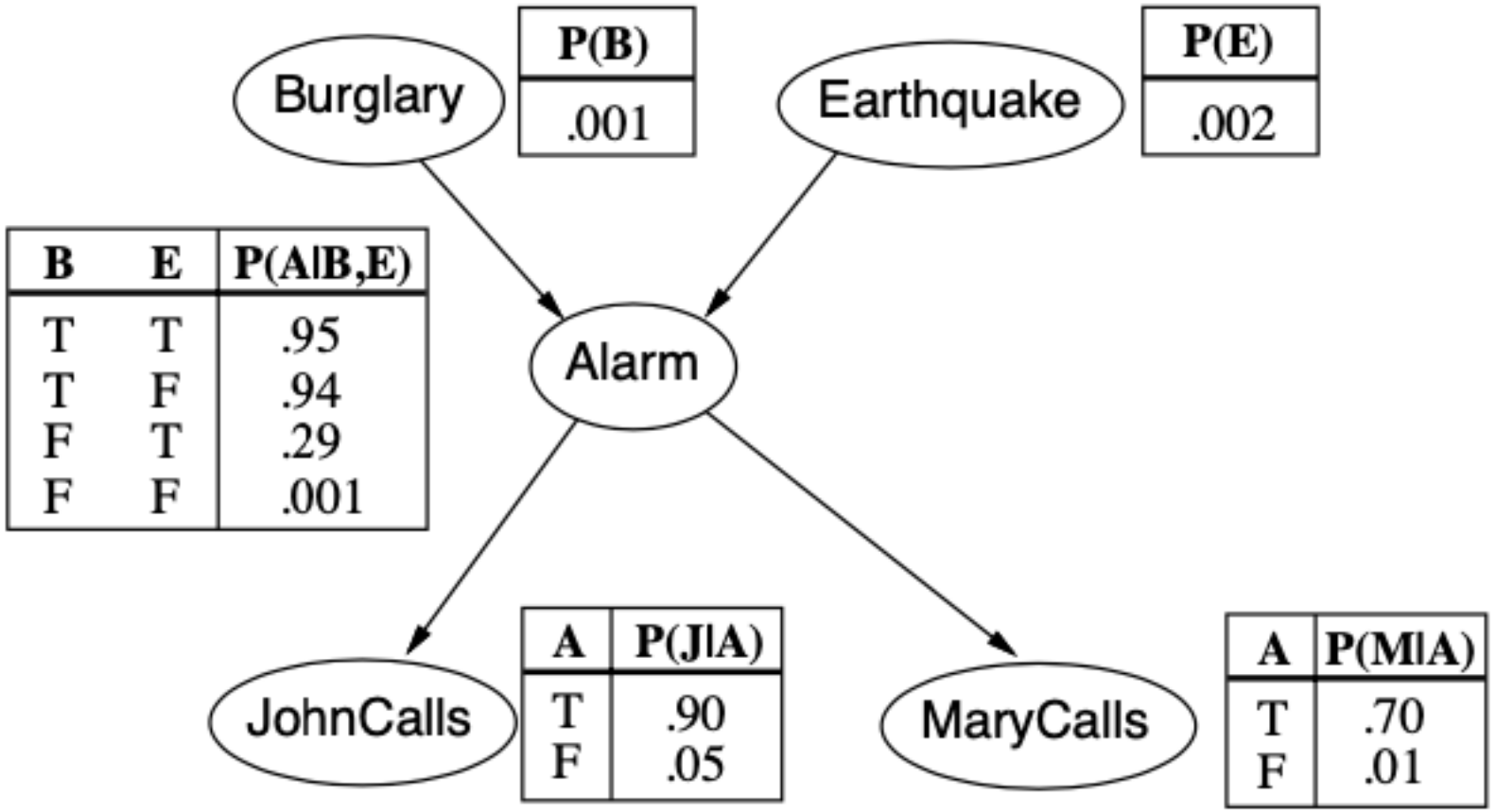
- Probabilities summarise potentially infinite set of possible circumstances

Conditional Probability Table

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

		$\mathbf{P}(Alarm \mid Burglary \wedge Earthquake)$	
Burglary	*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 Networks



Compactness

A CPT for Boolean Xi with k Boolean parents has 2^k rows for the combinations of parent values

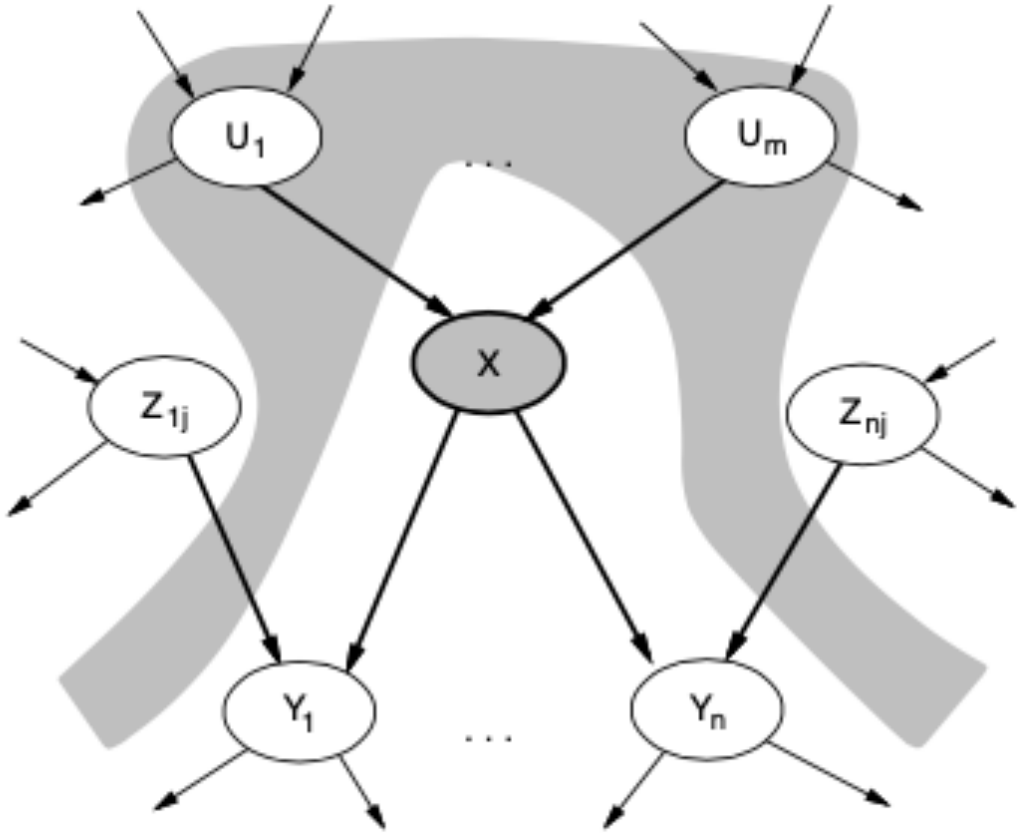
- Each row requires one number p for X_i = true
  - (the number for Xi = false is just 1 – p)
- If each variable has no more than k parents,
- the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

• For burglary net,

$1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Local semantics

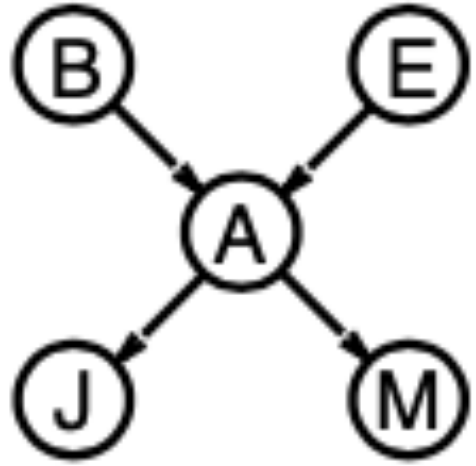
- Local semantics: each node is conditionally independent of its non-descendants given its parents
- Theorem: Local semantics ⇔ global semantics



Global semantics

- Global semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parent(X_i))$$



e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

Semantics of Bayesian Networks

- Bayesian network provides a complete description of the domain
- Joint probability distribution can be determined from the network

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

- For example,

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) =$$
$$P(J|A).P(M|A).P(A|\neg B \wedge \neg E).P(\neg B).P(\neg E) =$$
$$0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628$$

- Bayesian network is a complete and non-redundant representation of domain (and can be far more compact than joint probability distribution)

Calculation using Bayesian Networks

Fact 1: Consider random variable X with parents $Y_1, Y_2, \cdots, Y_n$

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

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

Fact 2 If $Y_1, \cdots, Y_n$ are pairwise disjoint and exhaust all possibilities

$$P(X) = \Sigma P(X \wedge Y_i) = \Sigma P(X|Y_i).P(Y_i)$$

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

e.g. $P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\Sigma P(J \wedge B \wedge e \wedge a \wedge m)}{\Sigma P(j \wedge B \wedge e \wedge a \wedge m)}$ 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

$$P(J \wedge B \wedge E \wedge A \wedge M) = P(J|A).P(B).P(E).P(A|B \wedge E).P(M|A) =$$
$$0.90 \times 0.001 \times 0.002 \times 0.95 \times 0.70 = 0.00000197$$

$$P(J \wedge B \wedge \neg E \wedge A \wedge M) = 0.00591016$$

$$P(J \wedge B \wedge E \wedge \neg A \wedge M) = 5 \times 10^{-11}$$

$$P(J \wedge B \wedge \neg E \wedge \neg A \wedge M) = 2.99 \times 10^{-8}$$

$$P(J \wedge B \wedge E \wedge A \wedge \neg M) = 0.000000513$$

$$P(J \wedge B \wedge \neg E \wedge A \wedge \neg M) = 0.000253292$$

$$P(J \wedge B \wedge E \wedge \neg A \wedge \neg M) = 4.95 \times 10^{-9}$$

$$P(J \wedge B \wedge \neg E \wedge \neg A \wedge \neg M) = 2.96406 \times 10^{-6}$$

Calculation using Bayesian Networks

$$P(\neg J \wedge B \wedge E \wedge A \wedge M) = 0.000000133$$

$$P(\neg J \wedge B \wedge \neg E \wedge A \wedge M) = 6.56684 \times 10^{-5}$$

$$P(\neg J \wedge B \wedge E \wedge \neg A \wedge M) = 9.5 \times 10^{-10}$$

$$P(\neg J \wedge B \wedge \neg E \wedge \neg A \wedge M) = 5.6886 \times 10^{-7}$$

$$P(\neg J \wedge B \wedge E \wedge A \wedge \neg M) = 0.000000057$$

$$P(\neg J \wedge B \wedge \neg E \wedge A \wedge \neg M) = 2.81436 \times 10^{-5}$$

$$P(\neg J \wedge B \wedge E \wedge \neg A \wedge \neg M) = 9.405 \times 10^{-8}$$

$$P(\neg J \wedge B \wedge \neg E \wedge \neg A \wedge \neg M) = 5.63171 \times 10^{-5}$$

Calculation using Bayesian Networks

- Can often simplify calculation without using full joint probabilities – but not always

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

$$\text{Therefore } P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\Sigma P(J \wedge B \wedge e \wedge a \wedge m)}{\Sigma P(j \wedge B \wedge e \wedge a \wedge m)} = \frac{0.00849017}{0.001}$$

$$P(J|B) = 0.849017$$

Inference in Bayesian Networks

- Diagnostic Inference From effects to causes

$$P(Burglary|JohnCalls) = 0.016$$

- Causal Inference From causes to effects

$$P(JohnCalls|Burglary) = 0.85; P(MaryCalls|Burglary) = 0.67$$

- Inter-causal Inference Explaining away

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

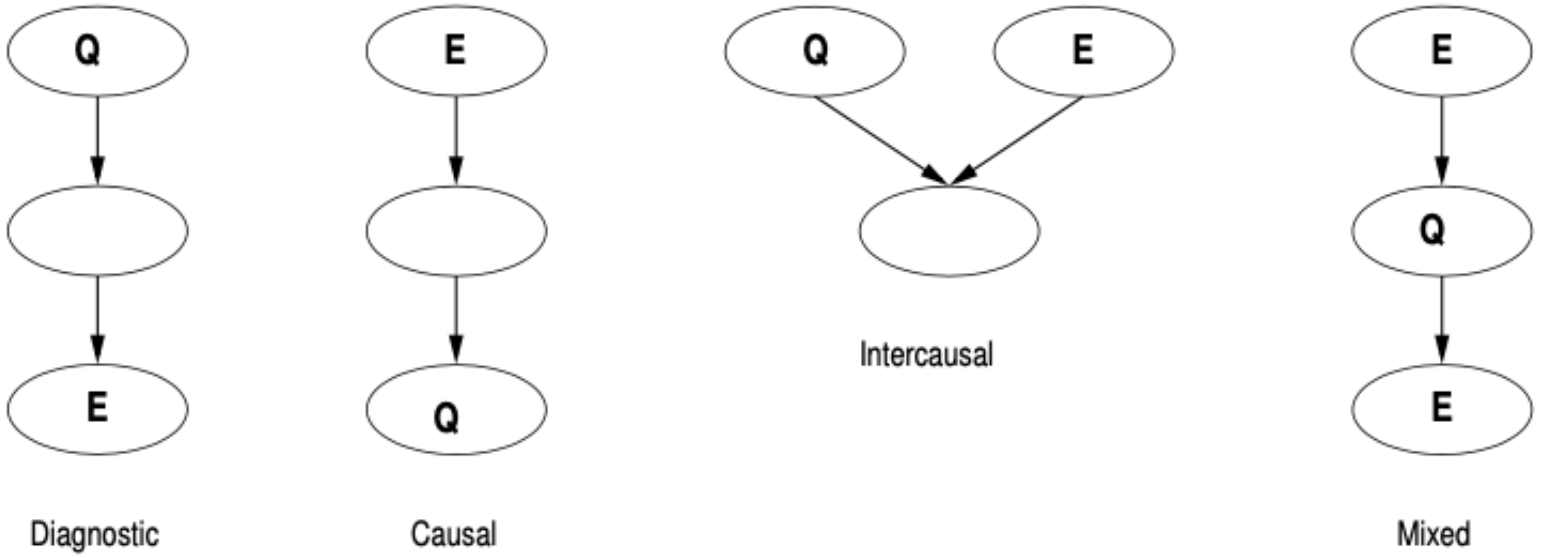
- Inference Combinations of the patterns above

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

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

Inference in Bayesian Networks

- Q- Query; E= evidence



Example – Diagnostic Inference

$$P(Earthquake|Alarm)$$

$$P(E|A) = \frac{P(A|E).P(E)}{P(A)}$$

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

$$= \frac{0.95\times 0.001\times 0.002+0.29\times 0.999\times 0.002}{P(A)} = \frac{5.8132\times 10^{-4}}{P(A)}$$

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

And $P(A|B\wedge\neg E).P(B).P(\neg E) + P(A|\neg B\wedge\neg E).P(\neg B).P(\neg E)$
$= 0.94\times 0.001\times 0.998 + 0.001\times 0.999\times 0.998 = 0.001935122$

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: $P(X\wedge Y) = P(X).P(Y)$ if X, Y are conditionally independent

Example – Causal Inference

$$P(JohnCalls|Burglary)$$

$$\begin{aligned} P(J|B) &= P(J|A \wedge B).P(A|B) + P(J|\neg A \wedge B).P(\neg A|B) \\ &= P(J|A).P(A|B) + P(J|\neg A).P(\neg A|B) \\ &= P(J|A).P(A|B) + P(J|\neg A).(1 - P(A|B)) \end{aligned}$$

$$\begin{aligned} \text{Now } P(A|B) &= P(A|B \wedge E).P(E|B) + P(A|B \wedge \neg E).P(\neg E|B) \\ &= P(A|B \wedge E).P(E) + P(A|B \wedge \neg E).P(\neg E) \\ &= 0.95 \times 0.002 + 0.94 \times 0.998 = 0.94002 \end{aligned}$$

$$\text{Therefore } P(J|B) = 0.90 \times 0.94002 + 0.05 \times 0.05998 = 0.849017$$

Fact 3: $P(X|Z) = P(X|Y \wedge Z).P(Y|Z) + P(X|\neg Y \wedge Z).P(\neg Y|Z)$, since $X \wedge Z \equiv (X \wedge Y \wedge Z) \vee (X \wedge \neg Y \wedge Z)$ (conditional version of Fact 2)

Belief network summary

- A belief network is a directed acyclic graph (DAG) where nodes are random variables.
- The **parents** of a node n are those variables on which n directly depends.
- A belief network is automatically acyclic by construction.
- A belief network is a graphical representation of **dependence** and **independence**:
  - A variable is independent of its non-descendants given its parents.

Constructing belief networks

- To represent a domain in a belief network, you need to consider:
- What are the relevant variables?
  - What will you observe?
  - What would you like to find out (query)?
  - What other features make the model simpler?
- What values should these variables take?
- What is the relationship between them? This should be expressed in terms of a directed graph, representing how each variable is generated from its predecessors.
- How does the value of each variable depend on its parents? This is expressed in terms of the conditional probabilities.

Conclusion

- Due to noise or **uncertainty** it is useful to reason with **probabilities**
- Calculating with joint probability distribution difficult due to the large number of values
  - **Joint probability** distribution specifies probability of every **atomic event**
- Use of Bayes' Rule and independence assumptions simplifies reasoning
  - Queries can be answered by summing over atomic events
  - For nontrivial domains, we must find a way to reduce the joint size
- **Independence** and **conditional independence** provide the tools

Conclusion

- A Bayesian network specifies a **full joint distribution**; each joint entry is defined as the product of the corresponding entries in the local conditional distributions.
  - A Bayesian network is often exponentially smaller than an explicitly enumerated joint distribution.
- Bayesian networks allow compact representation of probabilities and efficient reasoning with probabilities
- Elegant recursive algorithms can be given to automate the process of inference in Bayesian networks

References

- Poole &Mackworth, Artificial Intelligence: Foundations of Computational Agents, Chapter 8
- Russell & Norvig, *Artificial Intelligence: a Modern Approach*, Chapter 13 & 15.