Uncertainty

v 1.1

Dr Armin Chitizadeh

Inspired by the COMP9414 lectures by
Dr Alan Blair
Associate Professor Wayne Wobcke

UNSW SYDNEY

Corrections since the previous version

- The table on slide 19 is now corrected
- Superscripts in formulas are now compiled and demonstrated correctly in slides 44 and 45

Outline

- Uncertainty and Probability
- What shall we do when we receive new information
- What is Bayesian Networks and why do we need it
- Some examples

Uncertainty

Photo by Dan Freeman on Unsplash

Uncertainty

- In many situation, an AI agent has to choose an action based on incomplete information.
  - Incomplete Information: agent may not have the complete theory for the domain
  - Imperfect Information: agent may not have enough information about the domain
  - Noise: information agent does have may be unreliable
  - Non-determinism: environment itself may be stochastic
  - Multi-agent world: other agents act on their own interest

Planning under Uncertainty

Let action A_t = leave for the airport t minutes before the flight

Will A_t get me there on time? Problem:

- Partial observability, noisy sensors
- Uncertainty in action outcomes (flat tyre, etc.)
- Immense complexity of modelling and predicting traffic

Hence a purely logical approach assumes there is no uncertainty

Methods for handling Uncertainty

Probability

Probability gives a way of summarizing the uncertainty

- Given the available evidence,
  - Leaving 30 minutes in advance will get me there on time with probability 0.04
  - Leaving 90 minutes in advance will get me there on time with probability 0.75
  - Leaving 120 minutes in advance will get me there on time with probability 0.95
  - Leaving 1440 minutes in advance will get me there on time with probability 0.999

Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

Bell DF. Pascal: Casuistry, probability, uncertainty. Journal of Medieval and Early Modern Studies. 1998;28(1):37.

Random Variables

- **E.g. Weather**

•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

$$\langle x_1, x_2, \ldots, x_n \rangle$$

- Probabilities of all possible values $\mathbf{P}(Weather) = \langle 0.8, 0.1, 0.09, 0.01 \rangle$ is a probability distribution

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

Photo by Andy Henderson on Unsplash

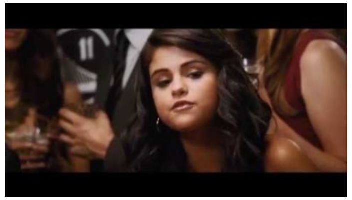
Objective View

Probabilities are real aspects of the world — objective

Photo by Luke Chesser on Unsplash

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; different agents may assign a different probability — subjective



https://www.youtube.com/watch?v=AUM59Eh6vTw

Sample Space and Events

- Flip a coin three times
- The possible outcomes are

TTT TTH THT THH

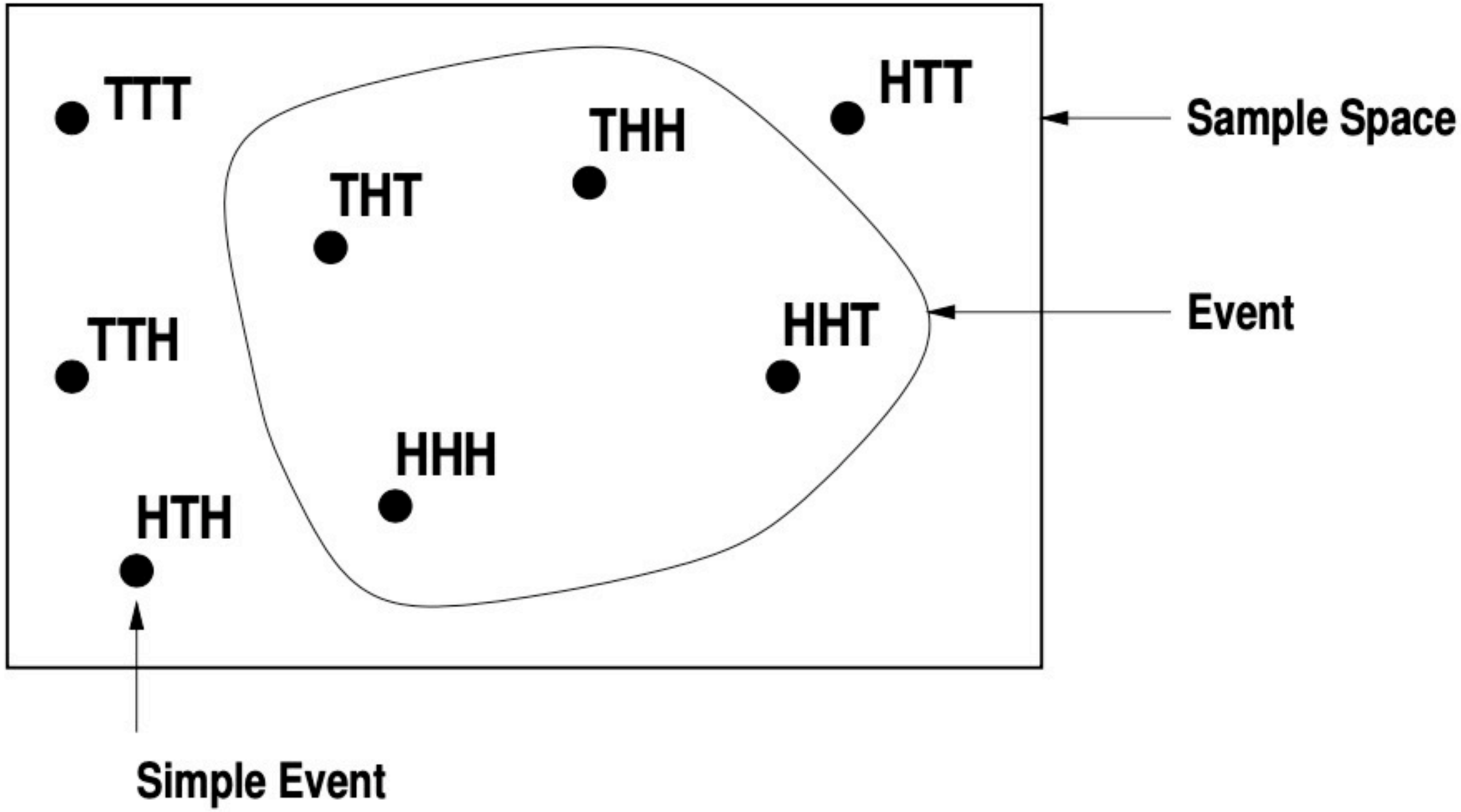
HTT HTH HHT HHH

- Set of all possible outcomes

$$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$

- Sample space is the set of all possible outcomes
- Any subset of the sample space is known as an event
- Any singleton subset of the sample space is know as sample point/possible world/atomic event/ simple event

Sample Space and Events

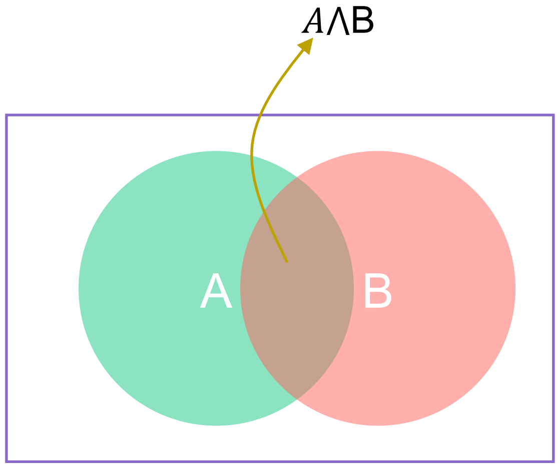


Prior Probability

- Probability before any new data is collected
- $P(A)$ is the prior or unconditional probability that an event A occurs
- For example, $P(Appendicitis = False)$=0.3
  - Other way to represent can be: $P(\neg Appendicitis)$=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

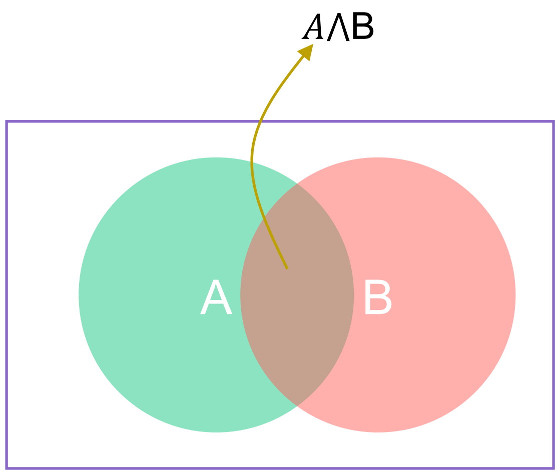
Axioms of Probability

- $0 \leq P(A) \leq 1$
  - All probabilities are between 0 and 1
- $P(True) = 1 \qquad P(False) = 0$
  - Valid propositions have probability 1
  - Unsatisfiable propositions have probability 0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - Can determine probabilities of all other propositions



Axioms of Probability

- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - Can determine probabilities of all other propositions



Joint probability

Probability of two atomic events co-occurring

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

Probabilities in table come from observation

Example: Tooth Decay

- 20% of people have a cavity in one of their teeth which needs a filling.

$$P(cavity) = 0.2$$

- Dentist catches a hole in a teeth of 34% of the people

$$P(\text{catch}) = 0.34$$

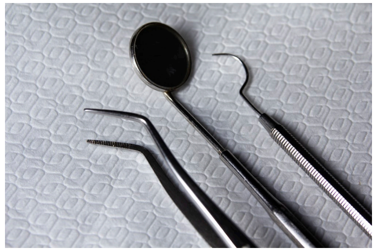


Photo by Dan Freeman on Unsplash

- 20% of people have toothache.

$$P(\text{tootache}) = 0.20$$

Joint Probability Distribution

Assume some underlying joint probably distribution over three random variables:

- Toothache, Cavity and Catch:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**0.064**	**.144**	**.576**

Note that the sum of the entries in the table is 1.0.

For any proposition, sum of simple events where it is true:

Inference by Enumeration

Start with the joint distribution

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**0.064**	**.144**	**.576**

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

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by Enumeration

Start with the joint distribution

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**0.064**	**.144**	**.576**

$$P(cavity \vee toothache) = ?$$

Inference by Enumeration

Start with the joint distribution

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	0.064	.144	.576

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

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

Outline

- Uncertainty and Probability
- What shall we do when we receive new information
- What is Bayesian Networks and why do we need it
- Some examples

Conditional Probability

Photo by Lina Kivaka on Pxeels

Photo by Mart Production on Pexels

Example: Tooth Decay

- Feeling a toothache, you think you have a cavity, perhaps as high as 60%.
- The conditional probability of cavity, given toothache, is 0.6, written as:

$$P(\texttt{cavity}|\texttt{toothache}) = 0.6$$

- Dentist's check will increase probability of a cavity, ca

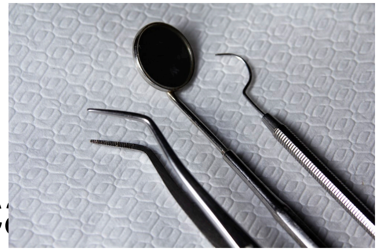


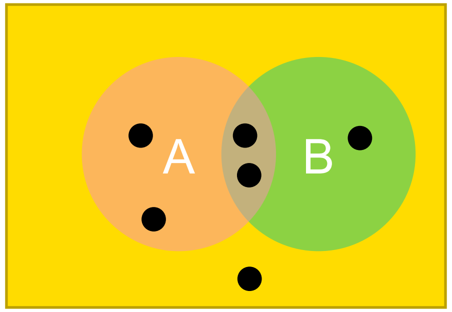
Photo by Dan Freeman on Unsplash

Conditional Probability

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

  e.g. $P(Appendicitis \mid AbdominalPain) = 0.75$

Conditional Probability

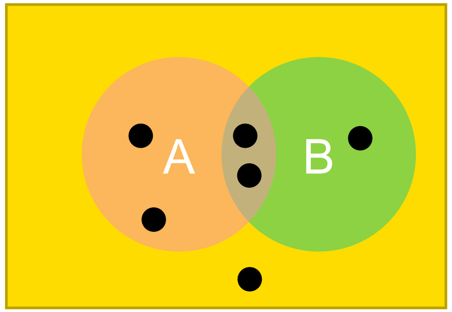


- Need to update probabilities based on new information
- Use conditional or posterior probability
- $P(A|B)$ is the probability of A given we know B

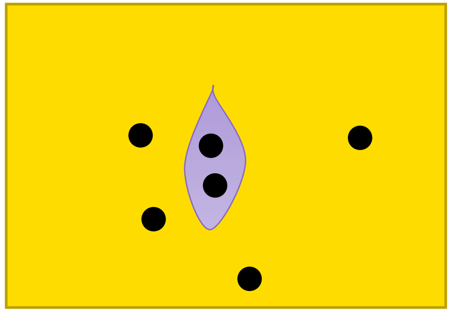
  e.g. $P(Appendicitis \mid AbdominalPain) = 0.75$
- Definition: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$. provided $P(B) > 0$
- Product Rule: $P(A \wedge B) = P(A|B)P(B)$

Conditional Probability

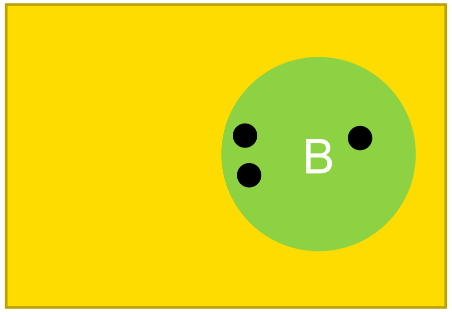
$P(A|B) = \frac{P(A \wedge B)}{P(B)}$. provided $P(B) > 0$



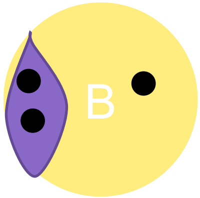
$P(A \wedge B) = 2/6$



$P(B) = 3/6$



$P(A|B) = 2/3$



$$\frac{P(A \wedge B)}{P(B)} = \frac{2/6}{3/6} = 2/3$$

Conditional Probability by Enumeration

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**0.064**	**.144**	**.576**

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.0064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Conditional Probability

Consider two random variable a and b, with $P(b) \neq 0$

- the conditional probability of a given b is

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

Alternative formulation:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

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

$$\begin{aligned} P(X_n, \ldots, X_1) &= P(X_n | X_{n-1}, \ldots, X_1) P(X_{n-1}, \ldots, X_1) \\ &= P(X_n | X_{n-1}, \ldots, X_1) P(X_{n-1} | X_{n-2}, \ldots, X_1) \\ &= \ldots = \prod_{i=1}^{n} P(X_i | X_{i-1}, \ldots, X_1) \end{aligned}$$

Bayes' Rule

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

- Deriving Bayes' Rule:
  $P(A \wedge B) = P(A|B)P(B)$ (Definition)
  $P(B \wedge A) = P(B|A)P(A)$ (Definition)
  So $P(A|B)P(B) = P(B|A)P(A)$ since $P(A \wedge B) = P(B \wedge A)$
  Hense: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ if $P(A) \neq 0$
  $P(A)$

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

Using Bayes' Rule

- Suppose there are two conditional probabilities for appendicitis

$$P(Appendicitis|AbdominalPain) = 0.8$$
$$P(Appendicitis|Nausea) = 0.1$$

- $P(Appendicitis|AbdominalPain \wedge Nausea)$

$$= \frac{P(AbdominalPain \wedge Nausea|Appendicitis).P(Appendicitis)}{P(AbdominalPain \wedge Nausea)}$$

- Need to know $P(AbdominalPain \wedge Nausea|Appendicitis)$
- With many symptoms that is a daunting task …

Outline

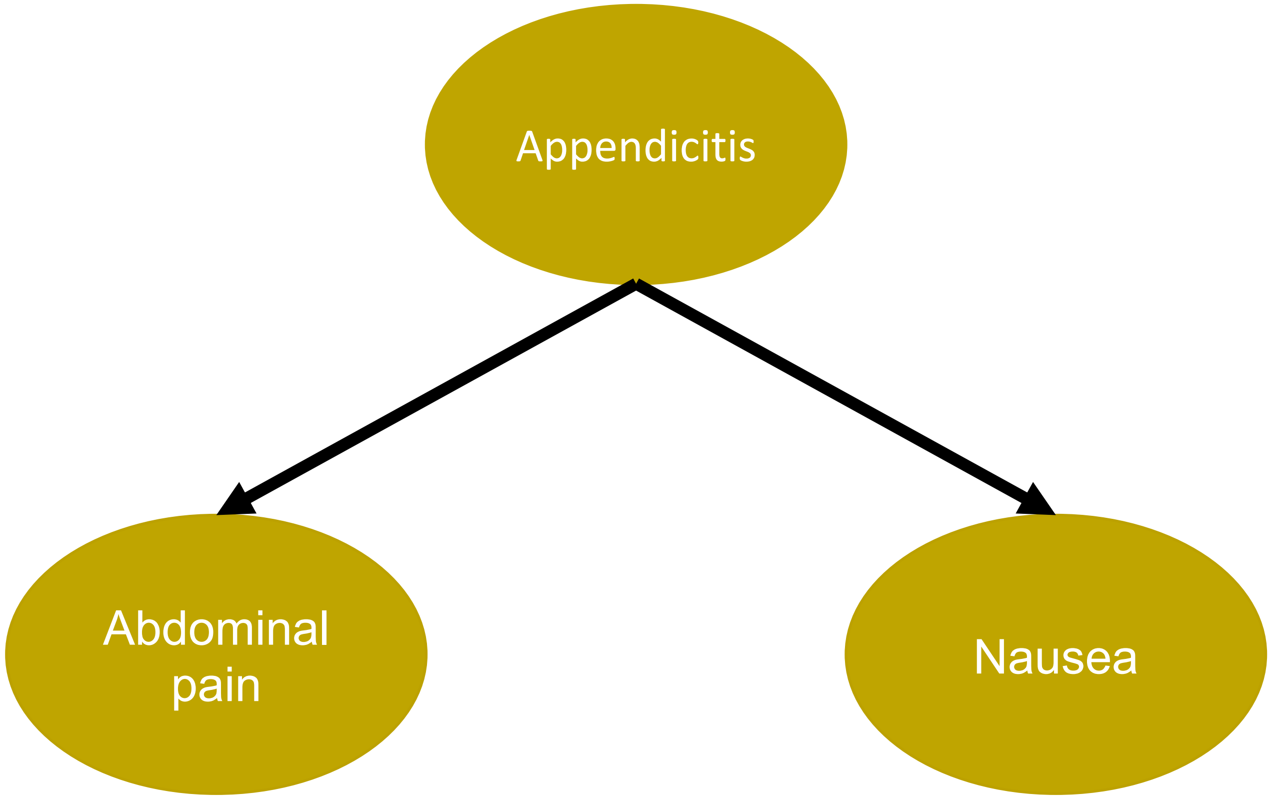
- Uncertainty and Probability
- What shall we do when we receive new information
- What is Bayesian Networks and why do we need it
- Some examples

Why shall we use Bayesian Networks

With many symptoms that is a daunting task …

Photo by Pixabay on Pexels

Conditional Independence



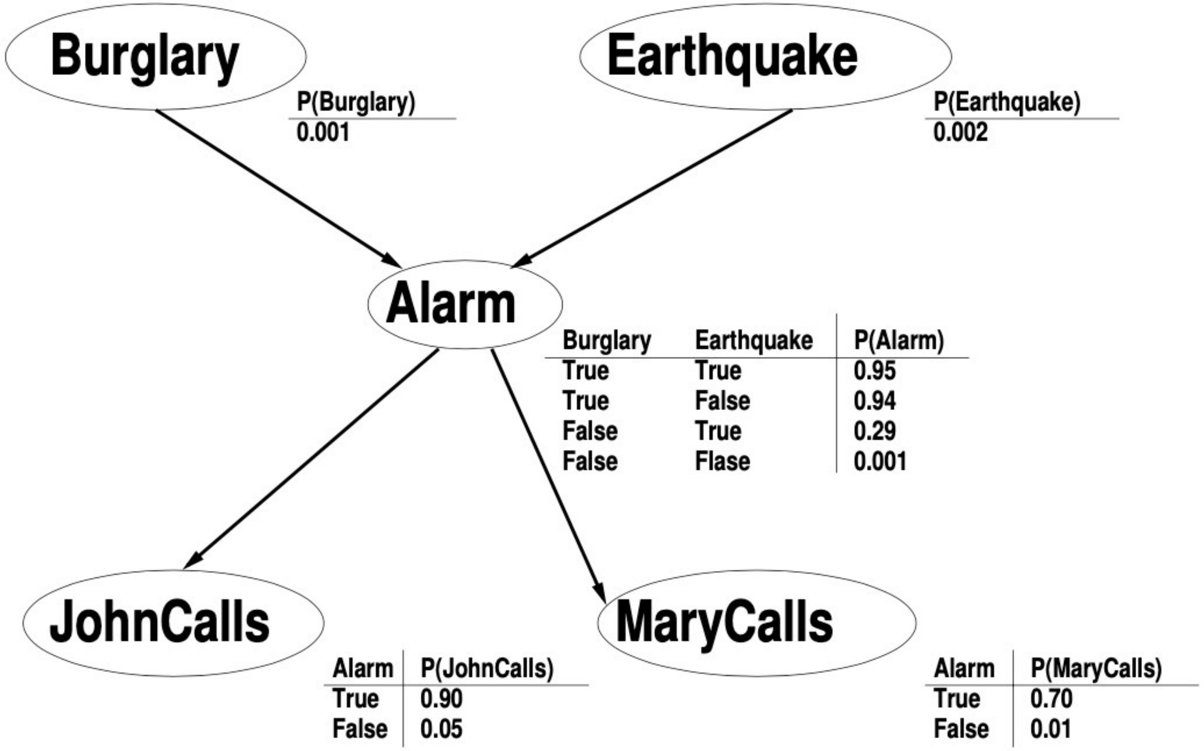
Conditional Independence

- 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*

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

Bayesian Networks

- Example (Pearl, 1988)



Probabilities summarize potentially infinite set of possible circumstances

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 non descendants given its parents

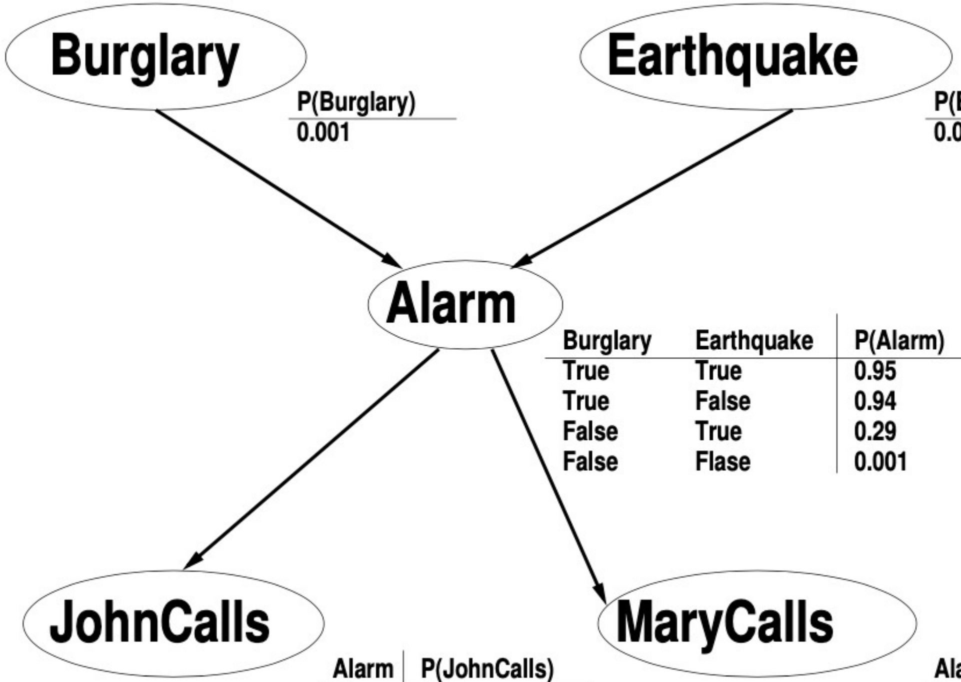
Bayesian Networks

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

Bayesian Networks

- Example (Pearl, 1988)



P(Burglary)
0.001

P(Earthquake)
0.002

Burglary	Earthquake	P(Alarm)
True	True	0.95
True	False	0.94
False	True	0.29
False	Flase	0.001

Alarm	P(JohnCalls)
True	0.90
False	0.05

Alarm	P(MaryCalls)
True	0.70
False	0.01

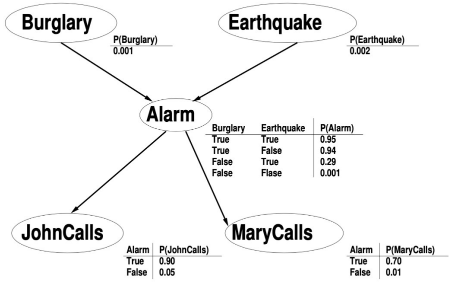
Probabilities summarize potentially infinite set of possible circumstances

Conditional Probability Table

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

		$P(Alarm \mid Burglary \wedge Earthquake)$
$Burglary$	$Earthquake$	True
True	True	0.950
True	False	0.940
False	True	0.290
False	False	0.001

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 domain (and can be far more compact than joint probability distribution)

Semantics of Bayesian Networks

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

$$P(X_1 \wedge X_2 \wedge \cdots \wedge X_n) = P(X_1).P(X_2|X_1).P(X_3|X_1 \wedge X_2). \cdots .P(X_n|X_1 \wedge X_2 \wedge \cdots \wedge X_{n-1})$$

- Now, order the variables $X_1, X_2, \cdots, X_2$ in a Bayesian network so that a variable comes after its parents – let π_{X_1} be the tuple of parents of variable *Xi* (this is a complex random variable)

- Using the chain rule,

$$P(X_1 \wedge X_2 \wedge \cdots \wedge X_n) = P(X_1).P(X_2|X_1).P(X_3|X_1 \wedge X_2).\cdots .P(Xn|X_1 \wedge X_2 \wedge \cdots \wedge Xn_{-1})$$

Semantics of Bayesian Networks

let π_{Xi} be the tuple of parents of variable X_i

Each $P(Xi \mid X_1 \wedge X_2 \wedge \cdots \wedge Xi_{-1})$ has the property that it is not conditioned on a descendant of X_i (given ordering of variables in Bayesian network)

Therefore, by conditional independence

> $P(Xi|X_1 \wedge X_2 \wedge \cdots \wedge Xi - 1) = P(Xi|\pi_{Xi})$

Rewriting gives the chain rule

> $P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{b} P(Xi|\pi_{Xi})$

Calculation using Bayesian Networks

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

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

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

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

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

> e.g. Type equation here. $P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\Sigma P(J \wedge B \wedge e \wedge a \wedge m)}{\Sigma P(j \wedge B \wedge e \wedge a \wedge m)}$ = where *j* ranges over *J*,¬*J*, e over E, ,¬*E, a over A, ¬A and m over M,* ¬M

Calculating 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

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

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

- $P(J|B) = 0.849017$

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

Inference in Bayesian Networks

Diagnostic Inference From effects to causes

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

Causal Inference From causes to effects

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

Intercausal Inference Explaining away

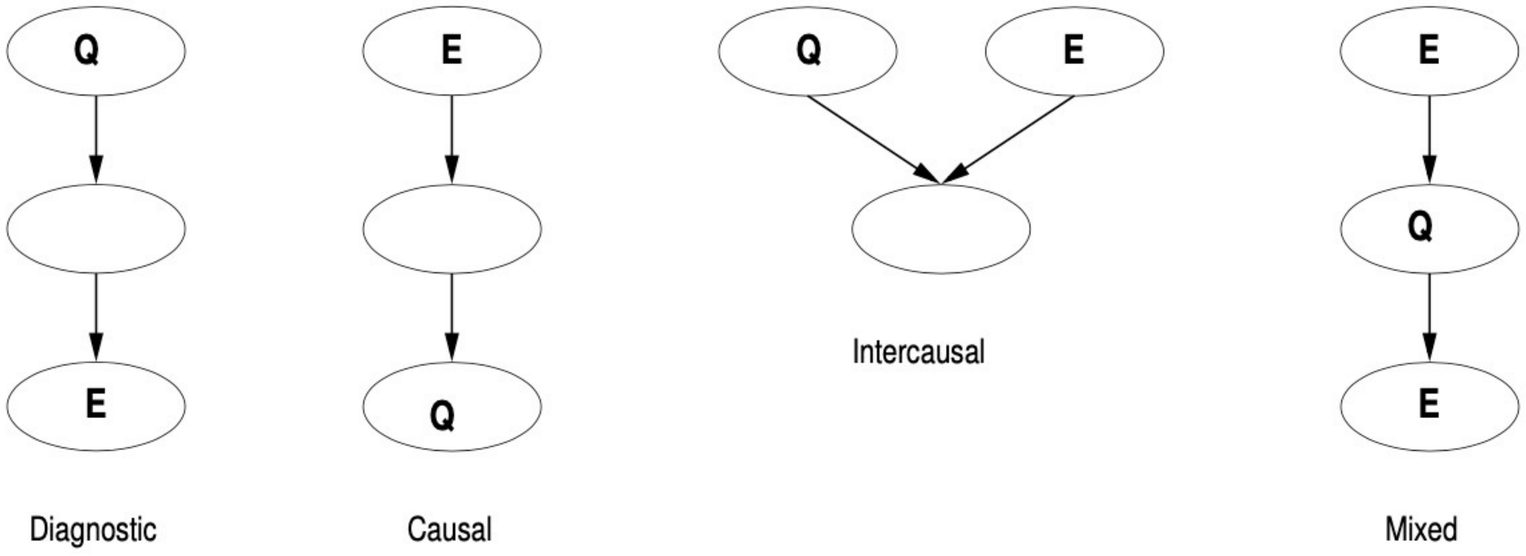
$P(Burglary|Alarm) = 0.3736$ but adding evidence, $P(Burglary|Alarm \wedge Earthquake) = 0.003$; despite the fact that burglaries and earthquakes are independent, the presence of one makes the other much less likely

Mixed 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

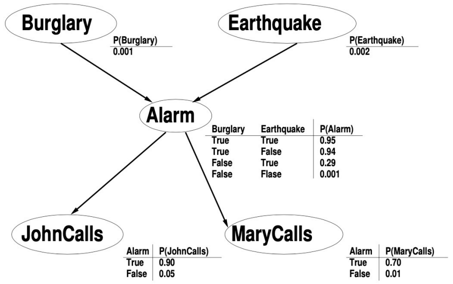


Q = query; E = evidence

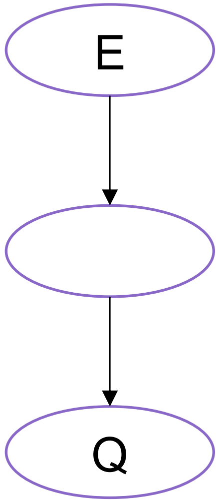
Outline

- Uncertainty and Probability
- What shall we do when we receive new information
- What is Bayesian Networks and why do we need it
- Some examples

Example – Causal Inference

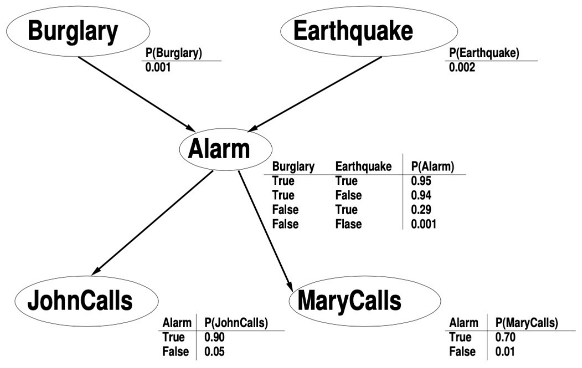


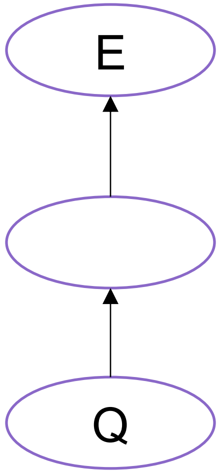
- $P(JohnCalls|Burglary)$
- $P(J|B) = P(J|A \wedge B).P(A|B) + P(J|\neg A \wedge B).P(\neg A|B)$
  $= P(J|A).P(A|B) + P(J|\neg A).P(\neg A|B)$
  $= P(J|A).P(A|B) + P(J|\neg A).(1 - P(A|B))$
- $Now\ P(A|B) = P(A|B \wedge E).P(E|B) + P(A|B \wedge \neg E).P(\neg E|B)$
  $= P(A|B \wedge E).P(E) + P(A|B \wedge \neg E).P(\neg E)$
  $= 0.95 \times 0.002 + 0.94 \times 0.998 = 0.94002$
- $Therefore\ P(J|B) = 0.90{\times}0.94002 + 0.05{\times}0.05998 = 0.849017$
- $Fact\ 3\text{:}\ P(X|Z) = P(X|Y \wedge Z).P(Y|Z) + P(X|\neg Y \wedge Z).P(\neg Y|Z), since$
- $X \wedge Z \equiv (X \wedge Y \wedge Z) \vee (X \wedge \neg Y \wedge Z)\ (conditional\ version\ of\ Fact\ 2)$



Example – Diagnostic Inference

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





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
- Use of Bayes’ Rule and independence assumptions simplifies reasoning
- 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