

Uncertainty

Chapter 13^{Russel}

"So far as the laws of Mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality."

Albert Einstein

(1879-1955)

"If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties."

Francis Bacon

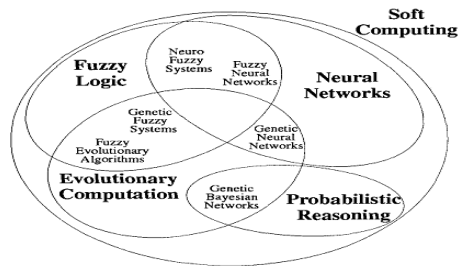
(1561-1626)

Uncertainty

Uncertainty is unavoidable when dealing with data. It may be due to

- The errors in measurements,
- limitations of measuring tools, or
- imprecise definition of linguistic variables

Each component of soft computing is complementary to another



Uncertainty

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Uncertainty

Hence a purely logical approach either

1. risks falsehood: “ A_{25} will get me there on time”, or
2. leads to conclusions that are too weak for decision making:

“ A_{25} 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.”

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport ...)

Methods for handling uncertainty

Default or **nonmonotonic** logic:

Assume my car does not have a flat tire

Assume A_{25} works unless contradicted by evidence

Issues: What assumptions are reasonable? How to handle contradiction?

Rules with fudge factors:

$A_{25} \rightarrow_{0.3}$ get there on time

Probability

Model agent's degree of belief

Given the available evidence,

A_{25} will get me there on time with probability 0.04

Probability

Probabilistic assertions **summarize** effects of

laziness: failure to enumerate exceptions, qualifications, etc.

ignorance: lack of relevant facts, initial conditions, etc.

Subjective probability:

Probabilities relate propositions to agent's own state of knowledge

e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are **not** assertions about the world

Probabilities of propositions change with new evidence:

e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Making decisions under uncertainty

Suppose I believe the following:

$$P(A_{25} \text{ gets me there on time} \mid \dots) = 0.04$$

$$P(A_{90} \text{ gets me there on time} \mid \dots) = 0.70$$

$$P(A_{120} \text{ gets me there on time} \mid \dots) = 0.95$$

$$P(A_{1440} \text{ gets me there on time} \mid \dots) = 0.9999$$

Which action to choose?

Depends on my **preferences** for missing flight vs. time spent waiting, etc.

Utility theory is used to represent and infer preferences

Decision theory = probability theory + utility theory

Expected utility of action a in state s

$$= \sum_{\text{outcome in Results}(s,a)} P(\text{outcome}) * \text{Utility}(\text{outcome})$$

- A rational agent acts to maximize expected utility

Making decisions under uncertainty (Example)

- Suppose I believe the following:
 - $P(\text{A25 gets me there on time} \mid \dots) = 0.04$
 - $P(\text{A90 gets me there on time} \mid \dots) = 0.70$
 - $P(\text{A120 gets me there on time} \mid \dots) = 0.95$
 - $P(\text{A1440 gets me there on time} \mid \dots) = 0.9999$
 - Utility(on time) = \$1,000
 - Utility(not on time) = -\$10,000
- **Expected utility** of action a in state s
 - = $\sum_{\text{outcome} \in \text{Results}(s,a)} P(\text{outcome}) * \text{Utility}(\text{outcome})$
 - $E(\text{Utility}(\text{A25})) = 0.04 * \$1,000 + 0.96 * (-\$10,000) = -\$9,560$
 - $E(\text{Utility}(\text{A90})) = 0.7 * \$1,000 + 0.3 * (-\$10,000) = -\$2,300$
 - $E(\text{Utility}(\text{A120})) = 0.95 * \$1,000 + 0.05 * (-\$10,000) = \450
 - $E(\text{Utility}(\text{A1440})) = 0.9999 * \$1,000 + 0.0001 * (-\$10,000) = \998.90

History of Probability in AI

- Early AI (1950's and 1960's)
 - Attempts to solve AI problems using probability met with mixed success
- Logical AI (1970's, 80's)
 - Abandoned probabilistic approaches
 - Focused on logic-based representations
 - Problem: Pure logic is “brittle” when applied to real-world problems.
- Probabilistic AI (1990's-present)
 - Judea Pearl invents Bayesian networks in 1988
 - Development of machine learning techniques to learn such models from data
 - Realization that approximate probability models are tractable and useful
 - Probabilistic techniques now widely used in vision, speech recognition, robotics, language modeling, game-playing, etc

Syntax

- Basic element: **random variable**
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- **Boolean** random variables
e.g., *Cavity* (= do I have a cavity?)
- **Discrete** random variables
e.g., *Weather* is one of $\langle \text{sunny, rainy, cloudy, snow} \rangle$
- Domain values must be exhaustive and mutually exclusive
- Elementary proposition is an assignment of a value to a random variable:
e.g., *Weather = sunny; Cavity = false* (abbreviated as $\neg \text{cavity}$)
- Complex propositions formed from elementary propositions and standard logical connectives :
e.g., *Weather = sunny \vee Cavity = false*

Probability

- $P(a)$ is the probability of proposition “a”
 - E.g., $P(\text{it will rain in London tomorrow})$
 - The proposition a is actually true or false in the real-world
 - $P(a)$ = “prior” or marginal or unconditional probability
 - Assumes no other information is available
- Axioms:
 - $0 \leq P(a) \leq 1$
 - $P(\text{NOT}(a)) = 1 - P(a)$
 - $P(\text{true}) = 1$
 - $P(\text{false}) = 0$
 - $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$
- Any agent that holds degrees of beliefs that contradict these axioms will act sub-optimally in some cases
 - e.g., de Finetti proved that there will be some combination of bets that forces such an unhappy agent to lose money every time.
- **Rational agents cannot violate probability theory.**

Probability and Logic

- Probability can be viewed as a generalization of propositional logic
- $P(a)$:
 - a is any sentence in propositional logic
 - Belief of agent in a is no longer restricted to *true*, *false*, *unknown*
 - $P(a)$ can range from 0 to 1
 - $P(a) = 0$, and $P(a) = 1$ are special cases
 - So logic can be viewed as a special case of probability

Axioms of probability

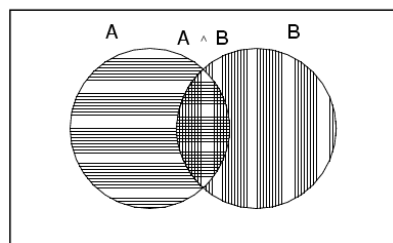
For any propositions A, B

$$0 \leq P(A) \leq 1$$

$$P(\text{true}) = 1 \text{ and } P(\text{false}) = 0$$

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

True



Conditional Probability

- $P(a|b)$ is the conditional probability of proposition a , conditioned on knowing that b is true,
 - E.g., $P(\text{rain in London tomorrow} \mid \text{raining in London today})$
 - $P(a|b)$ is a “posterior” or conditional probability
 - The updated probability that a is true, now that we know b
 - $P(a|b) = P(a \text{ AND } b) / P(b)$
 - Syntax: $P(a \mid b)$ is the probability of a given that b is true
 - a and b can be any propositional sentences
 - e.g., $p(\text{John wins OR Mary wins} \mid \text{Bob wins AND Jack loses})$
- $P(a|b)$ obeys the same rules as probabilities,
 - E.g., $P(a \mid b) + P(\text{NOT}(a) \mid b) = 1$
 - All probabilities in effect are conditional probabilities
 - E.g., $P(a) = P(a \mid \text{our background knowledge})$

Prior probability

Prior or unconditional probabilities of propositions

e.g., $P(\text{Cavity} = \text{true}) = 0.1$ and $P(\text{Weather} = \text{sunny}) = 0.72$ correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

$\mathbf{P}(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables

$\mathbf{P}(\text{Weather}, \text{Cavity}) =$ a 4×2 matrix of values:

<i>Weather</i> =	sunny	rainy	cloudy	snow
<i>Cavity</i> = true	0.144	0.02	0.016	0.02
<i>Cavity</i> = false	0.576	0.08	0.064	0.08

Conditional probability

Conditional or posterior probabilities

$$\text{e.g., } P(\text{cavity} \mid \text{toothache}) = 0.8$$

i.e., given that *toothache* is all I know

(Notation for conditional distributions:

$$\mathbf{P}(\text{Cavity} \mid \text{Toothache}) = \text{2-element vector of 2-element vectors}$$

If we know more, e.g., *cavity* is also given, then we have

$$P(\text{cavity} \mid \text{toothache}, \text{cavity}) = 1$$

Conditional probability

Definition of conditional probability:

$$P(a \mid b) = P(a \wedge b) / P(b) \text{ if } P(b) > 0$$

Product rule gives an alternative formulation:

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

A general version holds for whole distributions, e.g.,

$$\mathbf{P}(\text{Weather}, \text{Cavity}) = \mathbf{P}(\text{Weather} \mid \text{Cavity}) \mathbf{P}(\text{Cavity})$$

(View as a set of 4×2 equations, **not** matrix mult.)

Inference by enumeration

Start with the joint probability distribution:

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

For any proposition ϕ , sum the atomic events where it is true: $P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$

Inference by enumeration

Start with the joint probability distribution:

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

For any proposition ϕ , sum the atomic events where it is true: $P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$

$$P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by enumeration

Start with the joint probability distribution:

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

For any proposition ϕ , sum the atomic events where it is true: $P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$

$$P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by enumeration

Start with the joint probability distribution:

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

Can also compute conditional probabilities:

$$\begin{aligned}
 P(\neg \textit{cavity} \mid \textit{toothache}) &= \frac{P(\neg \textit{cavity} \wedge \textit{toothache})}{P(\textit{toothache})} \\
 &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064}
 \end{aligned}$$

Normalization

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

Denominator can be viewed as a **normalization constant** α

$$\begin{aligned}
 \mathbf{P}(\text{Cavity} \mid \text{toothache}) &= \alpha, \mathbf{P}(\text{Cavity}, \text{toothache}) \\
 &= \alpha, [\mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch}) + \mathbf{P}(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\
 &= \alpha, [<0.108, 0.016> + <0.012, 0.064>] \\
 &= \alpha, <0.12, 0.08> = <0.6, 0.4>
 \end{aligned}$$

General idea: compute distribution on query variable by fixing **evidence variables** and summing over **hidden variables**

Inference by enumeration, contd.

Typically, we are interested in
the posterior joint distribution of the **query variables** \mathbf{Y}
given specific values \mathbf{e} for the **evidence variables** \mathbf{E}

Let the **hidden variables** be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \sum_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$$

The terms in the summation are joint entries because \mathbf{Y} , \mathbf{E} and \mathbf{H} together exhaust the set of random variables

Independence

A and B are independent iff

$$\mathbf{P}(A/B, \text{Cavity, Toothache, Catch, Weather}) \text{ decomposes into } \mathbf{P}(A/B, \text{Cavity, Toothache, Catch}) \mathbf{P}(B, \text{Weather}) = \mathbf{P}(A) \mathbf{P}(B)$$

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

32 entries reduced to 12; for n independent biased coins, $O(2^n) \rightarrow O(n)$

Absolute independence powerful but rare

Conditional independence

$\mathbf{P}(\text{Toothache, Cavity, Catch})$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

$$(1) \mathbf{P}(\text{catch} / \text{toothache, cavity}) = \mathbf{P}(\text{catch} / \text{cavity})$$

The same independence holds if I haven't got a cavity:

$$(2) \mathbf{P}(\text{catch} / \text{toothache, } \neg \text{cavity}) = \mathbf{P}(\text{catch} / \neg \text{cavity})$$

Catch is **conditionally independent** of *Toothache* given *Cavity*.

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

Conditional independence contd.

Write out full joint distribution using chain rule:

$$\begin{aligned} & \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch}, \textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache} \mid \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity}) \end{aligned}$$

I.e., $2 + 2 + 1 = 5$ independent numbers

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n .

Joint Distributions

- Consider 2 random variables: A, B
 - $P(a, b)$ is shorthand for $P(A = a \text{ AND } B=b)$
 - $\sum_a \sum_b P(a, b) = 1$
 - Can represent $P(A, B)$ as a table of m^2 numbers
- Generalize to more than 2 random variables
 - E.g., A, B, C, ... Z
 - $\sum_a \sum_b \dots \sum_z P(a, b, \dots, z) = 1$
 - $P(A, B, \dots Z)$ is a table of m^K numbers, $K = \#$ variables
 - This is a **potential problem** in practice, e.g., $m=2, K = 20$

Linking Joint and Conditional Probabilities

- Basic fact:
 $P(a, b) = P(a | b) P(b)$
 - Why? Probability of a and b occurring is the same as probability of a occurring given b is true, times the probability of b occurring
- Bayes rule:
 $P(a, b) = P(a | b) P(b)$
 $= P(b | a) P(a)$ by definition

 $\Rightarrow P(b | a) = P(a | b) P(b) / P(a)$ [Bayes rule]

Why is this useful?

Often much more natural to express knowledge in a particular "direction", e.g., in the causal direction

e.g., b = disease, a = symptoms
More natural to encode knowledge as $P(a|b)$ than as $P(b|a)$

Using Bayes Rule

- Example:
 - $P(\text{stiff neck} | \text{meningitis}) = 0.5$ (prior knowledge from doctor)
 - $P(\text{meningitis}) = 1/50,000$ and $P(\text{stiff neck}) = 1/20$
(e.g., obtained from large medical data sets)

$$P(m | s) = P(s | m) P(m) / P(s)$$
$$= [0.5 * 1/50,000] / [1/20] = 1/5000$$

So given a stiff neck, and no other information,
 $p(\text{meningitis} | \text{stiff neck})$ is pretty small

But note that its 10 times more likely that it was before
- so it might be worth measuring more variables for this patient

More Complex Examples with Bayes Rule

- $P(a | b, c) = ??$
 $= P(b, c | a) P(a) / P(b, c)$
- $P(a, b | c, d) = ??$
 $= P(c, d | a, b) P(a, b) / P(c, d)$

Both are examples of basic pattern $p(x|y) = p(y|x)p(x)/p(y)$

(it helps to group variables together, e.g., $y = (a,b)$, $x = (c, d)$)

Note also that we can write $P(x | y)$ is proportional to $P(y | x) P(x)$
(the $P(y)$ term on the bottom is just a normalization constant)

Sequential Bayesian Reasoning

- h = hypothesis, $e_1, e_2, .. e_n$ = evidence
- $P(h)$ = prior
- $P(h | e_1)$ proportional to $P(e_1 | h) P(h)$
 $=$ likelihood of e_1 x prior(h)
- $P(h | e_1, e_2)$ proportional to $P(e_1, e_2 | h) P(h)$
in turn can be written as $P(e_2 | h, e_1) P(e_1 | h) P(h)$
 \sim likelihood of e_2 x "prior"(h given e_1)
- Bayes rule supports sequential reasoning
 - Start with prior $P(h)$
 - New belief (posterior) = $P(h | e_1)$
 - This becomes the "new prior"
 - Can use this to update to $P(h | e_1, e_2)$, and so on.....

Computing with Probabilities: Law of Total Probability

Law of Total Probability (aka "summing out" or marginalization)

$$\begin{aligned} P(a) &= \sum_b P(a, b) \\ &= \sum_b P(a | b) P(b) \quad \text{where B is any random variable} \end{aligned}$$

Why is this useful?

Given a joint distribution (e.g., $P(a,b,c,d)$) we can obtain any "marginal" probability (e.g., $P(b)$) by summing out the other variables, e.g.,

$$P(b) = \sum_a \sum_c \sum_d P(a, b, c, d)$$

We can compute any conditional probability given a joint distribution, e.g.,

$$\begin{aligned} P(c | b) &= \sum_a \sum_d P(a, c, d | b) \\ &= \sum_a \sum_d P(a, c, d, b) / P(b) \end{aligned}$$

where $P(b)$ can be computed as above

Computing with Probabilities: The Chain Rule or Factoring

We can always write

$$P(a, b, c, \dots z) = P(a | b, c, \dots z) P(b, c, \dots z)$$

(by definition of joint probability)

Repeatedly applying this idea, we can write

$$P(a, b, c, \dots z) = P(a | b, c, \dots z) P(b | c, \dots z) P(c | \dots z) \dots P(z)$$

This factorization holds for any ordering of the variables

This is the chain rule for probabilities

What does all this have to do with AI?

Logic-based knowledge representation

Set of sentences in KB

Agent's belief in any sentence is: true, false, or unknown

In real-world problems there is uncertainty

$P(\text{snow in New York on January 1})$ is not 0 or 1 or unknown

$P(\text{vehicle speed} > 50 \mid \text{sensor reading})$

$P(\text{Dow Jones will go down tomorrow} \mid \text{data so far})$

$P(\text{pit in square 2,2} \mid \text{evidence so far})$

Not acknowledging this uncertainty can lead to brittle systems and inefficient use of information

Uncertainty is due to:

Things we did not measure (which is always the case)

E.g., in economic forecasting

Imperfect knowledge

Agents, Probabilities, and Degrees of Belief

What we were taught in school

$P(a)$ represents the frequency that event a will happen in repeated trials

-> "relative frequency" interpretation

Degree of belief

$P(a)$ represents an agent's degree of belief that event a is true

This is a more general view of probability

Agent's probability is based on what information they have

E.g., based on data or based on a theory

Examples:

a = "life exists on another planet"

What is $P(a)$? We will all assign different probabilities

a = "Hillary Clinton will be the next US president"

What is $P(a)$?

a = "over 50% of the students in this class will get A's"

What is $P(a)$?

Probabilities can vary from agent to agent depending on their models of the

More on Degrees of Belief

- Our interpretation of $P(a | e)$ is that it is an agent's degree of belief in the proposition a , given evidence e
 - Note that proposition a is true or false in the real-world
 - $P(a|e)$ reflects the agent's uncertainty or ignorance
- The degree of belief interpretation does not mean that we need new or different rules for working with probabilities
 - The same rules (Bayes rule, law of total probability, probabilities sum to 1) still apply – our interpretation is different
- If Agent 1 has inconsistent sets of probabilities (violate axioms of probability theory) then there exists a betting strategy that allows Agent 2 to always win in bets against Agent 1
 - See Section 13.2 in text, de Finetti's argument

Decision Theory – why probabilities are useful

- Consider 2 possible actions that can be recommended by a medical decision-making system:
 - a = operate
 - b = don't operate
- 2 possible states of the world
 - c = patient has cancer, and also not(c)
- Given evidence so far, agent's degree of belief in c is $p(c|e)$
- Costs (to agent) associated with various outcomes:
 - Take action a and patient has cancer: cost = \$30k
 - Take action a and patient has no cancer: cost = -\$50k
 - Take action b and patient has cancer: cost = -\$100k
 - Take action b and patient has no cancer: cost = 0.

Maximizing expected utility (or minimizing expected cost)

What action should the agent take?

A rational agent should maximize expected utility, or equivalently minimize expected cost

Expected cost of actions:

$$E[\text{cost}(a)] = 30 p(c) - 50 [1 - p(c)]$$

$$E[\text{cost}(b)] = -100 p(c)$$

$$\text{Break even point? } 30p - 50 + 50p = -100p$$

$$100p + 30p + 50p = 50$$

$$\Rightarrow p(c) = 50/180 \sim 0.28$$

If $p(c) > 0.28$, the optimal decision is to operate

Original theory from economics, cognitive science (1950's)

- But widely used in modern AI, e.g., in robotics, vision, game-playing

Note that we can only make optimal decisions if we know the probabilities

Constructing a Propositional Probabilistic Knowledge Base

- Define all variables of interest: A, B, C, ... Z
- Define a joint probability table for $P(A, B, C, \dots Z)$
 - We have seen earlier how this will allow us to compute the answer to any query, $p(\text{query} \mid \text{evidence})$, where query and evidence = any propositional sentence
- 2 major problems:
 - Computation time:
 - $P(a|b)$ requires summing out over all other variables in the model, e.g., $O(m^{k-1})$ with k variables
 - Model specification
 - Joint table has $O(m^k)$ entries – where will all the numbers come from?
 - These 2 problems effectively halted the use of probability in AI research from the 1960's up until about 1990

Independence

2 random variables A and B are independent iff

$$P(a, b) = P(a) P(b) \quad \text{for all values } a, b$$

More intuitive (equivalent) conditional formulation

A and B are independent iff

$$P(a | b) = P(a) \quad \text{OR} \quad P(b | a) = P(b), \quad \text{for all values } a, b$$

Intuitive interpretation:

$P(a | b) = P(a)$ tells us that knowing b provides no change in our probability for a, i.e., b contains no information about a

Can generalize to more than 2 random variables

In practice true independence is very rare

"butterfly in China" effect

Weather and dental example in the text

Conditional independence is much more common and useful

Note: independence is an assumption we impose on our model of the world - it does not follow from basic axioms

Conditional Independence

2 random variables A and B are conditionally independent given C iff

$$P(a, b | c) = P(a | c) P(b | c) \quad \text{for all values } a, b, c$$

More intuitive (equivalent) conditional formulation

A and B are conditionally independent given C iff

$$P(a | b, c) = P(a | c) \quad \text{OR} \quad P(b | a, c) = P(b | c), \quad \text{for all values } a, b, c$$

Intuitive interpretation:

$P(a | b, c) = P(a | c)$ tells us that learning about b, given that we already know c, provides no change in our probability for a,

i.e., b contains no information about a beyond what c provides

Can generalize to more than 2 random variables

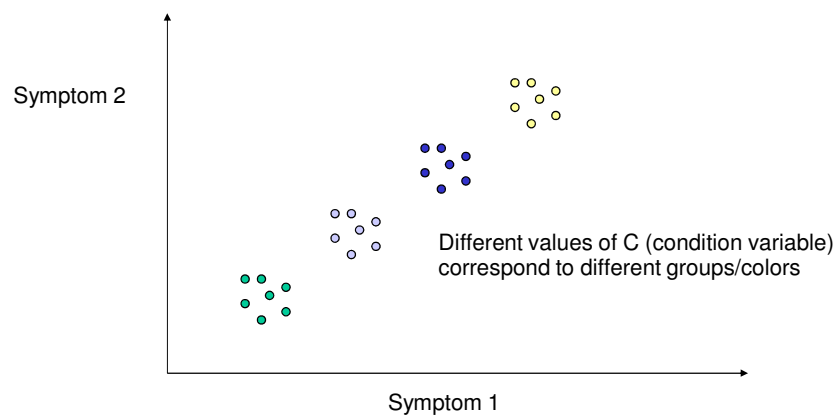
E.g., K different symptom variables X_1, X_2, \dots, X_K , and $C = \text{disease}$

$$P(X_1, X_2, \dots, X_K | C) = \prod P(X_i | C)$$

Conditional Independence vs. Independence

- Conditional independence does not imply independence
- Example:
 - A = height
 - B = reading ability
 - C = age
 - $P(\text{reading ability} \mid \text{age, height}) = P(\text{reading ability} \mid \text{age})$
 - $P(\text{height} \mid \text{reading ability, age}) = P(\text{height} \mid \text{age})$
- Note:
 - Height and reading ability are dependent (not independent) but are conditionally independent given age

Another Example



In each group, symptom 1 and symptom 2 are conditionally independent.
But clearly, symptom 1 and 2 are marginally dependent (unconditionally).

"...probability theory is more fundamentally concerned with the structure of reasoning and causation than with numbers."

Glenn Shafer and Judea Pearl
Introduction to Readings in Uncertain Reasoning,
Morgan Kaufmann, 1990

Conclusions...

- Representing uncertainty is useful in knowledge bases
 - Probability provides a coherent framework for uncertainty
- Full joint distributions are intractable to work with
- Conditional independence assumptions allow much simpler models of real-world phenomena
- Bayesian networks are a systematic way to construct parsimonious structured distributions
- **Rational agents cannot violate probability theory.**

Summary

Probability is a rigorous formalism for uncertain knowledge

Joint probability distribution specifies probability of every **atomic event**

Queries can be answered by summing over atomic events

For nontrivial domains, we must find a way to reduce the joint size