Part II
Methods of AI

Chapter 5
Uncertainty and Reasoning

Chapter 5 – Uncertainty and Reasoning

5.1 Uncertainty

5.2 Probabilistic Reasoning

5.3 Probabilistic Reasoning over Time

5.4 Making Decisions

AI: The Role of Statistics & Probability

„Who needs linguistics when we have statistics!“

„The desert Race of autonomous Vehicles“

(Sebastian Thrun)

→ Computer Vision


5.1 Uncertainty

Uncertainty – Introduction (1)

Let action $\mathbf{A_t}$ = “leave for airport t minutes before flight”

Will A_t get me there in time?

Problems:

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

Uncertainty – Introduction (2)

Hence a purely logical approach either

1) risks falsehood: "A_{25} will get me there on time" or

3) 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. . . . "

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

1. 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?

2. Rules with probability (fudge) factors:

   $A_{25} \rightarrow_{0.3}$ "get there on time"
   $Sprinkler \rightarrow_{0.99}$ "*WetGrass*"
   $WetGrass \rightarrow_{0.7}$ "*Rain*"

Methods for handling uncertainty (2)

Probability

Given the available evidence,

A_{25} will get me there on time with probability 0.40

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

Beliefs vs Truth:

Fuzzy logic handles *degree of truth* NOT uncertainty e.g.,

"*WetGrass* is true to degree 0.2" ?

Probability (1)

Probability assertions *summarize* 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., $P(A_{25} \mid \text{no reported accident}) = 0.60$

Probability (2)

These are not claims of some probabilistic tendency in the current situation
(but a bit might be learned from past experience of similar situations)

Probabilities of propositions change with new evidence:
e.g., $P(A_{25} \mid \text{no reported accident, 5 a.m.}) = 0.45$

(Analogous to logical entailment (status *KB)* $\models \alpha$, not truth.)

Making decisions under uncertainty

Suppose I believe the following:

$$P(A_{25} \text{ gets me there in time} \mid \ldots) = 0.40$$
$$P(A_{90} \text{ gets me there in time} \mid \ldots) = 0.70$$
$$P(A_{120} \text{ gets me there in time} \mid \ldots) = 0.95$$
$$P(A_{1440} \text{ gets me there in time} \mid \ldots) = 0.9999$$

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 dice.

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

A *probability space* or *probability model* is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ such that:

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

$\Sigma_{\omega} P(\omega) = 1$

e.g.,the dice: $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An *event A* is any subset of Ω

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

e.g., $P(\text{dice roll} < 4) = 1/6 + 1/6 + 1/6 = ½$

Random Variables

A random variable is a function from sample points to some range, e.g., the Reals or Booleans e.g., *Odd*(1) = *true*

P induces a *probability distribution* for any random variable *X*:

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

e.g., *P*(*Odd* = *true*) = 1/6 + 1/6 + 1/6 = 1/2

Propositions (1)


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 $A(\omega) = true$
event $\neg a$ = set of sample points where $A(\omega) = false$
event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$

Propositions (2)

Often in AI applications, the sample points are defined by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables.

Example:

With Boolean variables, sample point = propositional logic model
e.g., $A = true$, $B = false$, or $a \wedge \neg b$

Proposition = disjunction of atomic events in which it is true

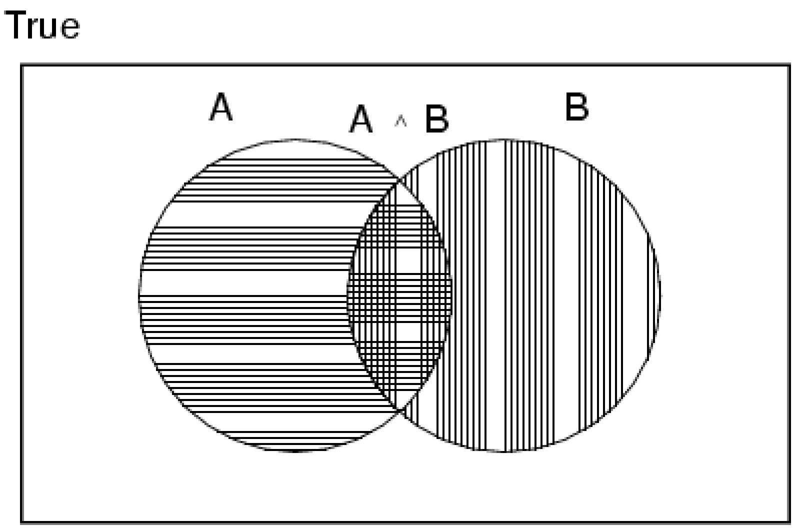
$$\text{e.g., } (a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$$

$$\Rightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$$

Why use probability?

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

E.g., $P(a \lor b) = P(a) + P(b) - P(a \land b)$



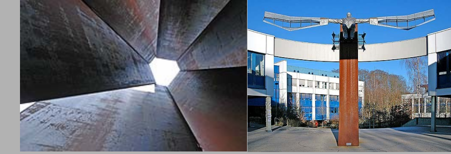
de Finetti (1931): An agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

De Finetti‘s Argument (Example)


Agent 1		Agent 2		Outcome for Agent 1			
Proposition	Belief	Bet	Stakes	$A \wedge B$	$A \wedge \neg B$	$\neg A \wedge B$	$\neg A \wedge \neg B$
A	0.4	A	4 to 6	-6	-6	4	4
B	0.3	B	3 to 7	-7	3	-7	3
$A \vee B$	0.8	$\neg(A \vee B)$	2 to 8	2	2	2	-8
				-11	-1	-1	-1

$\Rightarrow$ Agent 1 always looses, since 0.4 + 0.3 – 0.0 =/ 0.8

Syntax for propositions



Propositional or Boolean random variables

e.g., *Cavity* (do I have a cavity?) is ⟨*true, false*⟩

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 (1)

Prior or unconditional probabilities of propositions
e.g., P(*Cavity* = *true*) = 0.1 and *P*(*Weather* = *sunny*) = 0.72

California !

Or just: P(cavity) = 0.1

Probability distribution gives values for all possible assignments:

P(*Weather*) = ⟨0.72, 0.1, 0.08, 0.1⟩ = ⟨sunny, rain, cloudy, snow⟩
(*normalized*, i.e., sums to 1)

Prior probability (2)

Joint probability distribution for a set of random variables gives the probability of every atomic event on those variables. (i.e., every sample point)

P(*Weather*, *Cavity*) = a 4 x 2 matrix of values:

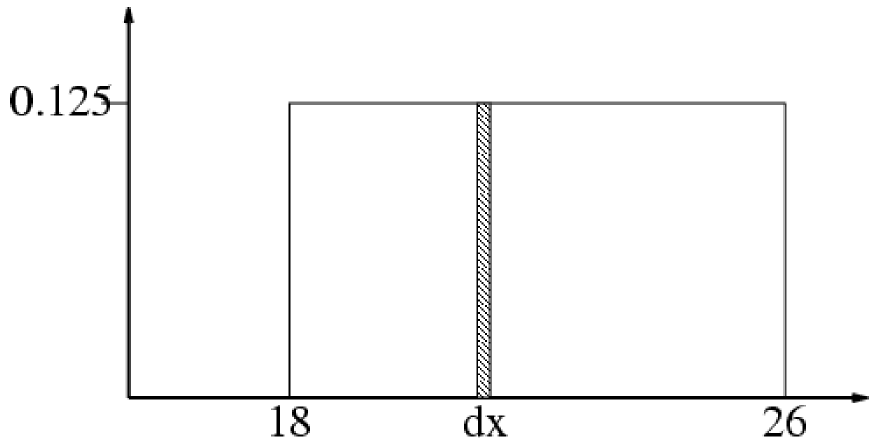
Weather =	*Sunny*	*rain*	*cloudy*	*snow*
Cavity = *true*	0.144	0.02	0.016	0.02
Cavity = *false*	0.576	0.08	0.064	0.08

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

Probability for continuous Variables

Express distribution as a parameterized function of value:

$P(X = x) = U[18.26](x)$ = "uniform density between 18 and 26"



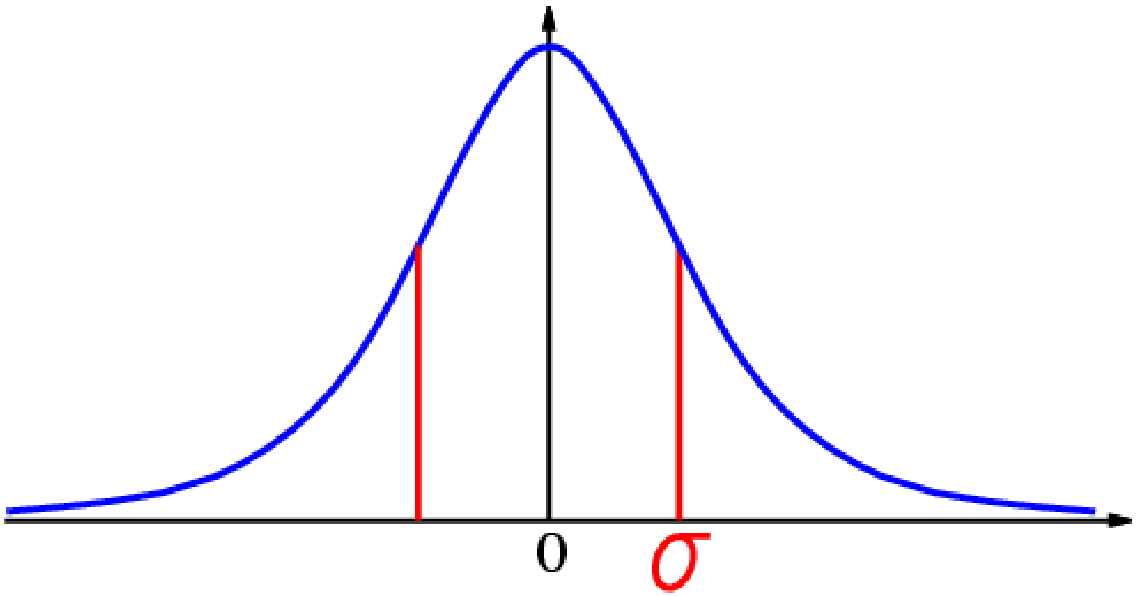
Here P is a density, integrates to 1.

$P(X = 20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Gaussian Density

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



Conditional probability (1)

Conditional or posterior probabilities

e.g., *P*(*cavity* | *toothache*) = 0.8

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

NOT: "if *toothache* then 80% chance of *cavity*"

discussion!

Notation for conditional distributions:

P(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors

Conditional probability (2)

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

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

Note: the less specific belief *remains valid* after more evidence arrives, but is not always *useful.*

New evidence may be irrelevant, allowing simplification, e.g.

$$P(cavity \mid toothache, 49ersWin) = P(cavity \mid toothache) = 0.8$$

This kind of inference, sanctioned by domain knowledge, is crucial.

Conditional probability (3)

Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \quad if \;\; P(b) \neq 0$$

Product rule gives an alternative formulation:

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

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

P(Weather, Cavity) = P(Weather | Cavity) P(Cavity)

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

Conditional Probability vs. Implication


Take care: $P(B/A) \neq P(A \Rightarrow B)$

Example:

$$P(A, B) = 0.25$$
$$P(A, \neg B) = 0.25$$
$$P(\neg A, B) = 0.25$$
$$P(\neg A, \neg B) = 0.25$$

A, B	$A, \neg B$
$\neg A, B$	$\neg A, \neg B$

$$P(A \Rightarrow B) = P(A, B) + P(A, \neg B) + P(\neg A, \neg B)$$
$$= 0.75$$

A, B	$(A, \neg B)$
$\neg A, B$	$(\neg A, \neg B)$

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{0.25}{0.5}$$
$$= 0.5$$

Conditional probability (4)

Chain rule is derived by successive application of product rule:

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

$$=P(X_1,\ldots,X_{n-2})P(X_{n_1}\;\;X_1,\ldots,X_{n-2})P(X_n\;\;X_1,\ldots,X_{n-1})$$

$$=\ldots$$

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

Inference by enumeration (1)


Start with the joint distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

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

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

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

Inference by enumeration (2)

Start with the joint distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

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

$$P(\Theta) = \sum_{\{\omega : \omega \models \Theta\}} P(\omega)$$

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

Inference by enumeration (3)

Start with the joint distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

Can also compute conditonal probabilities:

$$P(\neg\, cavity | toothache) = \frac{P(\neg\, cavity \wedge toothache)}{P(toothache)}$$

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

Normalization


Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Dominator can be viewed as a normalization constant

$\mathbf{P}(\text{Cavity}|\text{toothache}) = \alpha\ \mathbf{P}(\mathit{Cavity, toothache})$

$= \alpha\ [\mathbf{P}(\mathit{Cavity, toothache, catch}) + \mathbf{P}(\mathit{Cavity, toothache}, \neg\, \mathit{catch})]$

$= \alpha\ [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\ \langle 0.12, 0.08\rangle = \langle 0.6, \mathit{0.4}\rangle$

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 Y

given specific values e for the evidence variables E

Let the hidden variables be H = X – Y – E

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

$$P(Y|\mathrm{E}=\mathrm{e}) = \alpha\ \mathrm{P}(Y|\mathrm{E}=\mathrm{e}) = \alpha \sum_h P(\mathrm{Y}, \mathrm{E}=\mathrm{e}, \mathrm{H}=\mathrm{h})$$

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

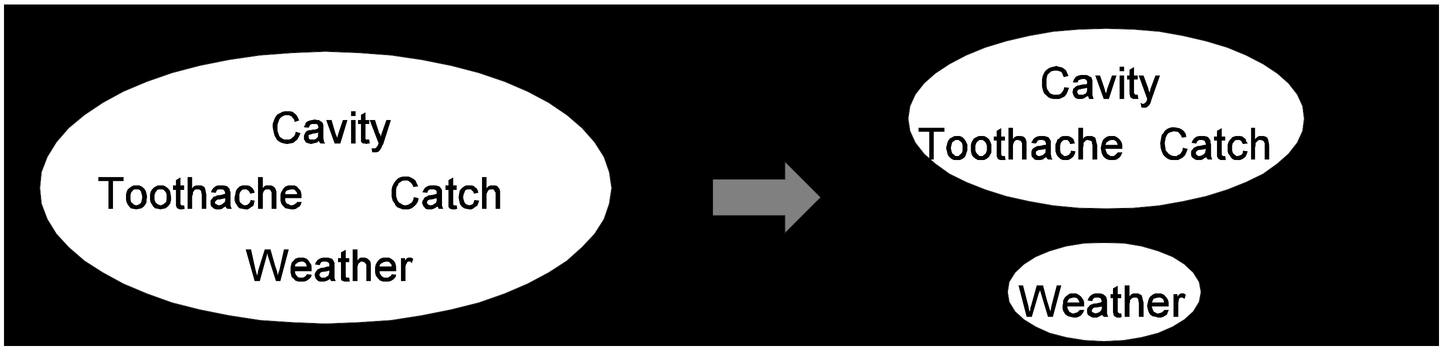
Obvious problems:

1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Independence

A and *B* are independent iff

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



P(*Toothache, Catch, Cavity, Weather)*

= **P**(*Toothache, Catch, Cavity)* **P***(Weather)*

32 entries reduced to 12 !

for *n* independent biased coins, $2^n \rightarrow n$

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence (1)

P(*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:

$$(1)\quad P(catch \mid toothache, cavity) = P(catch \mid cavity)$$

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

$$(2)\quad P(catch \mid toothache, \neg cavity) = P(catch \mid \neg cavity)$$

Catch is *conditionally independent of Toothache given Cavity:*

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

Conditional independence (2)

Equivalent statements:

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

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

Write out full joint distribution using chain rule:

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

I.e., 2 + 2 + 1 = 5 independent numbers (equation 1 and 2 remove2)

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.

Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes' Rule

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

$\Rightarrow$ **Bayes' Rule** $P(a\ \ b) = \dfrac{P(b\ \ a)P(a)}{P(b)}$

or in distribution form

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha\ P(X|Y)P(Y)$$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)} = \alpha\ P(X|Y)P(Y)$$

E.g., let M be meningitis, S be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small.

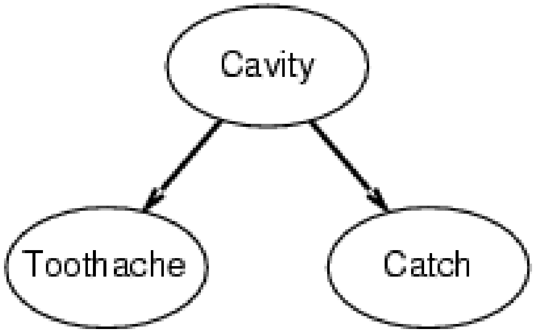
Bayes' Rule: conditional independence

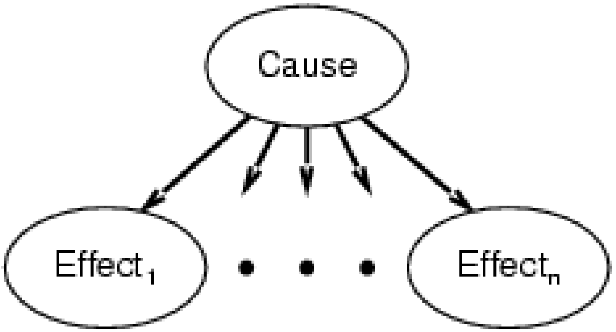
P(*cavity* | *toothache* ∧ *catch*)

= α **P**(*toothache* ∧ *catch* | *Cavity*) **P**(*Cavity*)
= α **P**(*toothache* | *Cavity*) **P**(*catch* | *Cavity*) **P**(*Cavity*)

This is an example of a naïve Bayes model:

$$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause) \prod_i P(Effect_i \;\; Cause)$$





Total number of parameter is linear in *n*.

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.

- Independence and conditional independence provide the tools.

Normalisation

Relative comparison of probabilities is often sufficient:

M := Meningitis, N := Nackensteife, S := Schleudertrauma (whiplash)

$$P(M|N) = \frac{P(N|M) * P(M)}{P(N)} \qquad P(S|N) = \frac{P(N|S) * P(S)}{P(N)}$$

$$\frac{P(M|N)}{P(S|N)} = \frac{P(N|M) * P(M)}{P(N|S) * P(S)}$$

$\Rightarrow$ Comparison of both diagnoses possible without knowledge on $P(N)$

$\Rightarrow$ Often decisions can be based on relative comparison of probabilities

Normalisation


Sometimes relative probabilities are weak for careful diagnoses;
Nevertheless, knowledge about basic probabilities like $P(N)$ can often be avoided.

$$P(M|N) = \frac{P(N|M) * P(M)}{P(N)} \qquad P(\neg M|N) = \frac{P(N|\neg M) * P(\neg M)}{P(N)}$$

$$P(M|N) + P(\neg M|N) = \frac{P(N|M) * P(M)}{P(N)} + \frac{P(N|\neg M) * P(\neg M)}{P(N)}$$
$$= 1/\,P(N) * (P(N|M) * P(M) + P(N|\neg M) * P(\neg M)) = 1$$

$$P(N) = P(N|M) * P(M) + P(N|\neg M) * P(\neg M)$$

$$P(M|N) = \frac{P(N|M) * P(M)}{P(N|M) * P(M) + P(N|\neg M) * P(\neg M))}$$

In general:

$$P(M|N) = \alpha * P(N|M) * P(M)$$

where α is normalisation constant, such that CPT entries for PCMIN sum up to 1.