Lecture 8
Uncertainty

Marco Chiarandini

Deptartment of Mathematics & Computer Science
University of Southern Denmark

Slides by Stuart Russell and Peter Norvig

Course Overview

- ✔ Introduction
  - ✔ Artificial Intelligence
  - ✔ Intelligent Agents
- ✔ Search
  - ✔ Uninformed Search
  - ✔ Heuristic Search
- ✔ Adversarial Search
  - ✔ Minimax search
  - ✔ Alpha-beta pruning
- ✔ Knowledge representation and Reasoning
  - ✔ Propositional logic
  - ✔ First order logic
  - ✔ Inference
- Uncertain knowledge and Reasoning
  - Probability and Bayesian approach
  - Bayesian Networks
  - Hidden Markov Chains
  - Kalman Filters
- Learning
  - Decision Trees
  - Maximum Likelihood
  - EM Algorithm
  - Learning Bayesian Networks
  - Neural Networks
  - Support vector machines

Summary

First-order logic:
- objects and relations are semantic primitives
- syntax: constants, functions, predicates, equality, quantifiers

Increased expressive power: sufficient to define wumpus world

Situation calculus:
- conventions for describing actions and change in FOL
- can formulate planning as inference on a situation calculus KB

Outline

1. Probability Calculus

Outline

◇ Uncertainty
◇ Probability
◇ Syntax and Semantics
◇ Inference
◇ Independence and Bayes' Rule

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 (KCBS traffic reports)
3) uncertainty in action outcomes (flat tire, etc.)
4) immense complexity of modelling and predicting traffic

Hence a purely logical approach either

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

Logic-based abductive inference: 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:
$A_{25} \mapsto_{0.3} AtAirportOnTime$
$Sprinkler \mapsto_{0.99} WetGrass$
$WetGrass \mapsto_{0.7} Rain$
Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??

Probability
Given the available evidence,
A_{25} will get me there on time with probability 0.04
Mahaviracarya (9th C.), Cardano (1565) theory of gambling

(Fuzzy logic handles **degree of truth** NOT uncertainty e.g.,
WetGrass is true to degree 0.2)

Probability

Probabilistic 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., $P(A_{25}|\text{no reported accidents}) = 0.06$

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., $P(A_{25}|\text{no reported accidents}, \text{5 a.m.}) = 0.15$
(Analogous to logical entailment status $KB \models \alpha$, not truth.)

Making decisions under uncertainty

Suppose I believe the following:

$$\begin{aligned} P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\ P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\ P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\ P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999 \end{aligned}$$

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

Interpretations of Probability

- Classical interpretation: probabilities can be determined a priori by an examination of the space of possibilities.
  It assigns probabilities in the absence of any evidence, or in the presence of symmetrically balanced evidence
- Logical interpretation: generalizes the classcial it in two important ways:
  - possibilities may be assigned unequal weights
  - probabilities can be computed whatever the evidence may be, symmetrically balanced or not
- Frequentist: the probability of an attribute A in a finite reference class B is the relative frequency of actual occurrences of A within B.
  issue of identity
- Propensity interpretation: innate property of the objects
- Subjective interpretation: subjective degree of belief + betting system to avoid unconstrained subjectivism

Probability basics

DEFINITION

ELEMENTARY EVENT

An *elementary* or *atomic event* is a happening or occurrence that cannot be made up of other events.

EVENT, E

An *event* is a set of elementary events.

SAMPLE SPACE, S

The set of all possible outcomes of an event E is the *sample space* S or *universe* for that event.

PROBABILITY, p

The *probability* of an event E in a sample space S is the ratio of the number of elements in E to the total number of possible outcomes of the sample space S of E. Thus, p(E) = |E| / |S|.

Probability basics

The probability of any event E from the sample space S is:

$0 \leq p(E) \leq 1$, where $E \subseteq S$

The sum of the probabilities of all possible outcomes is 1

The probability of the complement of an event is

$p(\overline{E}) = (|S| - |E|) \ / \ |S| \ = \ (|S| \ / \ |S|) - (|E| \ / \ |S|) = 1 - p(E).$

The probability of the contradictory or false outcome of an event

$p(\{ \}) = 1 - p(\overline{\{ \}}) = 1 - p(S) = 1 - 1 = 0$, or alternatively,

$= |\{ \}| / |S| = 0 / |S| = 0$

Probability basics

DEFINITION

INDEPENDENT EVENTS

Two events A and B are *independent* if and only if the probability of their both occurring is equal to the product of their occurring individually. This independence relation is expressed:

$$p(A \cap B) = p(A) * p(B)$$

We sometimes use the equivalent notation p(s,d) for $p(s \cap d)$. We clarify the notion of independence further in the context of conditional probabilities in Section 5.2.4.

The three Kolmogorov Axioms

1. The probability of event E in sample space S is between 0 and 1, ie, $0 \leq p(E) \leq 1$
2. When the union of all E gives S, $p(S) = 1$ and $p(\bar{S}) = 0$
3. The probability of the union of two sets of events A and B is:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Probability basics

DEFINITION

RANDOM VARIABLE

A *random variable* is a function whose domain is a sample space and range a set of outcomes, most often real numbers. Rather than using a problem-specific event space, a random variable allows us to talk about probabilities as numerical values that are related to an event space.

BOOLEAN, DISCRETE, and CONTINUOUS RANDOM VARIABLES

A *boolean random variable* is a function from an event space to {true, false} or to the subset of real numbers $\{0.0, 1.0\}$. A boolean random variable is sometimes called a *Bernoulli trial*.

A *discrete random variable*, which includes boolean random variables as a subset, is a function from the sample space to (a countable subset of) real numbers in [0.0, 1.0].

A *continuous random variable* has as its range the set of real numbers.

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

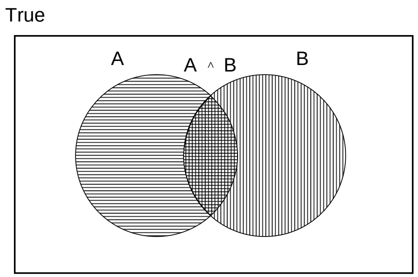
Often in AI applications, the sample points are **defined** by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables

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

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.

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 $\langle sunny, rain, cloudy, snow \rangle$
$Weather = rain$ 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., $Temp < 22.0$.

Prior probability

Prior or unconditional probabilities of propositions
e.g., $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$
correspond to belief prior to arrival of any (new) evidence

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

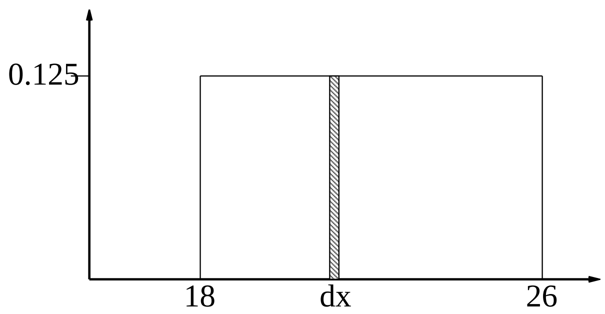
Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	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

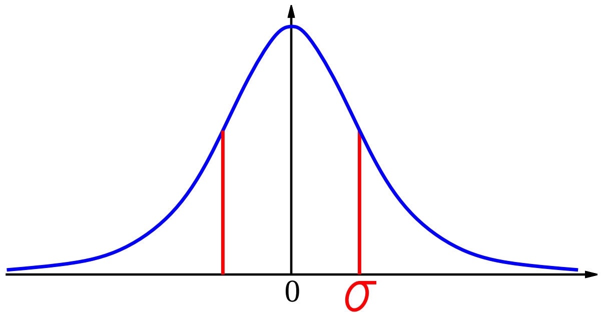


Here P is a density; integrates to 1.
$P(X = 20.5) = 0.125$ really means

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

Gaussian density

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



Conditional probability

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*" (Notation for conditional distributions:
$\mathbf{P}(Cavity|Toothache)$ = 2-element vector of 2-element vectors)

If we know more, e.g., *cavity* is also given, then we have
$P(cavity|toothache, cavity) = 1$
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.,
$P(cavity|toothache, 49ersWin) = P(cavity|toothache) = 0.8$
This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

Definition

Conditional probability:

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

Product rule gives an alternative formulation:
$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

A general version holds for whole distributions, e.g.,
$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$
(View as a 4×2 set of equations, **not** matrix mult.)

Definition

Chain rule is derived by successive application of product rule:

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

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.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 distribution:

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

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

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

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

Inference by enumeration

Start with the joint distribution:

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

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

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

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

Inference by enumeration

Start with the joint distribution:

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

Can also compute conditional probabilities:

$$\begin{aligned} P(\neg cavity | toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 \end{aligned}$$

Normalization

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

Denominator can be viewed as a normalization constant α

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

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

Inference by enumeration, contd.

Let $\mathbf{X}$ be all the variables. Typically, we want
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:

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

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

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???

Probability basics

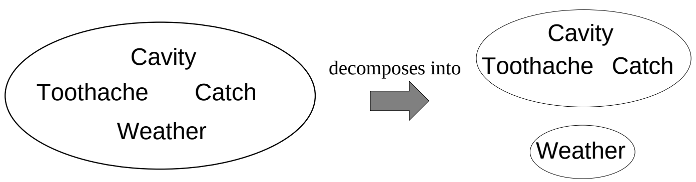
DEFINITION

INDEPENDENT EVENTS
Two events A and B are *independent* of each other if and only if $p(A \cap B) = p(A)\,p(B)$. When $p(B) \neq 0$ this is the same as saying that $p(A) = p(A|B)$. That is, knowing that B is true does not affect the probability of A being true.

CONDITIONALLY INDEPENDENT EVENTS
Two events A and B are said to be *conditionally independent* of each other, given event C if and only if $p((A \cap B) \mid C) = p(A \mid C)\,p(B \mid C)$.

Independence

A and B are independent iff
$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$



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

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

$\mathbf{P}(Toothache, Cavity, Catch)$ 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) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:
(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:
$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Equivalent statements:
$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$
$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$