Artificial Intelligence 2

Roman Barták
Department of Theoretical Computer Science and Mathematical Logic

Introduction

We construct rational agents.

An **agent** is an entity that perceives its **environment** through **sensors** and acts upon that environment through **actuators**.

A rational agent is an agent maximizing its expected performance measure.

In AI 1 we dealt mainly with a logical approach to agent design (no uncertainty).

We ignored

- interface to environment (sensors, actuators)
- uncertainty
- the possibility of self-improvement (learning)

Introduction

– motivation and background on probability

Probabilistic reasoning

– uncertainty, probabilistic reasoning, Bayesian networks, Hidden Markov Models

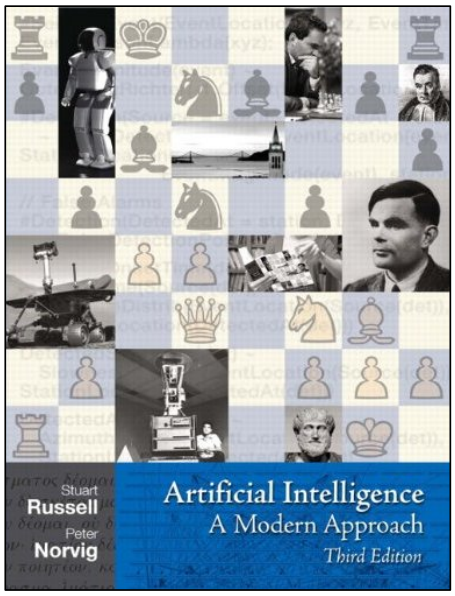
Rational decisions

– utility theory, Markov Decision Processes, game theory, mechanism design

Machine learning

– decision trees, regression, SVM, reinforcement learning

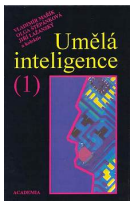
Artificial Intelligence: A Modern Approach

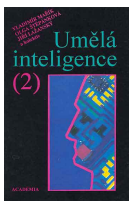


– S. Russell and P. Norvig
– Prentice Hall, 2010 (3rd ed.)
– http://aima.cs.berkeley.edu/

Umělá inteligence 1-6

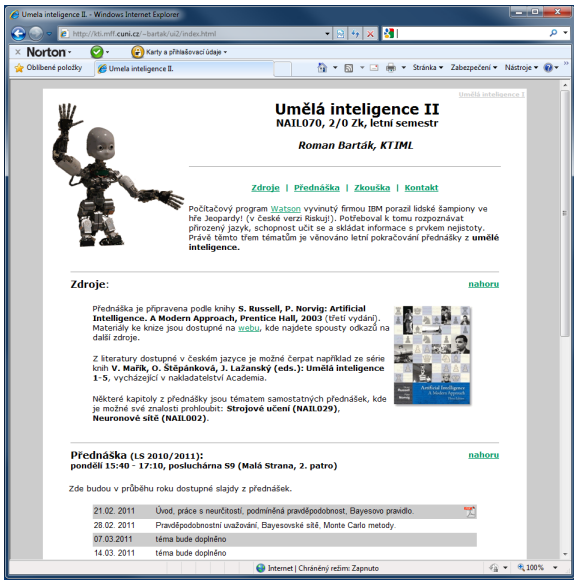
– Vladimír Mařík, Olga Štěpánková, Jiří Lažanský a kol.
– Academia





Course web site

http://ktiml.mff.cuni.cz/~bartak/ui2



You can find there:

– slides

– links and resources

– contacts

– quiz

– …

Links to other courses

Seminar on Artificial Intelligence II

– how to apply AI techniques in practice

Machine learning

– how can computers learn new things

Multi-agent systems

– how to handle multiple agents

Probabilistic graphical models

– how to do Bayesian inference efficiently etc.

Human-like artificial agents

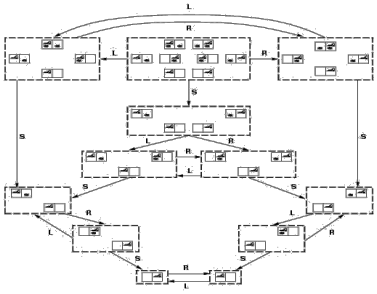
– how to design agents for virtual environments

Practical course on robotics

– how to design hardware agents

Uncertainty so far

Can we handle uncertain information in the pure logical approach?



belief states

– represent sets of all possible world states for the agent

Drawbacks

– a logical agent must consider every logically possible explanation for the observations, no matter how unlikely (large and complex representations)

– a correct contingent plan must consider arbitrary likely contingencies (big plans)

– sometimes there is no plan that is guaranteed to achieve the goal, yet the agent must act

Example

Diagnosing a dental patient's toothache

Let us try to apply propositional logic:

Toothache ⇒ Cavity

Hmm, is it really true?

– not all patients with toothaches have cavities; some of them have gum disease, an abscess, or other problems

Toothache ⇒ Cavity ∨ GumProblem ∨ Abscess ∨ …

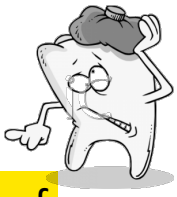
We could try turning the rule into a causal rule:

Cavity ⇒ Toothache

But this is not right either – not all cavities cause pain

The only way to fix the rule is to make it logically exhaustive!

Using a logical approach



Why does logic fail to cope with a domain like medical diagnosis?

- **laziness**: it is too much work to list the complete set of antecedent or consequents and too hard to use such rules
- **theoretical ignorance**: medical science has no complete theory for the domain
- **practical ignorance**: even if we know all the rules we might be uncertain because not all the necessary tests have been or can be run

We need another tool to deal with degrees of belief – **probability theory**.

A logical agent believes each sentence to be true or false or has no opinion. A probabilistic agent may have a numerical degree of belief between 0 (certainly false) and 1 (certainly true).

Basic probability notation

Like logical assertions, probabilistic assertions are about possible worlds – **sample space** Ω.

– the possible worlds are **mutually exclusive** and **exhaustive**

Each possible world ω is associated with a numerical probability $P(\omega)$ such that:

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

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

Example:

If we are about to roll two (distinguishable) dice, there are 36 possible worlds to consider: (1,1), (1,2),…, (6,6)

$P(\omega) = 1/36$

Events

The sets of possible worlds are called **events**.

Example: „doubles are rolled“ is an event

The probability of event is the sum of probabilities of possible worlds in the event.

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

Example:

P(doubles) = 1/36+1/36+1/36+1/36+1/36+1/36 = 1/6

These probabilities are called **unconditional** or **prior probabilities** („priors“ for short).

Conditional probability

Frequently, we have some information (**evidence**) and we are interested in probability of some event.

For example, what is the probability of double if we already know that first die rolled to 5?

*P(doubles | Die_1 = 5) = 1/36 / (6*1/36) = 1/6*

This is called **conditional** or **posterior probability**

$P(a \mid b) = P(a \wedge b) / P(b)$, whenever $P(B) > 0$

This can be also written in a different form called the **product rule**

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

Beware! If we have more evidence then the conditional probability needs to assume it.

P(doubles | Die_1 = 5, Die_2 = 5) = 1

Random variables

In a factored representation, a possible world is represented by a set of variable/value pairs.

Variables in probability theory are called **random variables**. Every random variable has a domain – the set of possible values it can take on (similarly to a CSP).

Die_1 – represents a value on the first die 1 (1,…,6)

Cavity – describes whether the patient has or has not cavity (true, false)

A possible world is fully identified by values of all random variables.

$P(Die_1 = 5, Die_2 = 5)$

Probability distribution

Probability of all possible worlds can be described using a table called a **full joint probability distribution** – the elements are indexed by values of random variables.

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

Given the table, we can calculate probabilities of values of any random variable:

P(toothache=true) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
P(toothache=false) = 0.072+ 0.008 + 0.144 + 0.576 = 0.8

We will describe the table in a short way as:

P(Toothache) = ⟨0.2, 0.8⟩

Probability axioms

$P(\neg a) = 1 - P(a)$

inclusion-exclusion principle

$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

chain rule

$\mathbf{P}(A,B,C,D)$

$= \mathbf{P}(A|B,C,D)\ \mathbf{P}(B,C,D)$

$= \mathbf{P}(A|B,C,D)\ \mathbf{P}(B|C,D)\ \mathbf{P}(C,D)$

$= \mathbf{P}(A|B,C,D)\ \mathbf{P}(B|C,D)\ \mathbf{P}(C|D)\ \mathbf{P}(D)$

Inference using full joint distributions

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

How to answer questions?

Knowledge base is represented using full joint distribution.

To compute posterior probability of a query proposition given observed evidence, we add up probabilities of possible worlds in which the proposition is true (**marginalization** or **summing out**).

$P(\phi) = \Sigma_{\omega:\omega|=\phi}\ P(\omega)$

$\mathbf{P}(\mathbf{Y}) = \Sigma_{z \in \mathbf{Z}}\ \mathbf{P}(\mathbf{Y},\mathbf{z})$

Example of probabilistic inference

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

$P(\phi) = \Sigma_{\omega:\omega|=\phi} P(\omega)$

$\mathbf{P}(\mathbf{Y}) = \Sigma_{z \in \mathbf{Z}} \mathbf{P}(\mathbf{Y}, z)$

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

P(cavity ∨ toothache)
= 0.108+0.012+0.072+0.008+0.016+0.064 = 0.28

P(¬cavity|toothache)
= P(¬cavity ∧ toothache) / P(toothache)
= (0.016 + 0.064) / (0.108 + 0.012 + 0.016 + 0.064)
= 0.4

Normalization

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

P(¬cavity|toothache)
= P(¬cavity ∧ toothache) / P(toothache)
= (0.016 + 0.064) / (0.108 + 0.012 + 0.016 + 0.064) = 0.4

P(cavity|toothache)
= P(cavity ∧ toothache) / P(toothache)
= (0.108 + 0.012) / (0.108 + 0.012 + 0.016 + 0.064) = 0.6

Notice that denominators are identical in both formulas!

We even do not need to know the exact value of denominator:
P(¬cavity|toothache) + P(cavity|toothache) = 1

We can use a **normalization constant** α instead, computed such that the evaluated distribution adds up to 1.

P(Cavity|toothache) = α **P**(Cavity,toothache)
= α [**P**(Cavity,toothache,catch) + **P**(Cavity,toothache,¬catch)]
= α [⟨0.108, 0.016⟩ + ⟨0.012, 0.064⟩]
= α [⟨0.12, 0.08⟩] = [⟨0.6, 0.4⟩]

Inference via enumeration

In a typical case, we know values e of random variables **E** from the **observation** and we are looking for probability distribution of random variables **Y** from the **query**.

The other random variables are **hidden H** = **X** – **Y** – **E**.

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

Some drawbacks of inference by enumeration:

- the **worst-case time complexity is $O(d^n)$,** where d is the number of values in domains of each random variable
- to store full joint probability distribution we need **$O(d^n)$ space**
- last but not least, it is not easy to obtain probabilities for all possible worlds

Independence

Let us expand the full joint distribution by adding a fourth variable Weather with the domain {cloudy, sunny, rain, snow} – the new full joint distribution has 2x2x2x4 = 32 elements (possible worlds).

P(toothache,catch,cavity,cloudy)
*= P(cloudy|toothache,catch,cavity) * P(toothache,catch,cavity)*

Do one's dental problems influence the weather?
P(cloudy|toothache,catch,cavity) = P(cloudy)

We can write in general:

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

Hence the full joint distribution can be constructed from two smaller tables, one with 8 elements and one with 4 elements.

This property is called **(absolute) independence**:

P(X|Y) = **P**(X) or **P**(Y|X) = **P**(Y) or **P**(X,Y) = **P**(X).**P**(Y)

Conditional independence

Full independence allows us **reducing the size** of the domain representation, but unfortunately full independence is rare and even independent subsets can be quite large.

When one has cavity, does catch depend on toothache?

P(catch | toothache, cavity) = P(catch | cavity)

P(catch | toothache, ¬ cavity) = P(catch | ¬ cavity)

Random variables Catch and Toothache are independent if we know the value of Cavity.

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

This property is called **conditional independence**:

P(X|Y,Z) = **P**(X|Y) or **P**(Z|X,Y) = **P**(Z|Y) or
P(Z,X|Y) = **P**(Z|Y) **P**(X|Y)

Exploiting conditional independence

Conditional independence can be used to further reduce the size of domain representation.

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

The full joint distribution can be constructed from three smaller tables of sizes 2 + 2 + 1 = 5 (only independent elements are represented).

Diagnostic systems

Let us go back to **diagnostic problems**.

Usually we are looking for disease (the source of problems) based on symptoms (observations).

- we are interested in the **diagnostic direction** expressed as conditional probability P(disease|symptoms)

However, from past experience we often have other information:

- the probability of disease P(disease)
- the probability of symptoms P(symptoms)
- the **causal relation** expressed as conditional probability P(symptoms|disease)

How can this information be exploited to get the probability of the diagnostic direction?

Bayes' rule

Recall the product rule

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

We can deduce a so called **Bayes' rule** (law or theorem):

$P(a|b) = P(b|a)\,P(a) / P(b)$

in general:

$$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\,\mathbf{P}(Y) / \mathbf{P}(X) = \alpha\,\mathbf{P}(X|Y)\,\mathbf{P}(Y)$$

It looks like two steps backward as now we need to know **P**(X|Y), **P**(Y), **P**(X).

But these are the values that we frequently have.

$P(\text{cause}|\text{effect}) = P(\text{effect}|\text{cause})\,P(\text{cause}) / P(\text{effect})$

- P(effect|cause) describes the **causal direction**
- P(cause|effect) describes the **diagnostic relation**

Using Bayes' rule

Medical diagnosis

- from past cases we know P(symptoms|disease), P(disease), P(symptoms)
- for a new patient we know symptoms and looking for diagnosis P(disease|symptoms)

Example:

- meningitis causes a stiff neck 70% of the time
- the prior probability of meningitis is 1/50 000
- the prior probability of stiff neck is 1%

What is the probability that a patient having a stiff neck has meningitis?

$P(m|s) = P(s|m).P(m) / P(s) = 0.7 * 1/50000 / 0.01 = 0.0014$

Why the conditional probability for the diagnostic direction is not stored directly?

- diagnostic knowledge is often more fragile than causal knowledge
- for example, if there is a sudden epidemic of meningitis, the unconditional probability of meningitis P(m) will go up so P(m|s) should also go up while the causal relation P(s|m) is unaffected by the epidemic, as it reflects how meningitis works

Naive Bayes model

What if there are more observations?

We can exploit conditional independence as follows

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

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

If all the effects are conditionally independent given the cause variable, we get:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i|Cause)$$

Such a probability distribution is called a **naive Bayes model** (it is often used even in cases where the "effect" variables are not actually conditionally independent given the value of the cause variable).

The Wumpus world revisited

Wumpus is back!

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

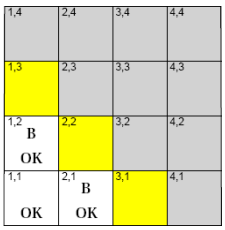
We have a maze with pits that are detected in neighboring squares via breeze (Wumpus and gold will not be assumed now).

Where does the agent should go, if there is breeze at (1,2) and (2,1)?

Pure logical inference can conclude nothing about which square is most likely to be safe!

To which square does the agent should go?

Wumpus: probabilistic model



Boolean variables:

$P_{i,j}$ – pit at square (i,j)

$B_{i,j}$ – breeze at square (i,j)
(only for the observed squares $B_{1,1}$, $B_{1,2}$ a $B_{2,1}$.

Full joint probability distribution

$\mathbf{P}(P_{1,2},\ldots,P_{4,4},B_{1,1},B_{1,2},B_{2,1})$

product rule

$= \mathbf{P}(B_{1,1},B_{1,2},B_{2,1} \mid P_{1,2},\ldots,P_{4,4},) * \mathbf{P}(P_{1,2},\ldots,P_{4,4})$

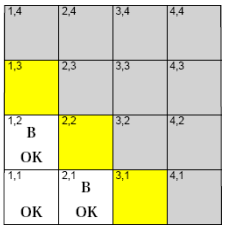
$\mathbf{P}(P_{1,2},\ldots,P_{4,4}) = \Pi_{i,j}\ \mathbf{P}(P_{i,j})$

pits are spread independently

$\mathbf{P}(P_{1,2},\ldots,P_{4,4}) = 0.2^{n} * 0.8^{16-n}$

probability of pit is 0.2 and there are *n* pits

Wumpus: query and simple reasoning



Assume that we have **evidence**:

$b = b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

We are interested in answering **queries** such as $\mathbf{P}(P_{1,3} \mid known, b)$.

Answer can be computed by enumeration of the full joint probability distribution.

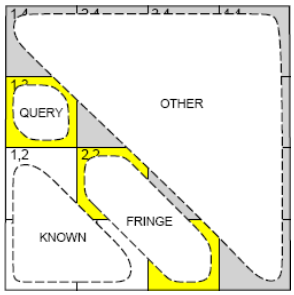
Let Unknown be the variables $P_{i,j}$ except $P_{1,3}$ and Known:

$\mathbf{P}(P_{1,3} \mid known, b)$
$= \Sigma_{unknown}\ \mathbf{P}(P_{1,3}, unknown, known, b)$

But it means to explore all possible values of variables Unknown and there are $2^{12} = 4096$ terms!

Can we do it better (faster)?

Wumpus: conditional independence



Observation:

The observed breezes are conditionally independent of the other variables given the known (white), frontier (yellow), and query variables.

We split the set of hidden variables into fringe and other variables:

Unknown = Fringe ∪ Other

From conditional independence we get:

$P(b \mid P_{1,3}, known, unknown) = P(b \mid P_{1,3}, known, fringe)$

Now, let us exploit this formula.

Wumpus: reasoning

$\mathbf{P}(P_{1,3} \mid known, b)$

$= \alpha \Sigma_{unknown} \mathbf{P}(P_{1,3}, known, unknown, b)$

product rule $\mathbf{P}(X,Y) = \mathbf{P}(X|Y)\ \mathbf{P}(Y)$

$= \alpha \Sigma_{unknown} \mathbf{P}(b \mid P_{1,3}, known, unknown) * \mathbf{P}(P_{1,3}, known, unknown)$

$= \alpha \Sigma_{fringe}\Sigma_{other} \mathbf{P}(b \mid P_{1,3}, known, fringe, other) * \mathbf{P}(P_{1,3}, known, fringe, other)$

$= \alpha \Sigma_{fringe}\Sigma_{other} \mathbf{P}(b \mid P_{1,3}, known, fringe) * \mathbf{P}(P_{1,3}, known, fringe, other)$

$= \alpha \Sigma_{fringe} \mathbf{P}(b \mid P_{1,3}, known, fringe) * \Sigma_{other} \mathbf{P}(P_{1,3}, known, fringe, other)$

$= \alpha \Sigma_{fringe} \mathbf{P}(b \mid P_{1,3}, known, fringe) * \Sigma_{other} \mathbf{P}(P_{1,3}) P(known) P(fringe) P(other)$

$= \alpha\ P(known) \mathbf{P}(P_{1,3}) \Sigma_{fringe} \mathbf{P}(b \mid P_{1,3}, known, fringe)\ P(fringe)\ \Sigma_{other}\ P(other)$

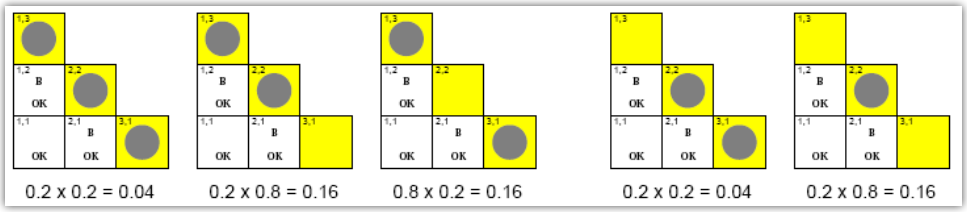
$= \alpha' \mathbf{P}(P_{1,3}) \Sigma_{fringe} \mathbf{P}(b \mid P_{1,3}, known, fringe)\ P(fringe)$

$\alpha' = \alpha .\ P(known)$
$\Sigma_{other}\ P(other) = 1$

Wumpus: solution

$\mathbf{P}(P_{1,3} \mid known, b) = \alpha' \mathbf{P}(P_{1,3}) \Sigma_{fringe} \mathbf{P}(b \mid P_{1,3}, known, fringe)\ P(fringe)$

Let us explore possible models (values) of Fringe that are compatible with observation b.



$\mathbf{P}(P_{1,3} \mid known, b)$

$= \alpha' \langle 0.2\ (0.04 + 0.16 + 0.16), 0.8\ (0.04 + 0.16) \rangle$

$= \langle\ 0.31, 0.69\ \rangle$

$\mathbf{P}(P_{2,2} \mid known, b) = \langle\ 0.86, 0.14\ \rangle$

Definitely avoid the square (2,2)!

Probability theory is a formal mechanism to handle **uncertainty**.

Full joint distribution describes probabilities of all possible worlds.

Answers to queries can be obtained by summing out probabilities of possible worlds consistent with the observation.

However, larger problems will require a better approach.

We are going to exploit **independence** and **conditional independence**.

Department of Theoretical Computer Science and Mathematical Logic
bartak@ktiml.mff.cuni.cz