Logical agents

Chapter 7

Outline

- Knowledge-based agents
- Wumpus world
- Logic in general—models and entailment
- Propositional (Boolean) logic
- Equivalence, validity, satisfiability
- Inference rules and theorem proving
  - forward chaining
  - backward chaining
  - resolution

A simple knowledge-based agent

function KB-AGENT() -> action
static: KB, a knowledge base
i, a counter, initially 0, indicating time

Tell(KB, MAKE-PECEPT-SENTENCE(percept, i))
action <- Ask(KB, MAKE-ACTION-QUERY(i))
Tell(KB, MAKE-ACTION-SENTENCE(action, i))
i <- i + 1
return action

The agent must be able to:
- Represent states, actions, etc.
- Incorporate new percepts
- Update internal representations of the world
- Deduce hidden properties of the world
- Deduce appropriate actions

Knowledge bases

- Inference engine: domain-independent algorithms
- Knowledge base: domain-specific content

Knowledge base = set of sentences in a formal language

Declarative approach to building an agent (or other system):
  TELL it what it needs to know

Then it can ASK itself what to do—answers should follow from the KB

Agents can be viewed at the knowledge level
  i.e., what they know, regardless of how implemented

Or at the implementation level
  i.e., data structures in KB and algorithms that manipulate them

Wumpus world PEAS description

Performance measure
  gold +1000, death -1000
  -1 per step, -10 for using the arrow

Environment
  Squares adjacent to wumpus are smelly
  Squares adjacent to pit are breezy
  Glitter iff gold is in the same square
  Shooting kills wumpus if you are facing it
  Shooting uses up the only arrow
  Grabbing picks up gold if in same square
  Releasing drops the gold in same square

Actuators
  Left turn, Right turn,
  Forward, Grab, Release, Shoot

Sensors
  Breeze, Glitter, Smell

Wumpus world characterization

Observable??
Wumpus world characterization

Observable??  No—only local perception
Deterministic?? Yes—outcomes exactly specified
Episodic??  No—sequential at the level of actions
Static?? Yes—Wumpus and Pits do not move
Discrete?? Yes

Single-agent?? Yes—Wumpus is essentially a natural feature
<table>
<thead>
<tr>
<th>OK</th>
<th>OK</th>
<th>OK</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OK</th>
<th>P?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OK</th>
<th>S</th>
<th>OK</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OK</th>
<th>P?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OK</th>
<th>P?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OK</th>
<th>S</th>
<th>OK</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>
Exploring a wumpus world

Logics are formal languages for representing information such that conclusions can be drawn

Syntax defines the sentences in the language

Semantics define the “meaning” of sentences; i.e., define truth of a sentence in a world

E.g., the language of arithmetic

$x + 2 \geq y$ is a sentence; $x2 + y >$ is not a sentence

$x + 2 \geq y$ is true if the number $x + 2$ is no less than the number $y$

$x + 2 \geq y$ is true in a world where $x=7, y=1$

$x + 2 \geq y$ is false in a world where $x=0, y=6$

Entailment

Entailment means that one thing follows from another:

$KB \models \alpha$

Knowledge base $KB$ entails sentence $\alpha$ if and only if $\alpha$ is true in all worlds where $KB$ is true

E.g., the KB containing “the Giants won” and “the Reds won” entails “Either the Giants won or the Reds won”

E.g., $x + y = 4$ entails $i=x + y$

Entailment is a relationship between sentences (i.e., syntax) that is based on semantics

Note: brains process syntax (of some sort)

Models

Logicians typically think in terms of models, which are formally structured worlds with respect to which truth can be evaluated

We say $m$ is a model of a sentence $\alpha$ if $\alpha$ is true in $m$

$M(\alpha)$ is the set of all models of $\alpha$

Then $KB \models \alpha$ if and only if $M(KB) \subseteq M(\alpha)$

E.g. $KB =$ Giants won and Reds won $\alpha =$ Giants won

Other tight spots

Breeze in (1,2) and (2,1) ⇒ no safe actions

Assuming pits uniformly distributed, (2,2) has pit w/ prob 0.86, vs. 0.31

Smell in (1,1)

⇒ cannot move

Can use a strategy of coercion:

shoot straight ahead

wumpus was there ⇒ dead ⇒ safe

wumpus wasn’t there ⇒ safe
Entailment in the wumpus world

Situation after detecting nothing in [1,1], moving right, breeze in [2,1]

Consider possible models for ?s assuming only pits

3 Boolean choices ⇒ 8 possible models

\[ KB = \text{wumpus-world rules} + \text{observations} \]

\( \alpha_1 = \text{"[1,2] is safe"}, \ KB \models \alpha_1, \text{proved by model checking} \)

\[ KB = \text{wumpus-world rules} + \text{observations} \]

\( \alpha_2 = \text{"[2,2] is safe"}, \ KB \nmodels \alpha_2 \)
Entailment

If $KB \vdash \alpha$, sentence $\alpha$ can be derived from $KB$ by procedure $i$.

Consequences of $KB$ are a haystack; $\alpha$ is a needle.

Entailment = needle in haystack; inference = finding it.

Soundness: $i$ is sound if whenever $KB \vdash \alpha$, it is also true that $KB \models \alpha$.

Completeness: $i$ is complete if whenever $KB \models \alpha$, it is also true that $KB \vdash \alpha$.

Preview: we will define a logic (first-order logic) which is expressive enough to say almost anything of interest, and for which there exists a sound and complete inference procedure.

That is, the procedure will answer any question whose answer follows from what is known by the $KB$.

Propositional logic: Syntax

Propositional logic is the simplest logic—illustrates basic ideas.

The proposition symbols $P_1, P_2$ etc are sentences.

If $S$ is a sentence, $\neg S$ is a sentence (negation).

If $S_1$ and $S_2$ are sentences, $S_1 \land S_2$ is a sentence (conjunction).

If $S_1$ and $S_2$ are sentences, $S_1 \lor S_2$ is a sentence (disjunction).

If $S_1$ and $S_2$ are sentences, $S_1 \Rightarrow S_2$ is a sentence (implication).

If $S_1$ and $S_2$ are sentences, $S_1 \iff S_2$ is a sentence (biconditional).

Propositional logic: Semantics

Each model specifies true/false for each proposition symbol.

E.g. $P_{i,2}$, $P_{i,2}$, $P_{i,1}$

true true false

(With these symbols, 8 possible models, can be enumerated automatically.)

Rules for evaluating truth with respect to a model $m$:

$\neg S$ is true iff $S$ is false.

$S_1 \land S_2$ is true iff $S_1$ is true and $S_2$ is true.

$S_1 \lor S_2$ is true iff $S_1$ is true or $S_2$ is true.

i.e., $S_1 \land \neg S_2$ is true iff $S_1$ is false and $S_2$ is true.

$S_1 \Rightarrow S_2$ is true iff $S_1$ is false or $S_2$ is true.

$S_1 \iff S_2$ is true iff $S_1$ is true and $S_2$ is true.

Simple recursive process evaluates an arbitrary sentence, e.g.,

$\neg P_{i,2} \land (P_{i,2} \lor P_{i,1})$ = true \land (false \lor true) = true \land true = true.

Truth tables for connectives

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$\neg P$</th>
<th>$P \land Q$</th>
<th>$P \lor Q$</th>
<th>$P \Rightarrow Q$</th>
<th>$P \Leftrightarrow Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
</tr>
</tbody>
</table>

Wumpus world sentences

Let $P_{i,j}$ be true if there is a pit in $[i,j]$.

Let $B_{i,j}$ be true if there is a breeze in $[i,j]$.

$\neg P_{i,j}$

$\neg B_{i,j}$

$B_{i,j}$

"Pits cause breezes in adjacent squares"
Truth tables for inference

<table>
<thead>
<tr>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
<th>$R_4$</th>
<th>$R_5$</th>
<th>$K_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>

Enumerate rows (different assignments to symbols), if $K_B$ is true in row, check that $\alpha$ is too

Validity and satisfiability

A sentence is **valid** if it is true in all models,
- e.g., $\text{True}, \ A \lor \lnot A, \ A \rightarrow A, \ (A \land (A \Rightarrow B)) \Rightarrow B$

Validity is connected to inference via the Deduction Theorem:
- $KB \models \alpha$ if and only if $(KB \Rightarrow \alpha)$ is valid

A sentence is **satisfiable** if it is true in **some** model
- e.g., $A \lor B, \ C$

A sentence is **unsatisfiable** if it is true in **no** models
- e.g., $A \land \lnot A$

Satisfiability is connected to inference via the following:
- $KB \models \alpha$ if and only if $(KB \land \lnot \alpha)$ is unsatisfiable
  - i.e., prove $\alpha$ by reductio ad absurdum

Inference by enumeration

Depth-first enumeration of all models is sound and complete

```python
function TT-ENTAILS?(KB, α) returns true or false
inputs: KB, the knowledge base, a sentence in propositional logic
α, the query, a sentence in propositional logic
symbols ← a list of the proposition symbols in KB and α
return TT-CHECK-ALL(KB, α, symbols, [])

function TT-CHECK-ALL(KB, α, symbols, model) returns true or false
if PL-TRUE?(KB, model) then return PL-TRUE?(α, model)
else return true
else do
    P ← FIRST(symbols); rest ← REST(symbols)
    return TT-CHECK-ALL(KB, α, rest, EXTEND(P, true, model)) and
    TT-CHECK-ALL(KB, α, rest, EXTEND(P, false, model))
```

$O(2^n)$ for $n$ symbols; problem is co-NP-complete

Proof methods

Proof methods divide into (roughly) two kinds:

**Application of inference rules**
- Legitimate (sound) generation of new sentences from old
- **Proof** = a sequence of inference rule applications
  - Can use inference rules as operators in a standard search alg.
  - Typically require translation of sentences into a normal form

**Model checking**
- truth table enumeration (always exponential in $n$)
  - improved backtracking, e.g., Davis–Putnam–Logemann–Loveland
  - heuristic search in model space (sound but incomplete)
    - e.g., min-conflicts-like hill-climbing algorithms

Logical equivalence

Two sentences are **logically equivalent** iff true in same models:
- $\alpha \equiv \beta$ if and only if $KB \models \alpha \land \beta$ and $KB \models \beta \land \alpha$

- $\alpha \land \beta \equiv (\beta \land \alpha)$: commutativity of $\land$
- $\alpha \lor \beta \equiv (\beta \lor \alpha)$: commutativity of $\lor$
- $(\alpha \land \beta) \land \gamma \equiv (\alpha \land (\beta \land \gamma))$: associativity of $\land$
- $(\alpha \lor \beta) \lor \gamma \equiv (\alpha \lor (\beta \lor \gamma))$: associativity of $\lor$
- $\lnot(\alpha) \equiv \alpha$: double-negation elimination
- $(\alpha \Rightarrow \beta) \equiv (\lnot \alpha \lor \beta)$: contraposition
- $(\alpha \Rightarrow \beta) \equiv ((\alpha \Rightarrow \beta) \land (\beta \Rightarrow \alpha))$: biconditional elimination
- $\lnot(\alpha \land \beta) \equiv (\lnot \alpha \lor \lnot \beta)$: De Morgan
- $\lnot(\alpha \lor \beta) \equiv (\lnot \alpha \land \lnot \beta)$: De Morgan
- $(\alpha \land (\beta \lor \gamma)) \equiv ((\alpha \land \beta) \lor (\alpha \land \gamma))$: distributivity of $\land$ over $\lor$
- $(\alpha \lor (\beta \land \gamma)) \equiv ((\alpha \lor \beta) \land (\alpha \lor \gamma))$: distributivity of $\lor$ over $\land$

Forward and backward chaining

**Horn Form (restricted)**
- $KB = \text{conjunction of Horn clauses}$
  - Horn clause:
    - $\Diamond$ proposition symbol; or
    - $\Diamond$ (conjunction of symbols) $\Rightarrow$ symbol
  - E.g., $C \land (B \Rightarrow A) \land (C \land D) \Rightarrow B$

**Modus Ponens** (for Horn Form): complete for Horn KBs
- $\alpha_1, \ldots, \alpha_n, \alpha_1 \land \ldots \land \alpha_n \Rightarrow \beta$
  - $\beta$

Can be used with **forward chaining** or **backward chaining**.
- These algorithms are very natural and run in linear time
Forward chaining

Idea: fire any rule whose premises are satisfied in the $KB$, add its conclusion to the $KB$, until query is found

$$P \Rightarrow Q$$
$$L \land M \Rightarrow P$$
$$B \land L \Rightarrow M$$
$$A \land P \Rightarrow L$$
$$A \land B \Rightarrow L$$
$$A$$
$$B$$

Forward chaining algorithm

```python
function PL-FC-ENTAILS(KB, q) returns true or false
inputs: KB, the knowledge base, a set of propositional Horn clauses
        q, the query, a propositional symbol
local variables: count, a table, indexed by clause, initially the number of premises
                inferred, a table, indexed by symbol, each entry initially false
                agenda, a list of symbols, initially the symbols known in KB
while agenda is not empty do
    p ← POP(agenda)
    unless inferred[p] do
        inferred[p] ← true
        for each Horn clause c in whose premise p appears do
            decrement count[c]
            if count[c] = 0 then do
                if HEAD[c] = q then return true
                PUSH(HEAD[c], agenda)
    return false
```

Forward chaining example
Forward chaining example

Proof of completeness

FC derives every atomic sentence that is entailed by $KB$

1. FC reaches a fixed point where no new atomic sentences are derived
2. Consider the final state as a model $m$, assigning true/false to symbols
3. Every clause in the original $KB$ is true in $m$
   
   Proof: Suppose a clause $a_1 \land \ldots \land a_k \Rightarrow b$ is false in $m$
   
   Then $a_1 \land \ldots \land a_k$ is true in $m$ and $b$ is false in $m$
   
   Therefore the algorithm has not reached a fixed point!

4. Hence $m$ is a model of $KB$
5. If $KB \models q$, $q$ is true in every model of $KB$, including $m$

   General idea: construct any model of $KB$ by sound inference, check $\alpha$

Backward chaining

Idea: work backwards from the query $q$:

   to prove $q$ by BC,
   
   check if $q$ is known already, or
   
   prove by BC all premises of some rule concluding $q$

Avoid loops: check if new subgoal is already on the goal stack

Avoid repeated work: check if new subgoal

1) has already been proved true, or
2) has already failed
Forward vs. backward chaining

FC is data-driven, cf. automatic, unconscious processing, e.g., object recognition, routine decisions

May do lots of work that is irrelevant to the goal

BC is goal-driven, appropriate for problem-solving, e.g., Where are my keys? How do I get into a PhD program?

Complexity of BC can be much less than linear in size of KB
Resolution

Conjunctive Normal Form (CNF—universal)

conjunction of disjunctions of literals

\[ E.g., (A \lor B) \land (B \lor C \land \neg D) \]

Resolution inference rule (for CNF): complete for propositional logic

\[ \ell_1 \lor \ldots \lor \ell_k \land m_1 \lor \ldots \lor m_n \]

where \( \ell_i \) and \( m_j \) are complementary literals. E.g.,

\[ P_{1,1} \lor P_{2,1}, \neg P_{2,2} \]

Resolution is sound and complete for propositional logic

Resolution example

\[ KB = (B_{1,1} \equiv (P_{1,2} \lor P_{2,1})) \land \neg B_{1,1} \land \neg P_{1,2} \]

Conversion to CNF

\[ B_{1,1} \equiv (P_{1,2} \lor P_{2,1}) \]

1. Eliminate \( \equiv \), replacing \( \alpha \equiv \beta \) with \( (\alpha \Rightarrow \beta) \land (\beta \Rightarrow \alpha) \).

\[ (B_{1,1} \equiv (P_{1,2} \lor P_{2,1})) \land (P_{1,2} \lor P_{2,1} \Rightarrow B_{1,1}) \]

2. Eliminate \( \equiv \), replacing \( \alpha \Rightarrow \beta \) with \( \neg \alpha \lor \beta \).

\[ (\neg B_{1,1} \lor P_{1,2} \lor P_{2,1}) \land (\neg (P_{1,2} \lor P_{2,1}) \lor B_{1,1}) \]

3. Move \( \neg \) inwards using de Morgan’s rules and double-negation:

\[ (\neg B_{1,1} \lor P_{1,2} \lor P_{2,1}) \land ((\neg P_{1,2} \land \neg P_{2,1}) \lor B_{1,1}) \]

4. Apply distributivity law (\( \lor \) over \( \land \)) and flatten:

\[ (\neg B_{1,1} \lor P_{1,2} \lor P_{2,1}) \land (\neg P_{1,2} \lor B_{1,1}) \land (\neg P_{2,1} \lor B_{1,1}) \]

Summary

Logical agents apply inference to a knowledge base
to derive new information and make decisions

Basic concepts of logic:
- syntax: formal structure of sentences
- semantics: truth of sentences wrt models
- entailment: necessary truth of one sentence given another
- inference: deriving sentences from other sentences
- soundness: derivations produce only entailed sentences
- completeness: derivations can produce all entailed sentences

Wumpus world requires the ability to represent partial and negated information, reason by cases, etc.

Forward, backward chaining are linear-time, complete for Horn clauses
Resolution is complete for propositional logic

Propositional logic lacks expressive power

Resolution algorithm

Proof by contradiction, i.e., show \( KB \land \neg \alpha \) unsatisfiable

```
function PL-RESOLUTION(KB, \alpha) returns true or false
inputs: KB, the knowledge base, a sentence in propositional logic
        \alpha, the query, a sentence in propositional logic
        clauses -- the set of clauses in the CNF representation of KB \land \neg \alpha
        new -- []
        loop do
            for each C_i, C_j in clauses do
                resolvents = PL-RESOLVE(C_i, C_j)
                if resolvents contains the empty clause then return true
                new = new \cup resolvents
            if new \subseteq clauses then return false
            clauses = clauses \cup new
```

Chapter 7