AI: from intuitive ideas to a formal framework

Ayoub El Hanchi

McGill University

April 1, 2019

Intelligent agent

- Agent: an entity that has a goal and that is able to interact with its environment.
- A rational agent: an agent that acts rationally, i.e. an agent that acts optimally in order to maximize its satisfaction, given the evidence at its disposal.
- One way to define an intelligent agent is therefore to associate it with a rational agent (although this may not be the only way).

Some Notation

- $S = \{S_1, S_2, ..., S_n\}$ the possible states of the environment.
- $D = \{D_1, D_2, ..., D_m\}$ the possible decisions that the agent can make.
- $L : S \times D \to \mathbb{R}^+$ a loss function that maps a current state of the environment and a decision to a positive real number, representing the amount of "discontent" the agent experiences by taking that decision in the specified environment.

A simple case

- Suppose that the agent knows its current environment, S_1 say. Then for it to act rationally, it must make the decision D_k satisfying:
$$L(S_1, D_k) = \min_{j=1,\dots,m} L(S_1, D_j)$$
- The agent will then be using all relevant information (the state of the environment) and will be acting optimally (by minimizing its loss).
- What if the agent does not know the current state of the environment ?

A more complex case

- Suppose now that the agent does not know its current environment, but it has some prior information about it, which we denote by I, and some new evidence from the environment, which we denote by E.
- For now, we will think of both I and E as a list of facts about the environment. More formally, we can think of them as sentences in propositional logic.

A more complex case

- In this scenario, we can attempt to logically infer the state of the environment from the prior information I, and new evidence E.
- Once the state of the environment is inferred, we can then choose the optimal decision as in the first case.

The realistic case

- What if the available information does not allow us to infer the state of the environment ?
- The information I and the evidence E contain nonetheless useful information that the agent must use in making its decision, otherwise it would not act rationally by our definition.
- We therefore need a way to express the plausibility of each state of the environment, in light of the information that the agent has.

A logic for uncertainty

- Our goal therefore is to define a mathematical framework in which we can assign plausibilities in a way consistent with what we perceive as rational.
- This framework should, among other things, allow us to combine our prior information with the evidence we collect to make the best decision possible.
- As we will see, if we impose quite simple and reasonable conditions, which coincide with our idea of rationality, we obtain a unique such framework.

The conditions

1. The plausibility of a proposition A depends on the information I we have related to it. (Dependence)
2. The plausibility of a proposition A given information I is a real number, and is denoted by $(A|I)$. (Comparability)
3. Agreement with propositional logic. For example, if $(A \Leftrightarrow B)$ then $(A|I) = (B|I)$ (Common sense)
4. If the plausibility of a proposition A given information I can be derived in multiple ways, then all results must be equal. (Consistency)

Cox's Theorem (1946, 1961)

Cox's theorem asserts that given the previous conditions on the plausability of a proposition, there exists a function $P : \mathbb{R} \to [0, 1]$ satisfying the following conditions:

- if A is known to be true given information I, then $P(A|I) = 1$.
- $P(A + B|I) = P(A|I) + P(B|I) - P(AB|I)$
- $P(AB|I) = P(A|I)P(B|AI) = P(B|I)P(A|BC)$

where AB means (A and B), and $A + B$ means (A or B). We call P a probability.

Consequences of Cox theorem

- An important consequence of the previous theorem is that ANY system which cannot be expressed in probabilistic terms must violate one of the previously mentioned conditions.
- The conditions on the probability P are equivalent to the standard axioms of probability theory in mathematics (with the addition of countable additivity).
- Coming back to our initial problem, now that we have a consistent system to manipulate probabilities, how can we use it to allow the agent to make the optimal decision ?

A general decision making procedure under uncertainty

1. Enumerate the possible states of the environment $\{S_1, ..., S_n\}$.
2. Assign prior probabilities $P(S_i|I)$ which reflect the prior information that the agent has about the current state.
3. Assuming the state S_i, assign the probability $P(E|S_i, I)$ which represents the prior knowledge of the agent about how the evidence is generated.
4. Compute the posterior probabilities of each of the S_i:

$$P(S_i|E, I) = \frac{P(E|S_i, I)P(S_i|I)}{P(E|I)}$$

5. Choose the decision D_k that minimizes the expected loss in light of the prior information I and evidence E:

$$\sum_{i=1}^{n} L(S_i, D_j)P(S_i|E, I)$$

Difficult questions

While the above procedure is very satisfying in both its generality and simplicity, it does not directly address many important questions:

- What if it is not possible to enumerate all possible states of the environment ?
- How do we express our prior knowledge in probability statements ?
- How do we choose our loss function ?
- Is this procedure computationally tractable ?

Parametric models

- It is difficult to give general answers to the questions we posed. Many of them depend on the details of the decision problem.
- Nonetheless, in practice, the most used class of models is that of parametric models, which instead of trying to enumerate all possible states of the environment, considers a set of hypotheses $H = \{H_\theta | \theta \in \mathbb{R}^d\}$.
- In addition to considering this set of hypotheses, one also usually assigns standard distributions to $P(E|H_\theta, I)$ and $P(H_\theta|I)$.

The computational challenge

- At this point, the challenge resides mainly in the computation of the expected loss, which now becomes an integral over R^d.
- Most choices of loss functions and distributions yield analytically intractable integrals, and when d, the dimension of the parameter θ, is large, standard numerical integration become prohibitively expensive.
- For the past year, my research has been about using gradient-based Markov chain Monte Carlo methods to compute this integral for large datasets with high dimensional parameters.

References

[1, 2]

E. T. Jaynes.
Probability theory: The logic of science.
Cambridge University Press, Cambridge, 2003.

Stuart J. Russell and Peter Norvig.
Artificial Intelligence: A Modern Approach.
Pearson Education, 2 edition, 2003.