Decision Making under Uncertainty Project

Sebastian Jobjörnsson
Stefan Walzer

December 19, 2013

1 Main Characteristics of the Agent

- The Agent assumes that the underlying process is an MDP and the observation is the state of this Markov Process (i.e. its not a POMDP).
- For each pair of state and action, we keep track of which subsequent states resulted (how often) from doing that action in that state. This represents at every point in time our knowledge of the process. Mathematically speaking, for each state and action we store a Dirichlet distribution of the following state and we update this belief occording to the observations we make.
- We store for each pair of state and action the average reward that we have received from them.
- Our strategy will with some probability ε_t (explained below) chose a random action (with distribution explained below) which makes sure that we explore the MDP. With $1 - \varepsilon_t$ we will play an action that maximises the expected utility, taking into account the expected reward for the next action and the expected value of the next state which is obtained by the value iteration algorithm.

2 Details

- In the initial state we imagine we had already observed for each pair of state and action exactly one transition to every other state. This initialisation corresponds to a prior Dirichlet distribution on the transition probabilities. We also assume that for each state and action we observed the maximum possible reward once (which is an optimistic intialisation and helps to ensure that many actions are explored).
- The transition probabilities are stored in a HashMap which saves memory if the underlying transition probability function (depending on state, action and next state) is sparse, i.e. if many transitions never occur.

- The probability for choosing a random action depends on the number of times the state was already visited. The idea is of course that there is less need for exploration in states that were already visited very often.
- If we choose to pick a random action, then the probabily of picking an action is proportional to the inverse of the number of times this action was already chosen in that state (+1). Again we want to favour using actions that were hardly used so far.
- The agent reacts to the `freeze learning` and `unfreeze learning` messages.

3 Mixed model for deterministic and stochastic MDPs

Our agent assumes that the environment is a fixed but unknown MDP. This MDP is either deterministic or stochastic. Determinism here means that a given state-action pair always leads to the same state and reward. As long as the observations and rewards obtained by interaction by the environment does not rule out the possibility of a deterministic MDP, the agent keeps and updates probabilities $p(M_D)$ and $p(M_S)$ over the two possible model types M_D and M_S. If, at some time step, an observation or reward is registered that is inconsistent with a previous observation or reward, then the possibility of a deterministic MDP is abandoned, i.e., $p(M_D)$ is set to 0 and $p(M_S)$ is set to 1.

For a deterministic MDP, the parameters that are kept in order to represent the current information state consists of a two-dimensional array ds mapping state-action pairs into fixed states. As soon as a state s has been observed and an action a taken in that state, the resulting state is stored as $ds[s][a]$. When the state action pair (s, a) is encountered in the future, the agent assigns probability 1 to make a transition to $ds[s][a]$ and probability 0 to any other transition (under the assumption of a deterministic MDP). For a state s and an action a such that a has never been executed in s, the subjective probability of the agent of a next state s' is set to be $1/|S|$, where S is the set of states.

The input to the value iteration part of the algorithm, which is used to compute an optimal policy given the current state of information, consists of an estimate of the transition probabilities $p(s' \mid s, a)$ and expected rewards $r(s, a)$ of the unknown MDP. Denoting the history of actions, observations and rewards up to time step t by h_t, this input is computed as

$$p(s' \mid s, a, h_t) = p(M_D \mid h_t)p(s' \mid s, a, h_t, M_D) + p(M_S \mid h_t)p(s' \mid s, a, h_t, M_S)$$
$$r(s, a \mid h_t) = p(M_D \mid h_t)r(s, a \mid h_t, M_D) + p(M_S \mid h_t)r(s, a \mid h_t, M_S).$$

In the above equations, $p(M_D, \mid h_t)$ and $p(M_S, \mid h_t)$ are the posterior probabilities of the respective models given what has been observed so far. By Bayes' rule, these posterior

probabilities may be written as

$$p(M_D \mid h_t) = \frac{p(M_D)p(h_t \mid M_D)}{p(M_D)p(h_t \mid M_D) + p(M_S)p(h_t \mid M_S)}$$
$$p(M_S \mid h_t) = \frac{p(M_S)p(h_t \mid M_S)}{p(M_D)p(h_t \mid M_D) + p(M_S)p(h_t \mid M_S)},$$

which, under the assumption of a prior $p(M_D) = p(M_S) = \frac{1}{2}$, is reduced to

$$p(M_D, \mid h_t) = \frac{p(h_t \mid M_D)}{p(h_t \mid M_D) + p(h_t \mid M_S)}$$
$$p(M_S \mid h_t) = \frac{p(h_t \mid M_S)}{p(h_t \mid M_D) + p(h_t \mid M_S)}.$$

There are now two cases. Either the history h_t is consistent with a deterministic model, or it is not. If it is not, then we immediately get that $p(h_t \mid M_D) = 0$, and therefore, after this time point, all computations will only involve the Dirichlet-Multinomial model for a stochastic MDP. Assume now that the history is consistent with a deterministic MDP (and also, of course, with a stochastic MDP). Then, for $p(h_t \mid M_D)$, we have

$$p(h_t \mid M_D) = \sum_{\eta} p(h_t \mid \eta, M_D)p(\eta \mid M_D).$$

In the above, each particular value of η corresponds to a specific specification of the transition table of a deterministic MDP. Since, for a state space S and action set A, there are $|S|^{|S||A|}$ such tables, under the assumption of a uniform prior, we have $p(\eta \mid M_D) = |S|^{-|S||A|}$. Further, for each fixed h_t, $p(h_t \mid \eta, M_D)$ will equal 1 if h_t is consistent with the transitions specified by η and 0 otherwise. Therefore, the value of the sum above is determined by the number values η for which this holds. Letting k be equal to the number of unique transitions in h_t, we get, since each such transition fixes a particular parameter in η, that

$$p(h_t \mid M_D) = \sum_{\eta} p(h_t \mid \eta, M_D)p(\eta \mid M_D)$$
$$p(h_t \mid M_D) = |S|^{-|S||A|} \sum_{\eta} p(h_t \mid \eta, M_D)$$
$$p(h_t \mid M_D) = |S|^{-|S||A|}|S|^{|S||A|-k} = |S|^{-k}.$$

Similarly, in the special case that h_t is consistent with a deterministic MDP, one may easily compute the following expression for $p(h_t \mid M_S)$, by making use of the assumed Dirichlet-Multinomial model:

$$p(h_t \mid M_S) = \prod_{s,a} \frac{Q_{s,a}!}{|S|(|S|+1)\dots(|S|+Q_{s,a}-1)},$$

where $Q_{s,a}$ is the number of times a transition has been observed for the state-action pair (s, a). The expressions derived for $p(h_t \mid M_D)$ and $p(h_t \mid M_D)$ may now be used to compute the posterior model probabilities $p(M_D, \mid h_t)$ and $p(M_S, \mid h_t)$.

$$p(M_D, \mid h_t) = \frac{1}{1 + |S|^k \prod_{s,a} \frac{Q_{s,a}!}{|S|(|S|+1)\ldots(|S|+Q_{s,a}-1)}}$$
$$p(M_S \mid h_t) = 1 - p(M_D, \mid h_t).$$

Algorithmically, this may be done in a stepwise manner. All that is required is to keep track of $Q_{s,a}$. I our code, an array was used, and this was updated after each action-observation step.