Mixed model for deterministic and stochastic MDPs

Sebastian Jobjörnsson
Stefan Walzer

December 19, 2013

1 General Idea: Two Underlying Models

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 with 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.

2 Deterministic Model M_D

For the deterministic model M_D, 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 considered in the future, the M_D assigns probability 1 to the transition to `ds[s][a]` and probability 0 to any other transition. 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 all states. Similarly to the transition probabilities, the observed rewards are stored in an array `rD[s][a]`. As long as the state action pair has not been encountered, the model predicts the maximum possible reward.

3 Stochastic Model

The stochastic model has a Dirichlet distribution with parameter $\vec{\alpha} \equiv 1$ as a prior and updates its beliefs according to the observations. To store which transitions occurred (and how often they occurred) we use a two dimensional array of hash tables, one hash table for each state action pair. Assuming that the set of possible transitions is sparse (in the set of all conceivable transitions) this means a significant improvement compared to the naive approach of using a three dimensional array since a transition that does not happen will just be represented by the corresponding entry not existing in the hash table.

If a certain state action pair has been encountered n times with an average reward of r then the model predicts a reward of $(n \cdot r + r_{max})/(n + 1)$where r_{max} is the maximum possible reward. This overestimates the reward in early stages and is supposed to lead to a better exploration of the MDP.

4 Value Iteration

The input to the value iteration part of the algorithm, which is used to compute an optimal policy given our current beliefs, 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

$$\begin{aligned} p(s' \mid s, a, h_t) &= p(M_D \mid h_t) \cdot p(s' \mid s, a, h_t, M_D) + p(M_S \mid h_t) \cdot p(s' \mid s, a, h_t, M_S) \\ r(s, a \mid h_t) &= p(M_D \mid h_t) \cdot r(s, a \mid h_t, M_D) + p(M_S \mid h_t) \cdot r(s, a \mid h_t, M_S). \end{aligned}$$

We have already mentioned how the models M_D and M_S estimate rewards and transition probabilities. In the following we outline how we calculate our beliefs about what the correct model is.

5 Probability for the Models given the data

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

$$\begin{aligned} p(M_D \mid h_t) &= \frac{p(M_D)p(h_t \mid M_D)}{p(M_D)p(h_t \mid M_D) + p(M_S)p(h_t \mid M_S)} \\ p(M_S \mid h_t) &= \frac{p(M_S)p(h_t \mid M_S)}{p(M_D)p(h_t \mid M_D) + p(M_S)p(h_t \mid M_S)} \end{aligned}$$

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 $p(M_S \mid h_t)$.

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

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.

6 The ε-Greedy Aspect

Orthogonal to what has been said before, our agent will with some probability ε_t instead of playing according to the results of the value iteration choose a random action instead which is supposed to ensure better exploration of the MDP.

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 probability of picking an action is proportional to the inverse of the number of times this action was already chosen in that state (plus one). Again we want to favour using actions that were hardly used so far.