Decision Making under Uncertainty Project

Sebastian Jobjörnsson
Stefan Walzer

December 16, 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.