Machine Learning - Assignment 5

Posted Friday, November 9, 2007
Due Friday, November 16, 2007

1. [30 points] **Defining Rewards**

   (a) [10 points] Suppose you are training a robot to find its way from the main entrance of McConnell to room 103. You give a reward of $+1$ for reaching the classroom and a reward of 0 at all other times. Suppose you treat this as an episodic task and set up the robot to maximize the expected sum of rewards. After running your learning algorithm for some time, you note that there is no improvement in the time taken to reach the classroom. What do you think is going wrong? Describe what changes you could make to this setup to fix the problem.

   (b) [10 points] What happens if you add a constant C to all rewards in the task above, and still ask the robot to maximize the sum of rewards? Will the optimal policy change or not?

   (c) [10 points] What happens if you add C to all rewards and ask the robot to maximize the sum of discounted rewards, with a discount factor $\gamma < 1$?

2. [30 points] **Defining a Markov Decision Process**

   You are hired by NASA to work on the team that is going to program the Mars rover for the next space mission. The Rover has a laser range finder sensor, which tells it the distance to different objects, a camera which can be oriented at different angles, and a gripper which can be used to pick up rocks. A rough chemical analysis can be performed even before a rock is picked up, if the rover is close enough. Picking up rocks costs more energy than not picking them. The rover has a limited energy supply. Formulate this problem as an MDP, specifying what are the states, actions, rewards, transition probabilities, and discount factor, if applicable. You do not need to give actual numbers, just explain qualitatively what these would be. Is the MDP formulation adequate in this case? Justify your answer.

3. [40 points] **Action-value functions**

   Consider the action-value function for an MDP, Q^π, defined as follows:

   $$Q^\pi(s,a) = E_\pi\left[r_{t+1} + \gamma r_{t+2} + \ldots | s_t = s, a_t = a\right]$$

   (a) [10 points] Show that Q^π satisfies the following set of equations:

   $$Q^\pi(s,a) = r_s^a + \sum_{s' \in S} p_{ss'}^a \sum_{a' \in A} \pi(s',a') Q^\pi(s',a')$$

   (b) [10 points] Write these equations in matrix-vector form and give a model-based learning algorithm for computing Q^π (analogously to what we discussed in class for V^π).

   (c) [10 points] Give a temporal-difference learning algorithm for estimating Q^π from data.

(d) [Extra credit 10 points] Prove that this algorithm converges to the correct action-value function

(e) [10 points] Explain how function approximation can be used to solve this problem (give the function of the approximator and a planning or learning algorithm for this case.