COMP-424 - Assignment 3

Posted Monday March 11, 2013
Due Wednesday March 20, 2013

1. [25 points] **Utility theory (adapted from J Pearl)**

   You are trying to decided if you want to buy a used car, which costs \$1500. If the car is in good shape, its market value is \$2000. If it is in bad shape, you will need to fix it at a cost of \$700, after which it will be in good shape. Your initial guess is that the car has 70% chance of being in good shape.

   (a) [5 points] Compute the expected value of buying the car

   (b) [10 points] Suppose you could take the car to the mechanic. If the car is in good shape, he will tell you that it is in good shape with probability 0.8. If the car is in bad shape he will still tell you that it is in good shape with probability 0.25. Compute the posterior probability of the car being in good or bad shape given each possible answer from the mechanic.

   (c) [5 points] Compute how much you should pay the mechanic for his test (rounded to the nearest dollar).

   (d) [5 points] Give the optimal choice of action conditioned on the mechanic's answer.

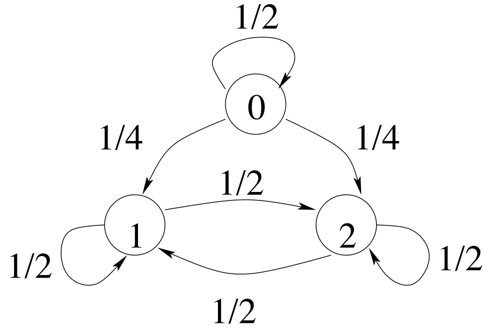
2. [15 points] **Bandit problems**

   (a) [5 points] Suppose that you use the optimistic initialization strategy for the action values. Is the learning algorithm which acts always greedily always guaranteed to converge to the optimal action? Justify your answer.

   (b) [10 points] Suppose you have a 2-armed bandit with one arm having a uniform distribution of rewards in $[-1, 1]$ and the other arm having a uniform distribution of rewards in $[-0.1, +0.12]$. Explain the difference between UCB and ϵ-greedy in how they will allocate their trials between the two arms.

3. [20 points] **Markov chains**

   Consider the simple Markov chain represented below:



(a) Suppose you start in state 0, at $t = 0$. What is the probability distribution over states at time step $t = 2$?

(b) What is the probability of being in state 0 at time step 5?

(c) What is the expected time until state 1 is reached first?

(d) If we let the chain run for an infinite amount of time, what will the probability over states be (this is called the stationary distribution of the chain, but for this particular example it is intuitively easy to determine, without knowing any further theory beyond what we discussed in class)

4. [15 points] **Defining rewards and returns**

   (a) [5 points] Suppose you are training a robot to find its way from the main entrance of McConnell to room 111N. You give a reward of +1 for reaching the room and a reward of 0 at all other times, and there is no discounting ($\gamma = 1$). Suppose you ask the robot to maximize the expected sum of rewards. After running your learning algorithm for some time, you notice 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) [5 points] What happens if you add a constant C to all rewards in the task above (the 1 and the 0s), and still ask the robot to maximize the sum of rewards? Will the optimal policy change or not?

   (c) [5 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 $\gamma < 1$?

5. [10 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.

6. [15 points] **Action-value function**

   We talked in class about the state-value function of an MDP, V^π. Now suppose we wanted a value function which depends on the action choice, as we use in bandit problems:

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

   In other words, on the very first time step you pick action a, then forever follow policy π.

(a) [10 points] Following a procedure similar to what we did in class, prove the following Bellman equation for action values:

$$Q^\pi(s, a) = r_a(s) + \gamma \sum_{s'} T_a(s, s') \sum_{a'} \pi(s', a') Q^\pi(s', a')$$

(b) [5 points] Give an iterative algorithm for computing Q^π based on the equation above