Probabilistic Reasoning in AI - Assignment 6

Due Thursday, April 8

1. [20 points] **Expected Utility**

   Data the android has been chosen to represent Starfleet in a power contest against the best Romulan android. The prize is $100 (old monetary units used on Earth in the 21st century). Data estimates that he has a 0.9 chance of winning if he is functioning correctly.

   The night before the contest, Data observes a malfunction in his controls. Based on a quick diagnosis, Lt. Geordi LaForge tells him that there are three possible causes for the problem. With probability 0.3, Data has a damaged neurotramsmitter. In this case, his chance of winning drops to 50%, and he might sustain further damage, valued at $70. If he does not enter the contest, the repair is valued at $20. With probability 0.2, this is just a minor problem, which will not appear again. In this case, Data's chances stay the same. With probability 0.3, there is a software malfunction, in which case Data will lose for sure. Geordi can later fix the problem at no charge.

   (a) [2 points] Draw the decision graph for this problem.

   (b) [5 points] Suppose that there is an entrance fee for the contest. How much should Data be willing to pay to enter?

   (c) [8 points] The Ferengi have a software analyzer that can detect whether the software is ok or not. How much should Data be willing to pay for using the analyzer?

   (d) [5 points] Draw the decision graph corresponding to the latter situation. Explain Data's optimal choices of action based on this graph.

2. [20 points] **Markov Decision Problem**

   Jack has a car dealership and is looking for a way to maximizes his profits. Every week, Jack orders a stock of cars, at the cost of d dollars per car. These cars get delivered instantly. The new cars get added to his inventory. Then during the week, he sells some random number of cars, k, at a price of c each. Jack also incurs a cost of u for every unsold car that he has to keep in inventory.

   Formulate this problem as a Markov Decision Process. What are the states and actions? What are the rewards? What are the transition probabilities? Describe the long-term return.

3. [10 points] **Optimal policies** Consider the general domain of gridworld navigation tasks, where there is a goal state, obstacles, and a discount factor $\gamma < 1$. The actions are stochastic, so the agent may "slip" into a different cell when trying to move. There are 5 possible actions (go north, south, east, west or stay). Consider the situation in which negative costs are incurred when bumping into walls. Can you draw a 3×3 example environment in which the best action in at least one state is to stay? If so, specify the actions, rewards and transition probabilities. If not, explain why.

4. [10 points] **Action-values**

   Write a Belman equation for the action-value function for a fixed policy, Q^π. Write a TD learning rule for Q^π.

5. [20 points] **Reinforcement learning**

Sometimes it is convenient in reinforcement learning to define an advantage function, instead of an action-value function. Given a policy π, the advantage of action a in state s is defined as:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

(a) [2 points] Write the advantage function in terms of action values only.

(b) [5 points] Suppose that in state s you have two available actions, a_1 and a_2, and $\pi(s,a_1) = \pi(s,a_2) = 0.5$. What can you say about $A(s,a_1)$ and $A(s,a_2)$?

(c) [5 points] Suppose that π is a deterministic policy, which picks the best action in state s. Let a^* be this best action. What can you say about $A(s,a^*)$?

(d) [8 points] Devise a learning algorithm which learns an advantage function from trajectories.

6. [20 points] **Exploration**

We discussed in class the issue of exploration, and we mentioned two basic strategies, ϵ-greedy and Boltzmann exploration. Another commonly used strategy is optimistic initialization. In this case, the action-values are initialized higher than any value that can ever be attained in the environment (e.g. you imagine that whatever you do, you will die a billionaire). In this case, all the actions that the agent does seem bad, so it is encouraged to try something different.

(a) [10 points] Suppose that the agent starts with optimistic initial values then always picks the greedy action according to its estimates. Give an example of a small environment in which this strategy would perform better than ϵ-greedy, in terms of the amount of time it takes to figure out what the best action is.

(b) [10 points] Now consider the case in which the rewards change over time (e.g. the goal state moves). Which of the two strategies would you expect to perform better and why?