COMS BC 3997 - F22: Problem Set 2

Introduction: Welcome to the second problem set of the semester! As you are hopefully already aware, this PDF comprises the written component of the first problem set. In addition to solving the problems found below, you will also need to complete the coding part of the assignment, found in the Github repo. Finally, we'd like to remind you that all work should be yours and yours alone. This being said, in addition to being able to ask questions at office hours, you are allowed to discuss questions with fellow classmates, provided 1) you note the people with whom you collaborated, and 2) you **DO NOT** copy any answers. Please write up the solutions to all problems independently.

Collaborators:

Markov Decision Processes (13 Points)

Annie is a 5-year old girl who loves eating candy and is ambivalent regarding vegetables. She can either choose to eat candy (Hershey's, Skittles, Peanut Butter Cups) or eat vegetables during every meal. Eating candy gives her +10 in happiness points, while eating vegetables only gives her +4 happiness points. But if she eats too much candy while sick, her teeth will all fall out (she won't be able to eat any more). Annie will be in one of three states: healthy, sick, and toothless. Eating candy tends to make Annie sick, while eating vegetables tends to keep Annie healthy. If she eats too much candy, she'll be toothless and won't eat anything else. The transitions are shown in the table below.

Health condition	Candy or Vegetables?	Next condition	Probability
healthy	vegetables	healthy	1
healthy	candy	healthy	1/4
healthy	candy	sick	3/4
sick	vegetables	healthy	1/4
sick	vegetables	sick	3/4
sick	candy	sick	7/8
sick	candy	toothless	1/8

Problem 1: (2 points) Model this problem as a Markov Decision Process: specify each state, action, and transition $T(s, a, s')$ and reward $R(a)$ functions.

Solution 1:

- The MDP states are:
- The MDP actions are:
- The MDP transitions are:
- The MDP rewards are:

Problem 2: (4 points) Write down the Value function $V(s)$ for this problem in all possible states under the following policies (the discount factor can be expressed as γ.

Remember $V^{\pi_k} = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_k}(s')\right]$

Hint: you may want to first write things down leaving $V^{\pi_k}(\cdot)$ as a variable and then solve a system of equations resulting in fractions related to γ – don't worry if it doesn't simplify well (and don't feel like you need to simplify it fully!

(a) π_1 in which Annie always eats candy

(b) π_2 in which Annie always eats vegetables

Solution 2:

(a) The Value function under π_1 is:

$$V^{\pi_1}(\cdot) = ?$$

(b) The Value function under π_2 is:

$$V^{\pi_1}(\cdot) = ?$$

Problem 3: (5 points) Start with a policy in which Annie always eats candy no matter what the her health condition is (π_1 from Problem 2). Simulate the first two iterations of the policy iteration algorithm. Show how the policy evolves as you run the algorithm. Set $\gamma = 0.9$.

Hint: You can start by plugging in γ to your solution from Problem 2, then extract the optimal policy, recompute the values, and re-extract the policy (see the solution template below)!

Remember $\pi_{k+1} = arg\max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^{\pi_k}(s')\right]$

Solution 3:
The Value function under π_1 is:

$$V^{\pi_1}(\cdot) =?$$

Using Policy extraction our new policy $\pi_{1.1}$ is:

$$\pi_{1.1}(\cdot) =?$$

The Value function under $\pi_{1.1}$ is:

$$V^{\pi_{1.1}}(\cdot) =?$$

Using Policy extraction our new policy $\pi_{1.2}$ is:

$$\pi_{1.2}(\cdot) =?$$

Problem 4: (2 points) Are the following statements true or false for an MDP? Briefly explain why (1-3 sentences).

(a) If one is using value iteration and the values have converged, the policy must have converged as well.

(b) For an infinite horizon MDP with a finite number of states and actions and with a discount factor that satisfies $0 < \gamma <= 1$, policy iteration is guaranteed to converge.

(c) There may be more than one optimal value function.

(d) There may be more than one optimal policy.

Solution 4:

(a)

(b)

(c)

(d)

Reinforcement Learning (7 points)

Problem 5: (1 point)
In class we learned about temporal-difference techniques for reinforcement learning. Now we'd like to take the next step (literally). Suppose we take two steps and get the state/action sequence s-a-s'-a'-s". Write the temporal-difference update equations for those transitions in terms of $V(\cdot)$.

Solution 5:

Problem 6: (3 points)
Consider the following deterministic Transition/Reward Model for an MDP with states (S1, S2, S3, S4, S5, S6) and actions (A1, A2, A3):

From	Action	To	Reward
S1	A3	S2	3
S2	A1	S1	2
S2	A2	S3	1
S3	A1	S4	2
S3	A2	S5	10
S4	A3	S3	5
S5	A1	S6	7
S6	A3	S5	2

(a) Suppose we start in state S3. We run Q-learning on this MDP using a greedy policy (always choose action with best Q-value). Ties are given to the action with the lower number ($A1 > A2$, etc). Assume $\alpha = 0.5$ and $\gamma = 0.9$. We initialize all Q-values to 0 and update the Q-values after each transition. What are the first 4 (state, action) pairs visited, including the start state and the following action? Remember that:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)[R(s,a,s') + \gamma \max_{a'} Q(s',a')]$$

(b) Why is this simple-greedy policy limited and what can we alter about this algorithm to overcome this limitation?

Solution 6:

(a)

(b)

Problem 7: (3 points)
Now consider the following Transition/Reward Model for an MDP with states (S1, S2, S3, S4, S5) and actions (A1, A2). Note that (S4, S5) are terminal states with no valid actions. Assume $\gamma = 1.0$: Initialize $Q(\cdot, \cdot) = 0$ for all (state, action) pairs for q-learning. With $\alpha = 0.5$, compute $Q(\cdot, \cdot)$ after seeing the following series of transitions:

From	Action	To	Reward	Probability
S1	A1	S2	0	0.5
S1	A1	S3	0	0.5
S1	A2	S2	0	0.25
S1	A2	S3	0	0.75
S2	A1	S4	6	1.0
S2	A2	S4	12	0.5
S2	A2	S5	4	0.5
S3	A1	S5	16	1.0

- $[(S1, A1, S2) \rightarrow (S2, A2, S4)]$ followed by
- $[(S1, A2, S3) \rightarrow (S3, A1, S5)]$ followed by
- $[(S1, A2, S2) \rightarrow (S2, A1, S4)]$.

Again, remember that:

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + (\alpha)[R(s, a, s') + \gamma \max_{a'} Q(s', a')]$$

Solution 7: