Artificial Intelligence (521495A), Spring 2022
Exercise 6 : Reinforcement learning
Solutions

This handout contains the example solutions for the problems 1-2.

Problem 1. Similar to Problem 4 in Exercise 5, consider a game where you repeatedly draw a card (with replacement) that is 2, 3, or 4. This time you don't know how many of 2, 3, or 4 cards are in the deck (i.e., the distribution of cards and due to that the transition function of MDP is unknown). You can either Draw or Stop if the total score of the cards you have drawn is less than 7. Otherwise, you must Stop. When you Stop, your reward is equal to your total cumulative score (up to 6), or zero if you exceed 6 (i.e., 7 or higher). When you Draw, you receive no reward. There is no discount ($\gamma = 1$). Apply model-based reinforcement learning to estimate the transition and reward functions based on training episodes below. Each line in an Episode is a tuple containing (s, a, s', r).

Episode 1	Episode 2	Episode 3
0, Draw, 3, 0	0, Draw, 4, 0	0, Draw, 2, 0
3, Draw, 6, 0	4, Draw, 7, 0	2, Draw, 4, 0
6, Stop, Done, +6	7, Stop, Done, 0	4, Stop, Done, +4

Episode 4	Episode 5	Episode 6
0, Draw, 2, 0	0, Draw, 2, 0	0, Draw, 3, 0
2, Draw, 4, 0	2, Draw, 5, 0	3, Draw, 5, 0
4, Draw, 6, 0	5, Draw, 8, 0	5, Stop, Done, +5
6, Stop, Done, +6	8, Stop, Done, 0	

Estimate the transition $T(s, a, s')$ and reward $R(s, a, s')$ function values for

(a) $T(0, Draw, 2)$

(b) $T(0, Draw, 3)$

(c) $T(4, Stop, Done)$

(d) $R(4, Draw, 6)$

(e) $R(6, Stop, Done)$

(f) $R(4, Stop, Done)$

Calculate average transition and reward occurences in training episodes:

(a) $T(0, Draw, 2) = 3/6 = 1/2$

(b) $T(0, Draw, 3) = 2/6 = 1/3$

(c) $T(4, Stop, Done) = 1$

(d) $R(4, Draw, 6) = 0$

(e) $R(6, Stop, Done) = (6 + 6)/2 = 6$

(f) $R(4, Stop, Done) = 4/1 = 4$

Problem 2. Consider the same setting as in Problem 1, but now apply model-free reinforcement learning. The learning rate is $\alpha = 0.5$ and there is no discount ($\gamma = 1$). Use fixed reward function (similar to Problem 4 in Exercise 5) as follows

$$\mathrm{R}\left(\mathrm{s},\ \mathrm{Stop}\ ,\mathrm{s}'\right) = \mathrm{s}\ \text{if}\ \mathrm{s} \leq 6$$
$$\mathrm{R}\left(\mathrm{s},\mathrm{a},\mathrm{s}'\right) = 0\ \text{otherwise.}$$

(a) Use direct evaluation and training episodes from Problem 1. to update the state policy values.

(b) Formulate Temporal Difference learning update equation and calculate state policy values using the observations from training episodes 1-6. V values are initialized to zeros.

(c) Use direct evaluation to calculate following Q-values (i.e., averaging the discounted reward).

1. $Q(0, Draw)$

2. $Q(2, Draw)$

(a) Calculate output values as cumulative rewards from each state averaged over occured episodes:

S	**0**	**2**	**3**	**4**	**5**	**6**	**Done**
V	21/6	10/3	11/2	10/3	5/2	12/2	0

(b) Update equation

$$V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + \alpha[R(s,a,s') + \gamma V^{\pi}(s')]$$

Values are updated using the iterative update equation and training episodes one by one:

Episode / S	**0**	**2**	**3**	**4**	**5**	**6**	**Done**
Init.	0	0	0	0	0	0	0
1	0	0	0	0	0	3	0
2	0	0	0	0	0	3	0
3	0	0	0	2	0	3	0
4	0	1	0	2 1/2	0	4 1/2	0
5	1/2	1/2	0	2 1/2	0	4 1/2	0
6	1/4	1/2	0	2 1/2	2 1/2	4 1/2	0

(c) Use direct evaluation to calculate following Q-values (i.e., averaging the discounted reward).

Similar to (a) but using Q-value representation:

1. $Q(0, Draw) = (6 + 0 + 4 + 6 + 0 + 5)/6 = 21/6$

2. $Q(2, Draw) = (4 + 6 + 0)/3 = 10/3$