CS461 Artificial Intelligence

Fall 2024: Homework 3

Due Date: 10.12.2024

1   HMMs

Consider a Markov Model with a binary state X (i.e., X_t is either 0 or 1). The transition probabilities are given as follows:

X_t	X_{t+1}	$P(X_{t+1}|X_t)$
0	0	0.9
0	1	0.1
1	0	0.5
1	1	0.5

(a) The prior belief distribution over the initial state X_0 is uniform, i.e., $P(X_0 = 0) = P(X_0 = 1) = 0.5$. After one timestep, what is the new belief distribution, $P(X_1)$?

X_1	$P(X_1)$
0	
1	

Now, we incorporate sensor readings. The sensor model is parameterized by a number $\beta \in [0, 1]$:

X_t	E_t	$P(E_t|X_t)$
0	0	β
0	1	$(1 - \beta)$
1	0	$(1 - \beta)$
1	1	β

(b) At $t = 1$, we get the first sensor reading, $E_1 = 0$. Use your answer from part (a) to compute $P(X_1 = 0|E_1 = 0)$. Leave your answer in terms of β.

(c) For what range of values of β will a sensor reading $E_1 = 0$ increase our belief that $X_1 = 0$? That is, what is the range of β for which $P(X_1 = 0|E_1 = 0) > P(X_1 = 0)$?

(d) Unfortunately, the sensor breaks after just one reading, and we receive no further sensor information. Compute $P(X_\infty|E_1 = 0)$, the stationary distribution very many timesteps from now.

X_∞	$P(X_\infty|E_1 = 0)$
0	
1	

(e) How would your answer to part (d) change if we never received the sensor reading E_1, i.e. what is $P(X_\infty)$ given no sensor information?

X_∞	$P(X_\infty)$
0	
1	

2   MDPs

Help Nathan Drake find the lost treasure of El Dorado that was hidden somewhere at the old frozen lake! For all parts of the problem, assume that value iteration begins with all states initialized to zero, and $\gamma = 1$.

Suppose that we are performing value iteration on the grid world MDP below. Nathan starts at the bottom-left part of the lake labeled 1. Nathan found an ancient manuscript detailing the rules of this MDP.

3	4	5 Treasure
1 Start	2 Gold Coin +2 reward	(black square)

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
1	Right	2	1.0	+2
1	Up	3	1.0	-5
2	Up	4	1.0	-5
3	Right	4	1-x	-5
3	Right	5	x	-2
4	Right	5	1.0	-5
5	Exit	End	1.0	+25

Please note that rewards are defined for transitions, i.e., $R(s,a,s')$.

1. Nathan can only go up or right, unless the action would cause Nathan to move off the board or into the black square.

2. Because the lake is frozen, if Nathan moves right from Square 3, it's possible (with probability x) that Nathan slips and actually moves two squares right.

3. From Squares 1, 2, and 4, Nathan deterministically moves one square in the chosen direction.

4. Once Nathan reaches Square 5, the only action available is Exit, which gives reward 25. (There are no actions available after exiting.)

5. Square 2 contains a coin, so any action that moves Nathan into Square 2 gives reward +2.

6. All other transitions give reward -2 if Nathan moves two squares, and -5 if Nathan moves one square.

(a) Find the optimal state values for Square 2 and Square 3. Your answer can be in terms of x.

(b) For which values of x would Nathan prefer to go up on Square 1? If there is no possibility, then write 'no possible value' (in any case, show your computation).

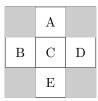(c) The rest of this question is independent of the previous subparts.

   Seeing how Nathan figures out the path to the treasure box, the ancient powers of the box start to confuse Nathan. Now, when Nathan takes an action, the action is changed to a different action according to a particular probability distribution denoted $p(a'|s, a)$. That is, given state s and action a, the action is changed to a' with probability $p(a'|s, a)$. To confuse him further, Nathan will receive reward as if his action did not change at all.

   Nathan tries to adopt those changes to Q-value iteration. Find the new formulation for the Q-values.

   Hint: Normally we have $Q(s, a) \leftarrow \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma \max_{a'} Q(s', a')]$.

3   Temporal Difference and Q-Learning

Consider the Gridworld below. We would like to use TD learning to find the values of these states.

	A	
B	C	D
	E	

Suppose we observe the following $(s, a, s', R(s, a, s'))^*$ transitions and rewards:

$(B, \text{East}, C, 2)$, $(C, \text{South}, E, 4)$, $(C, \text{East}, A, 6)$, $(B, \text{East}, C, 2)$

*Note that the $R(s, a, s')$ in this notation refers to observed reward, not a reward value computed from a reward function.

The initial value of each state is 0. Let $\gamma = 1$ and $\alpha = 0.5$.

(a) What are the learned values for each state from TD learning after all four observations?

(b) In class, we presented the following two formulations for TD-learning:

$$V^\pi(s) \leftarrow (1 - \alpha)V^\pi(s) + (\alpha)\text{sample} \tag{1}$$

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha(\text{sample} - V^\pi(s)) \tag{2}$$

Mathematically, these two equations are equivalent. However, they represent two conceptually different ways of understanding TD value updates. How could we intuitively explain each of these equations?

    (c) What are the learned Q-values from Q-learning after all four observations?
        Use the same $\alpha = 0.5$, $\gamma = 1$ as before.