COSC1125/1127 Artificial Intelligence

School of Computer Science and IT
RMIT University
Semester 1, 2019

Tutorial Sheet 9
Reinforcement Learning

1. In the passive *Adaptive Dynamic Programming* (ADP), an agent estimates the transition model T using tables of frequencies (model-based learning). Consider the 2 by 3 gridworld with the policy depicted as arrows and the terminal states are illustrated with X's.

	1	2
1	X	←
2	↑	←
3	X	↑

Compute the transition model after the passive ADP agent sees the following state transition observations:

Trial 1: $(3,2) \rightarrow (2,2) \rightarrow (2,1) \rightarrow (1,1)$
Trial 2: $(3,2) \rightarrow (2,2) \rightarrow (1,2) \rightarrow (1,1)$
Trial 3: $(3,2) \rightarrow (2,2) \rightarrow (2,1) \rightarrow (3,1)$
Trial 4: $(3,2) \rightarrow (2,2) \rightarrow (2,1) \rightarrow (2,2) \rightarrow (2,1) \rightarrow (1,1)$

With rewards perceived as follows:

State	**Reward**
$(3,2)$	-1
$(2,2)$	-2
$(2,1)$	-2
$(1,2)$	-1
$(3,1)$	-10
$(1,1)$	$+16$

Answer

We go through each trial, and count the frequency of each pair (state-action) and triple (state, action, next state) across all the trials.

Table of state-action frequencies/counts, $N(s, a)$:

state (s)	action (a)	Frequency
$(3, 2)$	N	4
$(2, 2)$	W	5
$(2, 1)$	N	4
$(1, 2)$	W	1

Table of state-action-state frequencies/counts, $N(s, a, s')$:

state (s)	action (a)	next state (s')	Frequency
$(3, 2)$	N	(2,2)	4
$(2, 2)$	W	(2,1)	4
$(2, 2)$	W	(1,2)	1
$(2, 1)$	N	(1,1)	2
$(2, 1)$	N	(3,1)	1
$(2, 1)$	N	(2,2)	1
$(1, 2)$	W	(1,1)	1

From these frequency tables, we can obtain an estimate of T, the transition model.

state (s)	action (a)	next state (s')	T(s,a,s')
$(3, 2)$	N	$(2, 2)$	$4/4 = 1.0$
$(2, 2)$	W	$(2, 1)$	$4/5 = 0.8$
$(2, 2)$	W	$(1, 2)$	$1/5 = 0.2$
$(2, 1)$	N	$(1, 1)$	$2/4 = 0.5$
$(2, 1)$	N	$(3, 1)$	$1/4 = 0.25$
$(2, 1)$	N	$(2, 2)$	$1/4 = 0.25$
$(1, 2)$	W	$(1, 1)$	$1/1 = 1.0$

2. Using the environment model and observations from question 1, now consider how a passive Temporal Difference agent will estimate the utility of the states. Using the first two trials, update the utilities as each observation comes in, using the temporal difference learning rule given in lectures. Use $\alpha = 0.5$ (recall α is the learning rate) and $\gamma = 1$, and for this question, assume α is a constant, i.e., α doesn't change as we visit a state more and more.

Answer

Trial 1:

- Starting state is $(3,2)$, $R[(3,2)] = -1$, $U[(3,2)] = -1$.
- First observation is $(3,2) \rightarrow (2,2)$, reward of $(2,2)$ is -2, hence $U[(2,2)] = -2$.

$$\begin{aligned} U[(3,2)] &= U[(3,2)] + \alpha(R[(3,2)] + \gamma U[(2,2)] - U[(3,2)]) \\ &= -1 + 0.5(-1 - 2 + 1) = -2 \end{aligned}$$

- Second observation is $(2,2) \rightarrow (2,1)$, reward of $(2,1)$ is -2, hence $U[(2,1)] = -2$:

$$\begin{aligned} U[(2,2)] &= U[(2,2)] + \alpha(R[(2,2)] + \gamma U[(2,1)] - U[(2,2)]) \\ &= -2 + 0.5(-2 - 2 + 2) = -3 \end{aligned}$$

- Third observation is $(2,1) \rightarrow (1,1)$, reward of $(1,1)$ is $+16$, hence $U[(1,1)] = +16$:

$$\begin{aligned} U[(2,1)] &= U[(2,1)] + \alpha(R[(2,1)] + \gamma U[(1,1)] - U[(2,1)]) \\ &= -2 + 0.5(-2 + 16 + 2) = 6 \end{aligned}$$

Taking a similar process to trial 2, we obtain the following:

- First observation is $(3,2) \rightarrow (2,2)$. As we visited $(2,2)$ already, no need to update utility of $(2,2)$; Update $U[(3,2)]$:

$$\begin{aligned} U[(3,2)] &= U[(3,2)] + \alpha(R[(3,2)] + \gamma U[(2,2)] - U[(3,2)]) \\ &= -2 + 0.5(-1 - 3 + 2) = -3 \end{aligned}$$

- Second observation is $(2,2) \rightarrow (1,2)$, reward of $(1,2)$ is -1, hence $U[(1,2)] = -1$; Update $U[(2,2)]$:

$$\begin{aligned} U[(2,2)] &= U[(2,2)] + \alpha(R[(2,2)] + \gamma U[(1,2)] - U[(2,2)]) \\ &= -3 + 0.5(-2 - 1 + 3) = -3 \end{aligned}$$

- Third observation is $(1,2) \rightarrow (1,1)$. As we visited $(2,2)$ already, no need to update utility of $(2,2)$; Update $U[(1,2)]$:

$$\begin{aligned} U[(1,2)] &= U[(1,2)] + \alpha(R[(1,2)] + \gamma U[(1,1)] - U[(1,2)]) \\ &= -1 + 0.5(-1 + 16 + 1) = 7 \end{aligned}$$

3. **(optional)** Consider the "occasionally-random" and exploration function methods to strike a balance between exploitation and exploration. Recall in the "occasionally-random" approach, $\frac{1}{t}$ of the time the agent selects a random action, otherwise follow the greedy policy. What would a high t value mean? What about a low value t?

   Contrast this with the exploration function concept. As an example, consider this exploration function,

$$f(u,n) = \begin{cases} R^+ \text{ , n } < N_e \\ u \text{ , otherwise} \end{cases}$$

What does high/low settings for R^+ and N_e result in?

Answer

For the occasionally-random scheme, a high t value means we do not often select a random action to take, resulting in less exploration. It means the agent is more likely to (greedily) select what it considers is the best policy, i.e., more exploitation, from its current estimation of the environment/world. Conversely a low t value means more focus on exploration over exploitation.

Using the exploration function given in the textbook and lectures, the first case/condition encourages exploration, while second case/condition encourages exploitation. Hence a high value of R^+ means the agent has an optimistic view about unknown or rarely states - the higher it is, the more optimistic. Unvisited states will be set this high value for its utility/Q-values. However, if this view is far from reality, the update rules will bring it back towards the true value, but will take longer. Similarly a small value will mean the agent has a pessimistic view about unknown or rarely visited states. N_e determines for how long we keep in exploration mode. The higher N_e is, the more willing the agent is to explore a state and its associated actions.

4. Consider the grid world from question 1, but now there are no policy specified.

   Apply Q-learning taught in class and textbook to learn Q-Values of all state-action pairs for two trials/episodes (one sequence from starting state to a terminal state). Initialise all Q-values to 0.

   The algorithm in the textbook (and which we follow in this course) does not update the Q-Values of terminal states. Instead, when the next state s' is a terminal state, rather then just setting all previous states, actions and rewards to null, we will additional set all Q-values for that terminal state to its reward. E.g., for state $(1, 1)$, its reward is $+16$, and the first time we visit $(1, 1)$ we will set $Q(North, (1, 1)) = Q(South, (1, 1)) = Q(East, (1, 1)) = Q(West, (1, 1)) = +16$.

   To simplify calculations, for this question assume actions are deterministic, i.e., if agent goes west, it will go west.

   Similar to the TD learning question, use $\alpha = 0.5$ and $\gamma = 1$. For the exploration function, use the one described in question 5, with $R^+ = +10$ and $N_e = 1$. In reality the starting state can vary, but for this question, assume we always start at $(3, 2)$. Note that there can be several answers possible, depending on the action selected when there are tie breakers.

Answer

First trial/episode:

- Start at state $(3,2)$, $R[(3,2)] = -1$.

  No previous state, so we don't need to update any Q-values. Instead, given all actions and next states have the same Q-value and we haven't visited any of them yet, we randomly select an action - say west (W) and arrive at $(3,1)$.

- We move to $(3,1)$, so current state is $(3,1)$ and previous state is $(3,2)$. Recall for terminal states, we will update the Q-values for all actions to its reward, but only after updating the Q-value of previous state. Update $Q[W,(3,2)]$:

  $$\begin{aligned} Q[W,(3,2)] &= Q[W,(3,2)] + \alpha(R[(3,2)] + \gamma \max_{a'} Q[a',(3,1)] - Q[W,(3,2)]) \\ &= 0 + 0.5(-1 + 0 - 0) = -0.5 \end{aligned}$$

  We update $Q[N,(3,1)] = Q[S,(3,1)] = Q[E,(3,1)] = Q[W,(3,1)] = -10$.

Second trial/episode:

- Start at state $(3,2)$, $R[(3,2)] = -1$.

  No previous state, so we don't need to update any Q-values.

  Instead, given all actions and next states have the same Q-value but we haven't gone north from $(3,2)$, we go north (N), and go to $(2,2)$.

- We move to $(2,2)$, so current state is $(2,2)$ and previous state is $(3,2)$. Non-terminal state. Update $Q[N,(3,2)]$:

  $$\begin{aligned} Q[N,(3,2)] &= Q[N,(3,2)] + \alpha(R[(3,2)] + \gamma \max_{a'} Q[a',(2,2)] - Q[N,(3,2)]) \\ &= 0 + 0.5(-1 + 0 - 0) = -0.5 \end{aligned}$$

  From $(2,2)$, all actions of $Q[*,(2,2)]$ are equally good according to exploration function, hence we random select one. Say we selected north (N), and go to $(1,2)$.

- We move to $(1,2)$, so current state is $(1,2)$ and previous state is $(2,2)$. Non-terminal state. Update $Q[N,(2,2)]$:

  $$\begin{aligned} Q[N,(2,2)] &= Q[N,(2,2)] + \alpha(R[(2,2)] + \gamma \max_{a'} Q[a',(1,2)] - Q[N,(2,2)]) \\ &= 0 + 0.5(-2 + 0 - 0) = -1 \end{aligned}$$

  From $(1,2)$, all actions of $Q[*,(1,2)]$ are equally good according to exploration function, hence we random select one. Say we selected west (W), and go to $(1,1)$.

- We move to (1,1), so current state is (1,1) and previous state is (1,2). (1,1) is a terminal state, so we will update the Q-values for all actions to its reward, but only after updating the Q-value of previous state. Update Q[W, (1,2)]:

  $$\begin{aligned} Q[W,(1,2)] &= Q[W,(1,2)] + \alpha(R[(1,2)] + \gamma \max_{a'} Q[a',(1,1)] - Q[W,(1,2)]) \\ &= 0 + 0.5(-1 + 0 - 0) = -0.5 \end{aligned}$$

  Afterwards, we also update $Q[N,(1,1)] = Q[S,(1,1)] = Q[E,(1,1)] = Q[W,(1,1)] = +16$.

5. For the following scenarios, briefly describe how we can model them as reinforcement learning problems (states, actions, type of rewards etc):

   a) Learning to play chess.

   b) Mary is about to graduate, and she decides to plan her finances for the next 40 years (until retirement). Consider what a reinforcement model might look like for financial planning.

Answer

a) States could be all the possible piece configurations. Actions are the moves of each individual piece. Non-terminal state rewards could be the value of a piece taken/lost, and terminal state (win or lose) can have a large positive or negative value.

b) This is an interesting problem, because there is time involved. Investing at 20 years old will likely have smaller returns (rewards) then investing at 40 years old, making the assumption a person will have more money to invest the older they get. To reflect this, the states can be Mary at different ages, actions are different types of investment, and rewards are tied with a state (age) and action (investment). There is a variant of Q-learning that associates rewards to a state-action pair, and it would be appropriate to use such a one.

6. Consider chess. If we wanted to approximate the utility of the states using a sum of weighted features, discuss the type of features that might be useful for playing chess.

Answer

Remember the utility is a indication of the “usefulness” of a state - in the case of chess, it is evaluating whether a state can lead to a winning or losing strategy. In this context, some features, not exhaustive, could be:

- Number of pieces agent has;
- Number of pieces opponent has;
- Total value of pieces (Queen = 9, Rook = 5 etc) agent has (similarly for opponent);
- Number of squares the agent’s King can move (if not many, King might not have much opportunity to break a check;
- Number of moves for pawn closest to been promoted;
- Many more!