cse574_hw4

Zack

November 7, 2018

Contents

CSE 574 HW 4 Zackary Crosley 1209001881

1 Passive Reinforcment Learning

Given the 4x3 grid world below, initialize all the utility values to 0. Set the utility of (4,3) to 100 and (4,2) to -100. $\gamma = 1$

1.1 Calculate Direct Utility and Adaptive Dynamic Programming Estimation.

Instead of setting reward function for each block with a fixed value define $R(n) = -n^2$. Let n denote the number of steps from the beginning, so the reward value for (1,1) is 1. Given two trials starting from (1,1):
a. (1,1) -> (1,2) -> (1,3) -> (2,3) -> (1,3) -> (2,3) -> (3,3) -> (4,3)
b. (1,1) -> (2,1) -> (3,1) -> (3,2) -> (4,2)
Apply direct utility and adaptive dynamic programming separately to update the value for each visited grid.

1.1.1 Direct Utility Estimation: Rewart-To-Go.

a.

$$
\begin{aligned}
U(4,3) &= -(7^2) + 100 = 51 \\
U(3,3) &= -(6^2) + 51 = 15 \\
U(2,3) &= -(5^2) + 15 = -10 \\
U(1,3) &= -(4^2) + -10 = -26 \\
U(2,3) &= -(3^2) + -26 = -35 \\
U(1,3) &= -(2^2) + -39 = -43 \\
U(1,2) &= -(1^2) + -43 = -44 \\
U(1,1) &= 0 + -44 = -44
\end{aligned}
$$

$$
\begin{aligned}
U(2,3) &= \frac{-10 + -35}{2} = -22.5 \\
U(1,3) &= \frac{-26 + -39}{2} = -32.5
\end{aligned}
$$

b.

$$
\begin{aligned}
U(4,2) &= -(4^2) + -100 = -116 \\
U(3,2) &= -(3^2) + -116 = -125 \\
U(3,1) &= -(2^2) + -125 = -129 \\
U(2,1) &= -(1^2) + -129 = -130 \\
U(1,1) &= -130
\end{aligned}
$$

$$
(1,1) = \frac{-44 + -130}{2} = -87
$$

-32.5	-22.5	15	100
-44	[HTML]000000	-125	-100
-87	-130	-129	0

Table 1: Table after direct utility estimation with two specified runs.

1.1.2 Adaptive Dynamic Programming:

U^{π} (s) = R(s) + γ $\sum_{s'}$ P(s' | s, π(s)) U^{π} (s')

$$
\begin{aligned}
&TransitionModel\\
&\quad (3,3) : P(4,3) = 1\\
&\quad (2,3) : P(3,3) = 0.5; P(1,3) = 0.5\\
&\quad (1,3) : P(2,3) = 1\\
&\quad (1,2) : P(1,3) = 1\\
&\quad (1,1) : P(1,2) = 0.5; P(2,1) = 0.5\\
&\quad (2,1) : P(3,1) = 1\\
&\quad (3,1) : P(3,2) = 1\\
&\quad (3,2) : P(4,2) = 1
\end{aligned}
$$

a.

$$
\begin{aligned}
U^\pi(4,3) &= R(4,3) = \mathbf{100} \\
U^\pi(3,3) &= R(3,3) + \gamma\left(P((4,3)|(3,3),\pi(3,3))U^\pi(4,3)\right) \\
&= 0 + 1 \times 1 \times 100 = \mathbf{100} \\
U^\pi(2,3) &= R(2,3) + \gamma \sum_{(3,3),(1,3)} P(s'|s,\pi(s))U^\pi(s') \\
&= 0 + 1 \times (P((3,3)|(2,3),\pi((2,3)))U^\pi((3,3)) + P((1,3)|(2,3),\pi((2,3)))U^\pi((1,3)) \\
&= 0.5 \times 100 + 0.5 \times 0 = \mathbf{50} \\
U^\pi(1,3) &= R(1,3) + \gamma P((2,3)|(1,3),\pi((1,3)))U^\pi((2,3)) \\
&= 0 + 1 \times 1 \times 50 = \mathbf{50} \\
U^\pi(2,3) &= R(2,3) + \gamma \sum_{(3,3),(1,3)} P(s'|s,\pi(s))U^\pi(s') \\
&= 50 + 1 \times (P((3,3)|(2,3),\pi((2,3)))U^\pi(3,3) + P((1,3)|(2,3),\pi((2,3)))U^\pi(1,3)) \\
&= 50 + (0.5 \times 100 + 0.5 \times 50) = 50 + 50 + 25 = \mathbf{125} \\
U^\pi(1,3) &= R(1,3) + \gamma P((2,3)|(1,3),\pi((1,3)))U^\pi(2,3) \\
&= 50 + 1 \times 1 \times 125 = \mathbf{175} \\
U^\pi(1,2) &= R(1,2) + \gamma P((1,3)|(1,2),\pi((1,2)))U^\pi(1,3) \\
&= 0 + 1 \times 1 \times 175 = \mathbf{175} \\
U^\pi(1,1) &= R(1,1) + \gamma \sum_{(1,2),(2,1)} P(s'|s,\pi(s))U^\pi(s') \\
&= 0 + 1 \times (P((1,2)|(1,1),\pi(1,1))U^\pi(1,2) + P((2,1)|(1,1),\pi(1,1))U^\pi(2,1)) \\
&= 0.5 \times 175 + 0.5 \times 0 = \ \mathbf{87.5}
\end{aligned}
$$

b.

$$
\begin{aligned}
U^\pi(4,2) &= R(4,2) = \mathbf{\text{-}100} \\
U^\pi(3,2) &= R(3,2) + \gamma P((4,2)|(3,2),\pi(3,2))U^\pi(4,2) \\
&= 0 + 1 \times 1 \times -100 = \mathbf{\text{-}100} \\
U^\pi(3,1) &= R(3,1) + \gamma P((3,2)|(3,1),\pi(3,1))U^\pi(3,2) \\
&= 0 + 1 \times 1 \times -100 = \mathbf{\text{-}100} \\
U^\pi(2,1) &= R(2,1) + \gamma P((3,1)|(2,1),\pi(2,1))U^\pi(3,1) \\
&= 0 + 1 \times 1 \times -100 = \mathbf{\text{-}100} \\
U^\pi(1,1) &= R(1,1) + \gamma \sum_{(1,2),(2,1)} P(s'|s,\pi(s))U^\pi(s') \\
&= R(1,1) + \gamma\left[P((1,2)|(1,1),\pi(1,1))U^\pi(1,2) + P((2,1)|(1,1),\pi(1,1))U^\pi(2,1)\right] \\
&= 87.5 + 1 \times 1 \times \left[P((1,2)|(1,1),\pi(1,1))U^\pi(1,2) + P((2,1)|(1,1),\pi(1,1))U^\pi(2,1)\right] \\
&= 87.5 + [0.5 \times -100 + 0.5 \times 175] = 87.5 - 50 + 87.5 = \mathbf{125}
\end{aligned}
$$

1.2 Calculate Minimum Value for (1,1)

Let R(n) be (n-λ)^2. Apply the direct utility method on both trails to compute the lambda which minimizes the utility of (1,1).

$$
\begin{aligned}
(4,3) &= (n-\lambda)^2 + 100 = (7-\lambda)^2 + 100 \\
(3,3) &= (6-\lambda)^2 + (7-\lambda)^2 + 100 \\
(2,3) &= (5-\lambda)^2 + (6-\lambda^2) + (7-\lambda)^2 + 100 \\
(1,3) &= 100 + \sum_{i=4,5,6,7} (i-\lambda)^2 \\
(2,3) &= 100 + \sum_{i=3,4,5,6,7} (i-\lambda)^2 \\
(1,3) &= 100 + \sum_{i=2,3,4,5,6,7} (i-\lambda)^2 \\
(1,2) &= 100 + \sum_{i=1,2,3,4,5,6,7} (i-\lambda)^2 \\
(1,1) &= 100 + \sum_{i=0,1,2,3,4,5,6,7} (i-\lambda)^2 \\
&= 100 + \lambda^2 + (1-\lambda)^2 + (2-\lambda)^2 + \dots \\
&= 100 + \lambda^2 + 1 - 2\lambda + \lambda^2 + 4 - 4\lambda + \lambda^2 + \dots \\
&= 100 + 8\lambda^2 + (-2-4-6-8-10-12-14)\lambda + (1+4+9+25+36+49) \\
&= 8\lambda^2 - 56\lambda + 224 \\
\frac{d(1,1)}{d\lambda} &= 16\lambda - 56 = 0 \\
\lambda &= \frac{56}{16} = \frac{7}{2} \\
\therefore min(1,1)_\lambda &= \frac{7}{2}
\end{aligned}
$$

2 Q Learning and Adaptive Dynamic Programming

Given the MDP with transitions as shown, where agents can either stay (100% probability of staying in current state) or move (80% chance of moving to next state), and reward of final state is 1 with all other states 0. Calculate the using learning rate equal to 1:

2.1 Given trial (Move, A) -> (Move, B) -> (Move, C) show the Q values for each state action pair.

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

$$
\begin{aligned}
Q(C, Move) &\leftarrow Q(C, Move) + \alpha R(C) = 0 + 1 \times 1 = 1 \\
Q(B, Move) &\leftarrow Q(B, Move) + \alpha\left(R(B) + \gamma max_a\left[Q(C, Move), Q(C, Stay)\right] - Q(B, Move)\right) \\
&\leftarrow 0 + 1 \times (0 + 0.9 max_a\left[1, 0\right] - Q(B, Move)) = 0.9 \times [1 - 0] = \mathbf{0.9} \\
Q(A, Move) &\leftarrow Q(A, Move) + \alpha\left(R(A) + \gamma max_a\left[Q(B, Move), Q(B, Stay)\right] - Q(A, Move)\right) \\
&\leftarrow 0 + 1 \times (0 + 0.9 max_a left[0.9, 0] - Q(A, Move) = \mathbf{0.81}
\end{aligned}
$$

2.2 Is Q-Learning or Adaptive Dynamic Programming (ADP) better for long chains?

TODO

2.3 What would be disadvantages for ADP compared to Q Learning?

Adaptive Dynamic Programming requires a policy for evaluation, whereas Q Learning learns an optimal policy. Problems where the user has no inside knowledge of an optimal policy the ADP algorithm will not help to formulate an evaluation of each state. Furthermore Adaptive Dynamic Programming requires an estimation of the transition model, which is usually gathered from looking at a number of runs. Q-Learning however doesn't require a transition model, and learns the value of each state action pair. ADP will perform worse if the transition model estimated from the data is inaccurate or incomplete.

3 Active Reinforcment Learning

Write out parameter update equations for TD Learning where:

$$U(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}$$

$$E_j(s) = \frac{\left(\hat{U}_\theta(s) - u_j(s)\right)^2}{2}$$

$$\theta_i \leftarrow \theta_i - \alpha \frac{\partial E_j(s)}{\partial \theta_i}$$

$$\theta_i \leftarrow \theta_i - \alpha \left(u_j(s) - U_\theta(s)\right) \frac{\partial U_\theta(s)}{\partial \theta_i}$$

$$\theta_i \leftarrow \theta_i - \alpha \left(u_j(s) - \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}\right) \frac{\theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y}}{\partial \theta_i}$$

$$\theta_0 \leftarrow \theta_0 - \alpha \left(u_j(s) - \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}\right)$$

$$\theta_1 \leftarrow \theta_1 - \alpha \left(u_j(s) - \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}\right) \times x$$

$$\theta_2 \leftarrow \theta_1 - \alpha \left(u_j(s) - \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}\right) \times y$$

$$\theta_3 \leftarrow \theta_3 - \alpha \left(u_j(s) - \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}\right) \times \sqrt{(x-x_g)^2 + (y-y_g)^2}$$

4 Partially Observable Markov Decision Processes

NOTE: Since not specified it is assumed that transition probability is 1. That is, each action is guaranteed to succeed.

4.1 Belief State After 1 Move.

$$b_i(s_i) = \frac{P(o|s_i, a) \sum_{s_j \in S} P(s_i|s_j, a) b_{i-1}(s_j)}{P(o|a, b)}$$

$$P(O = 2|s \in (1,1), (1,2), (1,3), (2,1), (2,3), (4,1)) = 0.9$$

$$P(O = 1|s \in (3,1), (3,2), (3,3)) = 0.9$$

$$b_0 = \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, 0, 0$$

$$\begin{aligned} P(o = 1|a, b = b_0) &= \sum_{i=1}^{N} b_0(i) \times p(o = 1|s') \\ &= \frac{1}{9} \times (0.1 \times 7 + 0.9 \times 2) \\ &= \frac{1}{9} \times 2.5 \\ &= 0.27777 \end{aligned}$$

$$
\begin{aligned}
b_1(1,1) &= \frac{0.1(P((1,1)|(1,2), left) \times \frac{1}{9}))}{0.2777} \\
&= 0.36 \\
b_1(1,2) &= 0 \\
b_1(1,3) &= \frac{0.1(P((1,3)|(2,3), left) \times \frac{1}{9})}{0.27777} \\
&= 0.36 \\
b_1(2,1) &= \frac{0.1(P((2,1)|(3,1), left) \times \frac{1}{9})}{0.27777} \\
&= 0.36 \\
b_1(2,3) &= \frac{0.1(P((2,3)|(3,3), left) \times \frac{1}{9})}{0.27777} \\
&= 0.36 \\
b_1(3,1) &= \frac{0.9(P((3,1)|(4,1), left) \times \frac{1}{9})}{0.27777} \\
&= 3.2409 \\
b_1(3,2) &= 0 \\
b_1(3,3) &= 0 \\
b_1(4,1) &= 0 \\
b_1(4,2) &= 0 \\
b_1(4,3) &= 0
\end{aligned}
$$

$$
\begin{aligned}
sum &= 0.36 \times 4 + 3.2409 = 4.681 \\
\frac{0.36}{4.681} &= 0.0769 \\
\frac{3.2409}{4.681} &= 0.6923 \\
b_1 &= \textbf{[0.077, 0, 0.077, 0.077, 0.077, 0.692, 0, 0, 0, 0, 0, 0]}
\end{aligned}
$$

4.2 Time Complexity of D Steps of POMDP Value Iteration.

In value iteration, each of S states must be updated individually. This iterates over each S again, however these these iterations can be eliminated using vector math. Thus doing a depth D value iteration should be O(DS).

5 Hidden Markov Models

$$
\begin{aligned}
States S &= A, B, C \\
Observations O &= G, H \\
Initial Probabilities \pi &= [0.2, 0.1, 0.7] \\
Transition Probabilities T &= \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.2 & 0.2 & 0.6 \\ 0.3 & 0.2 & 0.5 \end{bmatrix} \\
Observation Probabilities &= \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix} \\
State Val &= P(s)P(s'|s)P(Obs|s')
\end{aligned}
$$

Layer One

$$
\begin{aligned}
P(A_1) &= P(A|Start)P(H|A) = 0.2 \times 0.1 = \mathbf{0.02} \\
P(B_1) &= P(B|Start)P(H|B) = 0.1 \times 0.4 = \mathbf{0.04} \\
P(C_1) &= P(C|Start)P(H|C) = 0.7 \times 0.9 = \mathbf{0.63}
\end{aligned}
$$

Layer Two

$$
\begin{aligned}
P(A_1 \to A_2) &= P(A_1)P(A_2|A_1)P(H|A) = 0.02 \times 0.1 \times 0.1 = \mathbf{0.0002} \\
P(B_1 \to A_2) &= P(B_1)P(A_2|B_1)P(H|A) = 0.04 \times 0.2 \times 0.1 = \mathbf{0.0008} \\
P(C_1 \to A_2) &= P(C_1)P(A_2|C_1)P(H|A) = 0.63 \times 0.3 \times 0.1 = \mathbf{0.0189} \\
P(A_2) &= \mathbf{0.0189} \\
P(A_1 \to B_2) &= P(A_1)P(B_2|A_1)P(H|B) = 0.02 \times 0.4 \times 0.4 = \mathbf{0.0032} \\
P(B_1 \to B_2) &= P(B_1)P(B_2|B_1)P(H|B) = 0.04 \times 0.2 \times 0.4 = \mathbf{0.0016} \\
P(C_1 \to B_2) &= P(C_1)P(B_2|C_1)P(H|B) = 0.63 \times 0.2 \times 0.4 = \mathbf{0.0504} \\
P(B_2) &= \mathbf{0.0504} \\
P(A_1 \to C_2) &= P(A_1)P(C_2|A_1)P(H|C) = 0.02 \times 0.5 \times 0.9 = \mathbf{0.009} \\
P(B_1 \to C_2) &= P(B_1)P(C_2|B_1)P(H|C) = 0.04 \times 0.6 \times 0.9 = \mathbf{0.0216} \\
P(C_1 \to C_2) &= P(C_1)P(C_2|C_1)P(H|C) = 0.63 \times 0.5 \times 0.9 = \mathbf{0.2835} \\
P(C_2) &= \mathbf{0.2835}
\end{aligned}
$$

Layer Three

$$P(A_2 \to A_3) = P(A_2)P(A_3|A_2)P(G|A) = 0.0189 \times 0.1 \times 0.9 = \mathbf{0.001701}$$
$$P(B_2 \to A_3) = P(B_2)P(A_3|B_2)P(G|A) = 0.0504 \times 0.2 \times 0.9 = \mathbf{0.009072}$$
$$P(C_2 \to A_3) = P(C_2)P(A_3|C_2)P(G|A) = 0.2835 \times 0.3 \times 0.9 = \mathbf{0.076545}$$
$$P(A_3) = \mathbf{0.076545}$$
$$P(A_2 \to B_3) = P(A_2)P(B_3|A_2)P(G|B) = 0.0189 \times 0.4 \times 0.6 = \mathbf{0.004536}$$
$$P(B_2 \to B_3) = P(B_2)P(B_3|B_2)P(G|B) = 0.0504 \times 0.2 \times 0.6 = \mathbf{0.006048}$$
$$P(C_2 \to B_3) = P(C_2)P(B_3|C_2)P(G|B) = 0.2835 \times 0.2 \times 0.6 = \mathbf{0.03402}$$
$$P(B_3) = \mathbf{0.03402}$$
$$P(A_2 \to C_3) = P(A_2)P(C_3|A_2)P(G|C) = 0.0189 \times 0.5 \times 0.1 = \mathbf{0.000945}$$
$$P(B_2 \to C_3) = P(B_2)P(C_3|B_2)P(G|C) = 0.0504 \times 0.6 \times 0.1 = \mathbf{0.003024}$$
$$P(C_2 \to C_3) = P(C_2)P(C_3|C_2)P(G|C) = 0.2835 \times 0.5 \times 0.1 = \mathbf{0.014175}$$
$$P(C_3) = \mathbf{0.014175}$$

Layer Four

$$P(A_3 \to A_4) = P(A_3)P(A_4|A_3)P(H|A) = 0.076545 \times 0.1 \times 0.1 = \mathbf{0.000765}$$
$$P(B_3 \to A_4) = P(B_3)P(A_4|B_3)P(H|A) = 0.03402 \times 0.2 \times 0.1 = \mathbf{0.0006804}$$
$$P(C_3 \to A_4) = P(C_3)P(A_4|C_3)P(H|A) = 0.014175 \times 0.3 \times 0.1 = \mathbf{0.0004253}$$
$$P(A_4) = \mathbf{0.000765}$$
$$P(A_3 \to B_4) = P(A_3)P(B_4|A_3)P(H|B) = 0.076545 \times 0.4 \times 0.4 = \mathbf{0.012247}$$
$$P(B_3 \to B_4) = P(B_3)P(B_4|B_3)P(H|B) = 0.03402 \times 0.2 \times 0.4 = \mathbf{0.0027216}$$
$$P(C_3 \to B_4) = P(C_3)P(B_4|C_3)P(H|B) = 0.014175 \times 0.2 \times 0.4 = \mathbf{0.001134}$$
$$P(B_4) = \mathbf{0.012247}$$
$$P(A_3 \to C_4) = P(A_3)P(C_4|A_3)P(H|C) = 0.076545 \times 0.5 \times 0.9 = \mathbf{0.03444}$$
$$P(B_3C_4) = P(B_3)P(C_4|B_3)P(H|C) = 0.03402 \times 0.6 \times 0.9 = \mathbf{0.018371}$$
$$P(C_3 \to C_4) = P(C_3)P(C_4|C_3)P(H|C) = 0.014175 \times 0.5 \times 0.9 = \mathbf{0.0063788}$$
$$P(C_4) = \mathbf{0.03444}$$

5.1 Draw HMM.

See attached.

5.2 Find Maximum Joint Probability.

Maximum Joint Prob = 0.03444

5.3 Find Maximum Conditional Probability.

$$P(u|HHGH) = \frac{0.03444}{0.03444 + 0.012247 + 0.000765} = \frac{0.03444}{0.047452} = 0.7257$$

5.4 Give the Maximally Likely Hidden State Sequence.

Most Likely Sequence: C C A C