CS461 HW3 – Görkem Kadir Solun – 22003214

1 HMMs

Consider a Markov Model with a binary state X (i.e., Xt 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 X0 is uniform, i.e., P(X0 = 0) = P(X0 = 1) = 0.5. After one timestep, what is the new belief distribution, P(X1)?

$$P(X1) = \sum_k P(X1, X0 = k) = \sum_k P(X1|X0 = k) * P(X0 = k)$$

$$For\ P(X1 = 1) = \frac{1}{2} * \frac{1}{10} + \frac{1}{2} * \frac{1}{2} = \frac{6}{20}, \qquad For\ P(X1 = 0) = \frac{1}{2} * \frac{1}{9} + \frac{1}{2} * \frac{1}{2} = \frac{14}{20}$$

Now, we incorporate sensor readings. The sensor model is parameterized by a number β ∈ [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, E1 = 0. Use your answer from part (a) to compute P(X1 = 0|E1 = 0). Leave your answer in terms of β.

$$P(X1 = 0|E1 = 0) = \frac{P(X1 = 0, E1 = 0)}{P(E1 = 0)} = \frac{P(X1 = 0) * P(E1 = 0|X1 = 0)}{P(E1 = 0)}$$

$$= \frac{P(X1 = 0) * P(E1 = 0|X1 = 0)}{\sum_k P(E1 = 0|X1 = k)P(X1 = k)} = \frac{0.7 * \beta}{(1-\beta) * 0.3 + \beta * 0.7} = \frac{\beta * 0.7}{0.3 + \beta * 0.4}$$

(c) For what range of values of β will a sensor reading E1 = 0 increase our belief that X1 = 0? That is, what is the range of β for which P(X1 = 0|E1 = 0) > P(X1 = 0)?

$$Find\ P(X1 = 0|E1 = 0) > P(X1 = 0)$$

$$\frac{0.7 * \beta}{0.3 + \beta * 0.4} > 0.7 \rightarrow\ \beta > \frac{1}{2}$$

(d) Unfortunately, the sensor breaks after just one reading, and we receive no further sensor information. Compute P(X∞|E1 = 0), the stationary distribution very many timesteps from now.

$$P(X\infty = 0) = P(X\infty = 0) * 0.9 + P(X\infty = 1) * 0.5$$

$$P(X\infty = 1) = P(X\infty = 0) * 0.1 + P(X\infty = 1) * 0.5$$

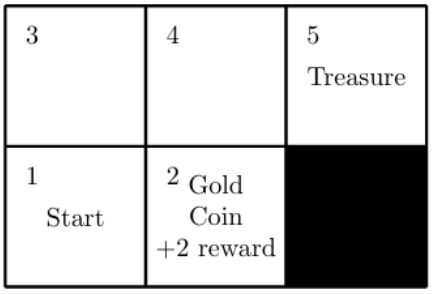
$$Let\ P(X\infty = 0) = p\ and\ P(X\infty = 1) = 1 - p$$

$$First\ equation: p = 0.9 * p + 0.5 * (1 - p) = \frac{5}{6}$$

$$P(X\infty = 0) = \frac{5}{6}, P(X\infty = 1) = \frac{1}{6}$$

Type equation here.

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 γ = 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.

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 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 a reward of 25. (There are no actions available after exiting.)

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

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

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

$$V(s) = \max_a \sum_{s'} T(s,a,s') * [R(s,a,s') + \gamma V(s')]$$

$$V(5) = 25$$

$$V(4) = \max_{right} \sum_{5} T(4,right,5) * [R(4,right,5) + V(5)] = 20$$

$$V(3) = \max_{right}\left[T(3, right, 4) * [-5 + V(4)] + T(3, right, 5) * [-2 + V(5)]\right] = 15 + 8x$$

$$V(2) = \max_{up}\left[T(2, up, 4) * [-5 + V(4)]\right] = 15$$

(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).

$$Q(s, a) = \sum_{s'} T(s, a, s') * [R(s, a, s') + \gamma V(s')]$$

$$Q(1, right) = T(1, right, 2) * [2 + V(2)] = 17$$

$$Q(1, up) = T(1, up, 3) * [-5 * V(3)] = 10 + 8x$$

$$10 + 8x = Q(1, up) > Q(1, right) = 17$$

$$x > \frac{7}{8}$$

(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 a 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.

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

Outer sum ($p(a'|s, a)$) accounts for the probability that action a is replaced with a'.

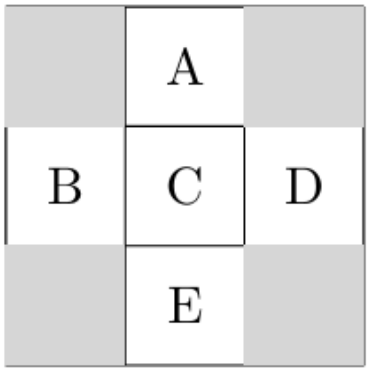
Inner sum ($\sum_{s'} T(s, a, s')$) is the usual summation over all possible next states s', considering that the actual action taken is a'.

Even though the action changes to a', the reward term ($R(s, a, s')$) is based on the original intended action a, as specified in the problem.

Future Q-values ($\gamma \max_{a''} Q(s', a'')$) represents the discounted future reward, if Nathan optimally chooses actions a'' in the next state s'.

$$Updated: Q(s, a) \leftarrow \sum_{a'} p(a'|s, a) * \sum_{s'} T(s, a', s') * \left[R(s, a, s') + \gamma \max_{a''} Q(s', a'')\right]$$

3 Temporal Difference and Q-Learning



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

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

$(B, East, C, 2), (C, South, E, 4), (C, East, A, 6), (B, East, C, 2)$

*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?

$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$ and $V(A) = V(B) = V(C) = V(D) = V(E) = 0$

$$V(B) = V(B) + 0.5\big(2 + V(C) - V(B)\big) = 1$$

$$V(C) = V(C) + 0.5\big(4 + V(E) - V(C)\big) = 2$$

$$V(C) = V(C) + 0.5\big(6 + V(A) - V(C)\big) = 4$$

$$V(B) = V(B) + 0.5\big(2 + V(C) - V(B)\big) = 3.5$$

$$V(A) = 0, V(B) = 3.5, V(C) = 4, V(D) = 0, V(E) = 0$$

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

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

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

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

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

This equation can be thought of as a weighted averaging process. Here, the new value estimate for state s is formed by taking a blend of what you previously believed $V^{\pi}(s)$ to be and the new sample you just observed. The factor $(1 - \alpha)$ keeps a portion of your old estimate, while α represents how much you trust the most recent sample. If α\alphaα is small, you're saying, I trust my old knowledge more, and if α is large, you're saying I trust this new observation more. In short, it's like continually updating your running average based on new pieces of evidence.

(2) $\boldsymbol{V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha\left(sample - V^{\pi}(s)\right)}$

This equation can be seen as an error-correction process. Here, the term $sample - V^{\pi}(s)$ represents how far off your current estimate is from the newly observed outcome. If the sample is larger than your current estimate, you increase your value; if it's smaller, you decrease it. The learning rate α tells you how aggressively to correct this error. Instead of thinking of it as a blend of old and new values, this perspective emphasizes adjusting the current estimate by a fraction of the observed prediction error. It's like saying, my old guess was off by some amount, and I will partially correct that error to get closer to what the truth seems to be.

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

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

$$Q(B,East) = Q(B,East) + 0.5 * \left[2 + \max_{a'} Q(C,a') - Q(B,East)\right] = 1$$

$$Q(C,South) = Q(C,South) + 0.5 * \left[4 + \max_{a'} Q(E,a') - Q(C,South)\right] = 2$$

$$Q(C,East) = Q(C,East) + 0.5 * \left[6 + \max_{a'} Q(A,a') - Q(C,East)\right] = 3$$

$$\begin{aligned} Q(B,East) &= Q(B,East) + 0.5 * \left[2 + \max_{a'} Q(C,a') - Q(B,East)\right] \\ &= Q(B,East) + 0.5 * \left[2 + \max_{a'}(Q(C,East), Q(C,South), 0) - Q(B,East)\right] = 3 \end{aligned}$$

$Q(B,East) = 3$, $Q(C,South) = 2$, $Q(C,East) = 3$, All other state-action values remain 0.