Homework 2

Jerry Duncan

September 16, 2020

1 Problem 1

Part a For part a, we're given a transition probability matrix for a 4 state problem. Each value (i, j) in the matrix represents the probability that the state is i given we were just in state j. More formally, that's $P(X_n = i|X_{n-1} = j) = M(i, j)$. Using this knowledge, we want to find $P(X_n = i)$ for all states. From the transition matrix, we can derive a set of equations to solve for each individually. For illustration purposes, I've written out the first equation in terms of P, then will use the matrix equations to write the rest.

$$P(X_n = 1) = \sum_{i=1}^{4} P(X_{n-1} = i)P(X_n = 1|X_{n-1} = i) \tag{1}$$

If we were to plug in the $P(X_n = 1|X_{n-1} = i)$ values we know from the transition matrix, we would have:

$$P(X_n = 1) = \frac{1}{4}P(X_{n-1} = 1) + 0P(X_{n-1} = 2) + \frac{1}{4}P(X_{n-1} = 3) + \frac{1}{4}P(X_{n-1} = 4)$$

Using the knowledge that this is a DTMC, we can assert that $\pi = \pi P$ and $\sum_i \pi(i) = 1$.

$$[\pi_1 \pi_2 \pi_3 \pi_4] = [\pi_1 \pi_2 \pi_3 \pi_4] \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}$$

We can multiply this out to get a system of independent linear equations we can solve.

$$\begin{aligned}
\pi_1 &= \frac{1}{4}\pi_1 + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 \\
\pi_2 &= \frac{1}{4}\pi_1 + \frac{1}{4}\pi_2 + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 \\
\pi_3 &= \frac{1}{2}\pi_1 + \frac{1}{2}\pi_2 + \frac{1}{4}\pi_3 \\
\pi_4 &= \frac{1}{4}\pi_2 + \frac{1}{4}\pi_3 + \frac{1}{2}\pi_4 \\
1 &= \pi_1 + \pi_2 + \pi_3 + \pi_4
\end{aligned}$$

Now we simply need to solve for each, one at a time. But first, let's simply a little.

$$\begin{aligned}
0 &= -\frac{3}{4}\pi_1 + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 && (2) \\
0 &= \frac{1}{4}\pi_1 + -\frac{3}{4}\pi_2 + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 && (3) \\
0 &= \frac{1}{2}\pi_1 + \frac{1}{2}\pi_2 + -\frac{3}{4}\pi_3 && (4) \\
0 &= \frac{1}{4}\pi_2 + \frac{1}{4}\pi_3 + -\frac{1}{2}\pi_4 && (5) \\
1 &= \pi_1 + \pi_2 + \pi_3 + \pi_4 && (6)
\end{aligned}$$

Now it should be easier to solve.

Using Eq. 3 and 6, we can multiply Eq. 6 by $\frac{3}{4}$ to get rid of π_2 in Eq. 3. Then we add them to get $\frac{3}{4} = \pi_1 + \pi_3 + \pi_4$. Using Eq. 6 again, we can deduce that $\pi_2 = \frac{1}{4}$.

Now we can plug π_2 in.

$$
\begin{aligned}
0 &= -\frac{3}{4}\pi_1 + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 \\
0 &= \frac{1}{4}\pi_1 + -\frac{3}{16} + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 \\
0 &= \frac{1}{2}\pi_1 + \frac{1}{8} + -\frac{3}{4}\pi_3 \\
0 &= \frac{1}{16} + \frac{1}{4}\pi_3 + -\frac{1}{2}\pi_4 \\
\frac{3}{4} &= \pi_1 + \pi_3 + \pi_4 \\
\pi_2 &= \frac{1}{4}
\end{aligned}
$$

We can now solve for π_1. Using Eq. 4, 5, and 6, we can multiply Eq. 6 by $\frac{1}{2}$ and add it to Eq. 5. This now gives us $\frac{3}{8} = \frac{1}{2}\pi_1 + \frac{3}{4}\pi_3 + \frac{1}{16}$. We can directly add that to Eq. 4, giving us $\frac{3}{8} = \pi_1 + \frac{3}{16}$ which can be reduced to $\pi_1 = \frac{3}{16}$.

We plug these back in to get our full system of equations.

$$
\begin{aligned}
0 &= -\frac{9}{64} + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 \\
0 &= \frac{3}{64} + -\frac{3}{16} + \frac{1}{4}\pi_3 + \frac{1}{4}\pi_4 \\
0 &= \frac{3}{32} + \frac{1}{8} + -\frac{3}{4}\pi_3 \\
0 &= \frac{1}{16} + \frac{1}{4}\pi_3 + -\frac{1}{2}\pi_4 \\
\frac{9}{16} &= \pi_3 + \pi_4 \\
\pi_1 &= \frac{3}{16} \\
\pi_2 &= \frac{1}{4}
\end{aligned}
$$

We can now directly solve for π_3 using Eq. 4. $3\pi_3 = \frac{7}{8} \Rightarrow \pi_3 = \frac{7}{24}$. And using π_3, we can solve for π_4. $\frac{9}{16} = \frac{7}{24} + \pi_4 \Rightarrow \frac{13}{48} = \pi_4$.

This gives us the answer that $\pi_1 = \frac{3}{16}$, $\pi_2 = \frac{1}{4}$, $\pi_3 = \frac{7}{24}$, and $\pi_4 = \frac{13}{18}$. We can also see that these all add up to 1, helping assert the correctness of our answer.

The answer to the question then is that $P(X=1)=\frac{3}{16}$, $P(X=2)=\frac{1}{4}$, $P(X=3)=\frac{7}{24}$, and $P(X=4)=\frac{13}{18}$.

Part b For part b, we want to figure out how often our factory breaks down. That is to say that from state 1, 2, how often do we transition to state 3 or 4? In other words, $P(X_n=3|X_{n-1}=1,2)+P(X_n=4|X_{n-1}=1,2)$. Using Eq. 1 from **part a** looking at just $X_n=3,4$ and $X_{n-1}=1,2$, we can calculate these values.

First let's solve for $P(X_n=3|X_{n-1}=1,2)$.

$$P(X_n=3|X_{n-1}=1,2)=\sum_{i=1}^{2}P(X_{n-1}=i)P(X_n=3|X_{n-1}=i)$$
$$=P(X_{n-1}=1)P(X_n=3|X_{n-1}=1)+P(X_{n-1}=2)P(X_n=3|X_{n-1}=2)$$
$$=\frac{3}{16}\cdot\frac{1}{2}+\frac{1}{4}\cdot\frac{1}{2}$$
$$=\frac{7}{32}$$

Next let's solve for $P(X_n=4|X_{n-1}=1,2)$.

$$P(X_n=4|X_{n-1}=1,2)=\sum_{i=1}^{2}P(X_{n-1}=i)P(X_n=4|X_{n-1}=i)$$
$$=P(X_{n-1}=1)P(X_n=4|X_{n-1}=1)+P(X_{n-1}=2)P(X_n=4|X_{n-1}=2)$$
$$=\frac{3}{16}\cdot 0+\frac{1}{4}\cdot\frac{1}{4}$$
$$=\frac{1}{16}$$

Therefore the probability that we move from a producing state to a non-producing one is $\frac{7}{32}+\frac{1}{16}=\frac{9}{32}$.

2 Problem 2

We need to determine the probability of moving from state i to state j in terms of our known model and a stochastic policy. More specifically, that is to calculate $P(X_n=j|X_{n-1}=i)$ using $\pi(a|S'=i)$ and $p(S=j,r|S'=i,a)$.

In order to take into account all possible actions at state i, we need to sum across $a \in A$. Similarly, we need to simultaneously sum the probabilities across all rewards given those actions for each $r \in R$. If we combine those all together, we get the following:

$$P(S = j | S' = i) = \sum_{a \in A} \pi(a | S' = i) \sum_{r \in R} p(S = j, r | S' = i, a)$$

3 Problem 3

Part a Because we're treating it as an episodic task where none of the future rewards are discounted, regardless of how long we take to exit the maze, we always receive a total reward of +1. What is going wrong is that it doesn't matter if we take 100 steps or 1 step, the reward is always +1 so the robot has no real incentive to find a shorter path to exit the maze. We haven't effectively communicated our desire for the robot to leave the maze as quickly as possible because we aren't punishing it for taking a long time / aren't rewarding it aptly for taking a short time. The most simple way to fix this would be to give the robot a -1 reward every time it moves without finding the exit, that way it can maximize its reward by getting to the end faster.

Part b We're asked to write the expected future value from state s in terms of the value of the expected leaf node. The purpose of this is to mathematically explain the intuition that the value of a state depends on the value of the actions possible from that state and how likely it is we take those actions under the current policy.

$$v_\pi(s) = E[G_t | S_t = s]$$

This can be rewritten to show how it takes each action into account like so:

$$\sum_a E[G_t | S_t = s, A_t = a] P(A_t = a | S_t = s)$$

We are also asked to rewrite this using $q_\pi(s, a)$ and $\pi(a|s)$.

$$v_\pi(s) = \sum_a \pi(a|s) \cdot q_\pi(s, a)$$

Part c First we want to express $q_\pi(s,a)$ in terms of expected next reward R_{t+1} and expected next state value $v_\pi(S_{t+1})$, given $S_t = s$ and $A_t = a$.

$$q_\pi(s,a) = \sum_{s'} E[R_{t+1}+G_{t+1}|S_t = s, A_t = a, S_{t+1} = s']P(S_{t+1} = s'|S_t = s, A_t = a)$$

We can then rewrite this explicitly in terms of $p(s', r|s, a)$.

$$\sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

4 Problem 4

For problem 4, we're given the optimal policy for grid world. That is that we know what move we're taking given each state, no matter what. We need to calculate the expected return, ignoring the already calculated $V*(s)$. We want to calculate the expected return at A, that is to say more specifically we want to calculate $V(2,1)$. To calculate the $V(s)$ of any state, we use the formula below.

$$V(s) = \sum_a \pi(a|s) \sum_{s\prime,a} p(s\prime, r|s, a)[r + \gamma V(s\prime)]$$

This formula says that for all possible actions we might take, given a state, using policy π, sum the probability that we end up at state $s\prime$ and receive reward r given we were at state s and took action a multiplied by the reward we received plus the discounted future reward we expect to receive from state $s\prime$.

First, we must note that for $V(1,2)$ specifically, regardless of what action we take, $s\prime$ is always $(5,2)$ and r is always 10, simplifying $V(2,1)$ greatly. Similarly, for $V(2,3,4,5,2)$, the only action we take is up and always with 100% certainty so we only need to look at $r + \gamma V(s\prime)$ for each $V(x,2)$. So let's write out all of our equations with this in mind:

$$
\begin{gathered}
V(1,2) = 10 + \gamma V(5,2) \\
V(2,2) = 0 + \gamma V(1,2) \\
V(3,2) = 0 + \gamma V(2,2) \\
V(4,2) = 0 + \gamma V(3,2) \\
V(5,2) = 0 + \gamma V(4,2)
\end{gathered}
$$

We can see that these are circular in nature, so we should try to write $V(1,2)$ in terms of itself.

$V(1,2) = 10 + \gamma(0 + \gamma V(4,2)) \Rightarrow V(1,2) = 10 + \gamma(0 + \gamma(0 + \gamma V(3,2)) \Rightarrow V(1,2) = 10 + \gamma(0 + \gamma(0 + \gamma(0 + \gamma V(2,2)) \Rightarrow V(1,2) = 10 + \gamma(0 + \gamma(0 + \gamma(0 + \gamma(0 + \gamma V(1,2)) = 10 + \gamma^5 V(1,2)$

Now that we have it in terms of $V(1,2)$ on both sides, we can expand it to see if we can find an infinite series.

$V(1,2) = 10 + \gamma^5(10 + \gamma^5(10 + ...)) = 10\gamma^0 + 10\gamma^5 + 10\gamma^{10} + ... = \sum_{i=0}^{\infty} 10\gamma^{5i}$

Because γ is assumed to be less than 1, this is a geometric series of the form $\sum_{i=0}^{\infty} ar^i$ where $a = 10$ and $r = \gamma^5$ that will converge. Assuming convergence, the solution to a geometric series is $\frac{a}{1-r} = \frac{10}{1-\gamma^5}$. Assuming that $\gamma = 0.9$ like the examples in class, then the answer is $\frac{10}{1-0.9^5} = 24.419428097 = 24.419$.