CSCI-UA 472 Artificial Intelligence

Muhammad Wajahat Mirza

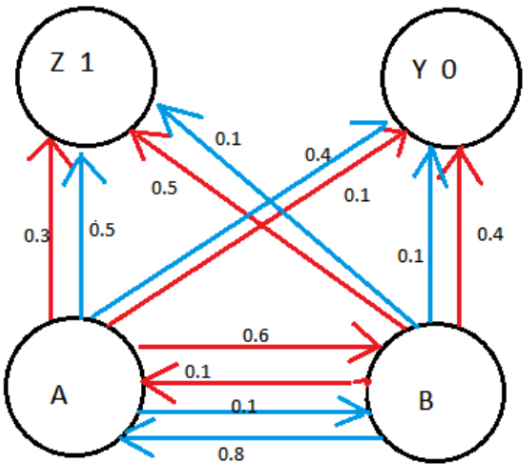
mwm356@nyu.edu

Homework 05

October 15, 2020

Problem 1

Consider the simple Markov Decision Process shown below:



The two states Z and Y are terminal states with values 1 and 0. The states A and B are non-terminal states. In each state, two actions are possible: red and blue. The transition probabilities are shown in this table and on the graph.

	Z	Y	A	B
A blue	0.5	0.4	0	0.1
A red	0.3	0.1	0	0.6
B blue	0.1	0.1	0.8	0
B red	0.5	0.4	0.1	0

That is, if the state is A and the agent chooses the blue action, then there is a 0.5 probability that the next state is Z; a 0.4 probability that the next state is Y; and a 0.1 probability that

the next state is Y. Likewise for the other states and actions.

Suppose that initially the policy is that both agents choose the blue action. Under that policy, the expected values of the two states A and B observe the equations:

$$\begin{cases} A = 0.5 \cdot 1 + 0.4 \cdot 0 + 0.1 \cdot \mathrm{B} \\ B = 0.1 \cdot 1 + 0.1 \cdot 0 + 0.8 \cdot \mathrm{A} \end{cases}$$

The solution is A = 0.5543; B = 0.5435.

Trace the execution of the MDP algorithm from that point. (Note: You may use any computational aids you want to solve the simultaneous linear equations; you do not have to do that by hand.)

Solution to Problem 1

Original given policy: both agents choose blue action {A:b, B:b}

$$\begin{cases} A = 0.5 \cdot 1 + 0.4 \cdot 0 + 0.1 \cdot \mathrm{B} \\ B = 0.1 \cdot 1 + 0.1 \cdot 0 + 0.8 \cdot \mathrm{A} \end{cases}$$

The solution is A = 0.5543; B = 0.5435.

Using A = 0.5543; B = 0.5435, under this policy, recalculate the expected values using state names.

$$\begin{cases} \{A:b\} = 0.5 \cdot 1 + 0.4 \cdot 0 + 0.1 \cdot 0.5435 & = 0.5544 \\ \{A:r\} = 0.3 \cdot 1 + 0.1 \cdot 0 + 0.6 \cdot 0.5435 & = 0.6261 \\ \{B:b\} = 0.1 \cdot 1 + 0.1 \cdot 0 + 0.8 \cdot 0.5543 & = 0.5434 \\ \{B:r\} = 0.5 \cdot 1 + 0.4 \cdot 0 + 0.1 \cdot 0.5543 & = 0.5554 \end{cases}$$

Using these new expected values, we get the maximal expected values of agents A, B and their actions to choose next policy.

$$\{A:r\} > \{A:b\}$$

i.e.

$$0.6261 > 0.5543$$

Also

$$\{B:r\} > \{B:b\}$$

i.e.

$$0.5554 > 0.5435$$

Therefore, new policy is both agents choose the red action i.e. {A:r, B:r} and get the new equations.

$$\begin{cases} A = 0.3\cdot\ 1 + 0.1 \cdot 0 + 0.6 \cdot \mathrm{B} \\ B = 0.5\cdot\ 1 + 0.4 \cdot 0 + 0.1 \cdot \mathrm{A} \end{cases}$$

The solution is A = 0.6383; B = 0.5638.

Using A = 0.6383; B = 0.5638, under this policy, recalculate the expected values using state names.

$$\begin{cases} \{A:b\} = 0.5\cdot\ 1 + 0.4 \cdot 0 + 0.1 \cdot 0.5638 & = 0.5564 \\ \{A:r\} = 0.3\cdot\ 1 + 0.1 \cdot 0 + 0.6 \cdot 0.5638 & = 0.6383 \\ \{B:b\} = 0.1\cdot\ 1 + 0.1 \cdot 0 + 0.8 \cdot 0.6383 & = 0.6106 \\ \{B:r\} = 0.5\cdot\ 1 + 0.4 \cdot 0 + 0.1 \cdot 0.6383 & = 0.5638 \end{cases}$$

Using these new expected values, we get the maximal expected values of agents A, B and their actions to choose next policy.

$$\{A:r\} = \{A:r\}$$

i.e.

$$0.6383 = 0.6383$$

Also

$$\{B:b\} > \{B:r\}$$

i.e.

$$0.6106 > 0.5638$$

Therefore, new policy is agent A chooses the red action and agent B chooses the blue actions i.e. {A:r, B:b} and get the new equations.

$$\begin{cases} A = 0.3\cdot\ 1 + 0.1 \cdot 0 + 0.6 \cdot \mathrm{B} \\ B = 0.1\cdot\ 1 + 0.4 \cdot 0 + 0.8 \cdot \mathrm{A} \end{cases}$$

The solution is A = 0.6923; B = 0.6538.

Using A = 0.6923; B = 0.6538, under this policy, recalculate the expected values using state names.

$$\begin{cases} \{A:b\} = 0.5\cdot\ 1 + 0.4 \cdot 0 + 0.1 \cdot 0.6538 & = 0.5654 \\ \{A:r\} = 0.3\cdot\ 1 + 0.1 \cdot 0 + 0.6 \cdot 0.6538 & = 0.6923 \\ \{B:b\} = 0.1\cdot\ 1 + 0.1 \cdot 0 + 0.8 \cdot 0.6923 & = 0.6538 \\ \{B:r\} = 0.5\cdot\ 1 + 0.4 \cdot 0 + 0.1 \cdot 0.6923 & = 0.5692 \end{cases}$$

Using these new expected values, we get the maximal expected values of agents A, B and their actions to choose next policy.

$$\{A : r\} = \{A : r\}$$

i.e.

$$0.6923 = 0.6923$$

Also

$$\{B : b\} = \{B : b\}$$

i.e.

$$0.6538 = 0.56538$$

Therefore, our new policy is agent A chooses the red action and agent B chooses the blue actions i.e. {A:r, B:b}. Since the policy is unchanged in this iteration, we terminate our MDP algorithm. Hence, in pursuit of best strategy, our policy evolves as such:

Step 01:

$$\{A : b, B : b\}$$

Step 02:

$$\{A : r, B : r\}$$

Step 03:

$$\{A : r, B : b\}$$

Step 04:

$$\{A : r, B : b\}$$

Terminate!

End of Assignment. Thank you!