task_12.3

Machine Learning (WiSe 2025/2026)

Author: Suvansh Shukla
Matriculation No. 256245

Assignment 12 Task 3

Part A

Given:

- γ = 0.8
- S_8 is an absorbing goal state

Equation for Value function =

$$V(s) \leftarrow max_a \{r(s,a) + \sum_{s'} Pr[s'|s,a] \gamma V(s')\}$$

However, since we're given that the world is deterministic we can simplify that into:

$$V(s) \leftarrow \underset{a}{max}[r(s,a) + \gamma V(s')]$$

This can be done because a deterministic world means that every state-action pair (s,a) leads to exactly one succeeding state δ(s,a) with 100% probability.

Starting with S_8 then moving in decreasing order:

- S_9 = 10
- S_8 = 0
- S_7 = 10
- S_6 = 2 + (0.8 * 10) = 10
- S_5 = 10
- S_4 = 0 + (0.8 * 10) = 8
- S_3 = 2 + (0.8 * 8) = 8.4
- S_2 = 0 + (0.8 * 8) = 6.4
- S_1 = 0 + (0.8 * 8.4) = 6.72

An optimal policy (π*) is the action selection rule that maximizes the expected sum of future discounted rewards for an agent, starting from any given state

Q function represents the expected reward of taking a specific action a in state s and thereafter following the optimal policy.

To calculate values for the Q table we'll approach things in a backward manner:

Starting from S9 we see that the reward for transitioning to S8 is 10 and since it has no further actions after the goal it's V* is 0.

So, the Equation becomes:

$$S_9(left) = 10 + 0 = 10$$

Using the same logic for the following states:

$$S_7(right) = 10 + 0 = 10$$

$$S_5(down) = 10 + 0 = 10$$

Now for State 6 we see that it has a transitioning reward of 2 with a maximum further action reward of 10.

So the Equation becomes:

$$S_6(down) = 2 + (0.8 * 10) = 10$$

Using this we can calculate the transitioning Q for

$$S_9 \; to \; S_6 \; upwards$$

:

$$S_9(up) = 0 + (0.8 * 10) = 8$$

And now for

$$S_6 \; leftward$$

:

$$S_6(left) = 0 + (0.8 * 10) = 8$$

Now carrying onwards for other states:

$$S_4(right) = 0 + (0.8 * 10) = 8$$

$$S_4(down) = 0 + (0.8 * 10) = 8$$

$$S_5(left) = 0 + (0.8 * 8) = 6.4$$

Notice how the transition value went down for going from S5 to S4.

$$S_3(right) = 2 + (0.8 * 8) = 8.4$$

$$S_2(down) = 0 + (0.8 * 8) = 6.4$$

$$S_4(left) = 0 + (0.8 * 8.4) = 6.72$$

$$S_4(up) = 0 + (0.8 * 6.4) = 5.12$$

$$S_1(down) = 0 + (0.8 * 8.4) = 6.72$$

$$S_1(right) = 0 + (0.8 * 6.4) = 5.12$$

$$S_3(up) = 0 + (0.8 * 6.72) = 5.376$$

$$S_2(left) = 0 + (0.8 * 6.72) = 5.376$$

Part B

Perhaps taking the episodic reward function would bring about the changes that are required.