1.

Question 1

The learner and decision maker is the _______.

Environment

State

Reward

Agent

Correct!

1 / 1 point

2.

Question 2

At each time step the agent takes an _______.

State

Reward

Action

Correct!

Environment

1 / 1 point

3.

Question 3

Imagine the agent is learning in an episodic problem. Which of the following is true?

The number of steps in an episode is stochastic: each episode can have a different number of steps.

Correct!

The number of steps in an episode is always the same.

The agent takes the same action at each step during an episode.

1 / 1 point

4.

Question 4

If the reward is always +1 what is the sum of the discounted infinite return when

$\gamma<1$

$\gamma<1$

gamma, is less than, 1

$G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$

G

t

$=\sum$

$k=0$

∞

γ

k

R

$t+k+1$

G, start subscript, t, end subscript, equals, sum, start subscript, k, equals, 0, end subscript, start superscript, infinity, end superscript, gamma, start superscript, k, end superscript, R, start subscript, t, plus, k, plus, 1, end subscript

Gt=1*γk

G

t

$=1*\gamma$

k

G, start subscript, t, end subscript, equals, 1, times, gamma, start superscript, k, end superscript

Gt=11−γ

G

t

$=$

$1-\gamma$

1

G, start subscript, t, end subscript, equals, start fraction, 1, divided by, 1, minus, gamma, end fraction

Correct!

Infinity.

Gt=γ1−γ

G

t

$=$

$1-\gamma$

γ

G, start subscript, t, end subscript, equals, start fraction, gamma, divided by, 1, minus, gamma, end fraction

1 / 1 point

5.

Question 5

How does the magnitude of the discount factor (gamma/

γ

γ

gamma

) affect learning?

With a smaller discount factor the agent is more far-sighted and considers rewards farther into the future.

The magnitude of the discount factor has no effect on the agent.

With a larger discount factor the agent is more far-sighted and considers rewards farther into the future.

Correct!

1 / 1 point

6.

Question 6

Suppose

γ=0.8

γ=0.8

gamma, equals, 0, point, 8

and the reward sequence is

R1=5

R

1

$=5$

R, start subscript, 1, end subscript, equals, 5

followed by an infinite sequence of 10s. What is

$G0$

G

0

G, start subscript, 0, end subscript

?

45

Correct!

$G2=10/(1-0.8)=50$

G

2

$=10/(1-0.8)=50$

G, start subscript, 2, end subscript, equals, 10, slash, left parenthesis, 1, minus, 0, point, 8, right parenthesis, equals, 50

$G1=10+.8*(50)=50$

G

1

=10+.8*(50)=50

G, start subscript, 1, end subscript, equals, 10, plus, point, 8, times, left parenthesis, 50, right parenthesis, equals, 50

G0=5+.8*50=45

G

0

$=5+.8*50=45$

G, start subscript, 0, end subscript, equals, 5, plus, point, 8, times, 50, equals, 45

55

15

1 / 1 point

7.

Question 7

What does MDP stand for?

Meaningful Decision Process

Markov Deterministic Policy

Markov Decision Protocol

Markov Decision Process

Correct!

1 / 1 point

8.

Question 8

Suppose reinforcement learning is being applied to determine moment-by-moment temperatures and stirring rates for a bioreactor (a large vat of nutrients and bacteria used to produce useful chemicals). The actions in such an application might be target temperatures and target stirring rates that are passed to lower-level control systems that, in turn, directly activate heating elements and motors to attain the targets. The states are likely to be thermocouple and other sensory readings, perhaps filtered and delayed, plus symbolic inputs representing the ingredients in the vat and the target chemical. The rewards might be moment-by-moment measures of the rate at which the useful chemical is produced by the bioreactor.

Notice that here each state is a list, or vector, of sensor readings and symbolic inputs, and each action is a vector consisting of a target temperature and a stirring rate.

Is this a valid MDP?

Yes. Assuming the state captures the relevant sensory information (inducing historical values to account for sensor delays). It is typical of reinforcement learning tasks to have states and actions with such structured representations; the states might be constructed by processing the raw sensor information in a variety of ways.

Correct!

No. If the instantaneous sensor readings are non-Markov it is not an MDP: we cannot construct a state different from the sensor readings available on the current time-step.

1 / 1 point

9.

Question 9

Case 1: Imagine that you are a vision system. When you are first turned on for the day, an image floods into your camera. You can see lots of things, but not all things. You can't see objects that are occluded, and of course you can't see objects that are behind you. After seeing that first scene, do you have access to the Markov state of the environment?

Case 2: Imagine that the vision system never worked properly: it always returned the same static imagine, forever. Would you have access to the Markov state then? (Hint: Reason about

$P(St+1 \mid St,...,S0)$

$P(S$

$t+1$

$\mid S$

t

$,\ldots,S$

0

)

P, left parenthesis, S, start subscript, t, plus, 1, end subscript, vertical bar, S, start subscript, t, end subscript, comma, point, point, point, comma, S, start subscript, 0, end subscript, right parenthesis

, where

St

S

t

S, start subscript, t, end subscript

= AllWhitePixels)

You have access to the Markov state in both Case 1 and 2.

You have access to the Markov state in Case 1, but you don't have access to the Markov state in Case 2.

Incorrect. Because there is no history before the first image, the first state has the Markov property. The Markov property does not mean that the state representation tells all that would be useful to know, only that it has not forgotten anything that would be useful to know.

The case when the camera is broken is different, but again we have the Markov property. All the possible futures are the same (all white), so nothing needs to be remembered in order to predict them.

You don't have access to the Markov state in Case 1, but you do have access to the Markov state in Case 2.

You don’t have access to the Markov state in both Case 1 and 2.

1 point

10.

Question 10

What is the reward hypothesis?

Ignore rewards and find other signals.

That all of what we mean by goals and purposes can be well thought of as the minimization of the expected value of the cumulative sum of a received scalar signal (called reward)

That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)

Correct!

Always take the action that gives you the best reward at that point.

1 / 1 point

11.

Question 11

Imagine, an agent is in a maze-like gridworld. You would like the agent to find the goal, as quickly as possible. You give the agent a reward of +1 when it reaches the goal and the discount rate is 1.0, because this is an episodic task. When you run the agent its finds the goal, but does not seem to care how long it takes to complete each episode. How could you fix this? (**Select all that apply**)

Give the agent -1 at each time step.

> Correct! Giving the agent a negative reward on each time step, tells the agent to complete each episode as quickly as possible.

Give the agent a reward of +1 at every time step.

Give the agent a reward of 0 at every time step so it wants to leave.

Set a discount rate less than 1 and greater than 0, like 0.9.

Correct! From a given state, the sooner you get the +1 reward, the larger the return. The agent is incentivized to reach the goal faster to maximize expected return.

1 / 1 point

12.

Question 12

When may you want to formulate a problem as continuing?

When the agent-environment interaction does not naturally break into sequences. Each new episode begins independently of how the previous episode ended.

Correct!

When the agent-environment interaction naturally breaks into sequences and each sequence begins independently of how the previous sequence ended.

1 / 1 point

Like

Dislike

Report an issue