﻿

Question 1

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

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

2. The agent takes the same action at each step during and

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

Question 2

How does the magnitude of the discount factor (gammay) affecting? 1. The magnitude of the decount factor has no effect on the agent

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

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

﻿ Question 3

An NDP is based on the Markov assumption, which can be formulated as "The future is independent of the past given the present

﻿

Question 4

Case 1 imagine that you are a vision system. When you are turned on for the day an image foods 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 cant see objects that are befund you. After seeing that test scene, do you have access to the Markov state of the environment?

Case 2 Imagine that the vision system never worked property it always retumed the same static imagine, forever. Would you have access to the Markov state then?

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

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

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

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

﻿

Question 5

2 points

Let's say that you have a robot and you want that robot to learn how to cook a pancake and how to knead some dough and how to mix you a drink and how to do the laundry and whatever night, so you have many different tasks that you want to learn. So perhaps we learn at these tasks individually, one after the other, maybe all the different tasks can benet from one another and therefore accelerate each other.

True Or False

﻿

Question 6

How is a model-free prediction different from a model-based prediction?

1. Model-based prediction is a prediction which is based on a model

2.For model the prediction, the agent has full knowledge of the environment while for model-based prediction, the agent has no idea of what the environment looks like.

3. Model-free prediction is the prediction of a free model

4. For model prediction, the agent has no idea of what the environment looks like while for model based prediction, the agent has full knowledge of the environment.

﻿ ﻿ ﻿

Question 7

Reaching diverse goals is same as performing diverse tasks

True or False

Question 8

At least one deterministic optimal policy exists in every Markov decision process

True or False

Question 9

Does adding a constant to all rewards change the set of optimal policies in episodic tasks?

True or False

Question 10

The goal of a reinforcement learning agent is to maximize the immediate reward

True or False

﻿

Question 11

Q-learning is a value-based method

True or False

﻿

Question 12

Which one could be true while performing supervised learning in RL (Select one or more options)

a. The tasks we can perform in source domain, same we can perform in the target domain

b.Fine tuning process may still need to explore, but optimal policy may be determinist

c.Representations learned in source domain works well in target domain

d.None

﻿

Question 13

2 points

See Answer Consider using enforcement learning to control the motion of a robot arm to pick up objects and place them into new positions. The actions in this case might be the voltages appled to each motor at each joint, and the states might be the latest readings of joint angles and velocities. The reward might be for each object successfully picked up and placed. To encourage smooth movements, on each time step a small, negative reward can be given as function of the moment to moment jerkness of the motion is this a valid MDP? True or False

Question 14:

﻿In \_\_\_\_\_\_\_\_\_\_we explore using counts and utlize different models to estimate densities.

﻿

Question 15 (Picture based Question)

Consider the undiscounted episode MDP below There are four actions possible in each state, A- up, down, night, left, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid intact leave the state unchanged. The right half of the figure shows the value of each state under the equprobable random policy pit is the equprobable random policy, what is down?

﻿

Question 16

In Reinforcement Learning which of the following are theoretically intractable?

1. ALL
2. Multi-armed bandits
3. Contextual bandits
4. Large, Infinite MDPs

﻿

Question 17

Suppose lambda - 0.8 and we observe the following sequence of rewards: R-3, R2-5, R-2.R.-7. and R-1, with T-5. What is Go? (Hint: Work Backwards and recall that G, that Gt = Rt+1 + lambda Gt+1.)

a.1.11502

b.2,-3

c.3,6.2736

d.4,8.24

﻿

Question 18

Define the optimal state-value function Vs) for an MDP. write the equation

﻿

Question 19

An experience sample is composed of?

1.State,action, discount factor reward

2.State, action, next state, next action

3. State, action, next state, reward

﻿ Question 20

What brings about the "Deep" in Deep Reinforcement Leaming?

1. a dеереr reinforcement learning

2.the use of deep learning which involves neural networks

3. it has no meaning

4. represents more complex reinforcement learning problems

﻿

Question 21

How is the expected SARSA algorithm similar to the Sarsa algorith

1. They both update the policy after every episode

2. They are not similar

3. They are both on policy temporal difference algorithms

4.They are both off policy TD algorithms

﻿

Question 22

Which of the following are not true about optimistic exploration (Mutiple options can be selected)

1.Learn distribution over-functions or policies

2 requires estimating state visitation frequencies or novelty

3.typically realized by means of exploration bonuses

4 sample and act according to sample

﻿

Question 23

The leaner and decision maker is the

01. Reward

02.State

03. Agent

04. Environment

﻿

Question 24

if the exploration is so great why did epsilon of 0.0 greedy agent) perform better than epsilon of 0.4?

1.Epsilon of 0.4 does not explore often enough to find the optimal action

2.None

3.Epsilon of 0.0 is greedy, thus it will always choose the optimal am

4.Epsilon of 0.4explores too often that takes many sub-optimal actions causing it to do worse over the long term

﻿ ﻿

Question 25

Agents are the main part of Reinforcement Lang How many agents can eat?

a.As many as possible

b.5

c.10

d.2

Question 26

Which kind of approaches are more suited for continuous action problems?mple answers)

01.Policy based methods

02. Value-based methods

03.Actor methods

04.None

﻿

Question 27

Assertion Actor-critic updates have lesser variance than enforce updates

Reason Actor-critic methods use TD target instead of Gt

1.Both Assertion and Reason are true, and Reason is correct explanation for Assertion

2.Both Assertion and Reason are true, and Reason is not correct explanation for Assertion

3.Assertion is true Reasons is False

4.Both Assertion and Reason are False

﻿

Question 28

While learning behaviors to prepare for an unknown future goal, we need to have knowledge of reward function

True Or False

﻿

Question 29

Image, 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 reward of 1 when it reaches the goal and the discount rate is 1.0, because this is an episode task. When you run the agent it finds the goal but does not seems to care how long it takes to complete each episode. How could you fix this?

1.Set a discount rate less than 1 and greater than 0.9

2. Give the agent-1 at each time step

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

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

Question 30

Select all the option(s) that are exploration examples

a. all the options

b.in online Ad placement showing the successful advertisement

c. Showing random advent

d. Drill the of at a new location

e. Trying a different Hotel

Question 31

In RL we use the knowledge from prior tasks to acquire new knowledge on new task. How is this knowledge stored?

Match them in left column with the right column:

a.Policy

b.Q-function

c. Models

d.Features

Answer should be provided as: a-1, b-2, c-3,

1. Provides us with a good representation

2. Tells us some of the laws of physics that govern the world

3 It tells us which actions or states are good

4.It tells us which actions are potentially useful

﻿Question 32

Value iteration and Policy iteration algorithms can be used even when the state transition model is unknown

True Or False﻿

Question 33

Where can you use dynamic programming? Discuss problems that you have encountered that could be solved with dynamic programming methods. What are the advantages of using DP to solve these? What are the disadvantages?

Question 34

why did epsilon of 0.1 perform better of 1000 steps than epsion of 0.017

a. 0.01 agent didn't explore enough. Thus it ended up selecting a suboptimal for longer

b. Epsilon of 0.1 is the optimal value for epsilon general

c. The 001 agent explored too much causing the arm to choose bad action too often

d. All of the options

﻿

Question 35

What is a Bellman Equation?﻿

Question 36

Consider an undiscounted Markov Reward Process with two states A and B. The transition matrix and reward function are unknown, but you have observed two sample episodes:

X+3->X+2🡪T-4->A+4->terminate

B-2->A+3->B-3->terminate

In the above episodes, sample state transitions and sample rewards are shown at each step For instance, X-3X indicates a transition from state X to state X with a reward of 3

1. Using first visit Monte-Carlo evaluation estimate the state value function(A), V(B)
2. Using every-vis Monte-Carlo evaluation, estimate the state-value function V(A), V(B)

﻿

Question 37

If the reward is always what is the sum of the discounted infinite when lambda < 1

Ans : ()

﻿

Question 38

Which of these does not make up the SARSA algorithm?

1. Neural Networks

2.Actions

3. Rewards

4. States

Question 39

What makes a policy on greedy?

1.Greedy policies on the other hand only take actions that lead to the best rewards

2.Apolicy is epsilon greedy when it does not take only actions that lead to high rewards.

Question 40

Out of the following options, which one is are the best objects) to?

1. Policy

b. Model

C. all

d. Value Function