|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Roll No. |  |  |  |  |  |  |  |  |  |  |  |  |



**PRESIDENCY UNIVERSITY**

**Bengaluru**

|  |
| --- |
| **End - Term Examinations –JANUARY 2025** |
| **Date:** 16- 01- 2025 **Time:** 9:30 am – 12:30 pm |

|  |  |
| --- | --- |
| **School:** SOIS | **Program:** BSD |
| **Course Code :**CSA3074 | **Course Name :**REINFORCEMENT LEARNING |
| **Semester**: V | **Max Marks**:100 | **Weightage**:50% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CO - Levels** | **CO1** | **CO2** | **CO3** | **CO4** | **CO5** |
| **Marks** | **16** | **26** | **34** | **24** |  |

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *Do not write anything on the question paper other than roll number.*

**Part A**

|  |
| --- |
| **Answer ALL the Questions. Each question carries 2marks. 10Q x2M=20M** |
| **1** | Define ‘reward’ and ‘return’ for an episodic task with an example for each | **2 Marks** | **L1** | **CO1** |
| **2** | Differentiate deterministic and stochastic environments with an example each | **2 Marks** | **L2** | **CO1** |
| **3** | Classify model based learning and model free learning | **2 Marks** | **L2** | **CO2** |
| **4** | Define the following 1. Goal b)Action space
 | **2 Marks** | **L1** | **CO1** |
| **5** | Write the equation to find V(S) in the Monte Carlo method | **2 Marks** | **L2** | **CO2** |
| **6** | Define a policy | **2 Marks** | **L1** | **CO2** |
| **7** | State the SARSA update rule | **2 Marks** | **L1** | **CO3** |
| **8** | Distinguish between SARSA and Q-Learning algorithms | **2 Marks** | **L2** | **CO3** |
| **9** | Illustrate upper confidence bound of an arm in MAB Problem | **2 Marks** | **L2** | **CO4** |
| **10** | Interpret Thompson Sampling | **2 Marks** | **L2** | **CO4** |

**Part B**

|  |
| --- |
|  **Answer all the Questions Total 80 Marks.** |
| **11.** | **a.** | Identify Bellman equation to the value function of a state in a deterministic environment and stochastic environment . Explain each term in it. Find the value of all the states in the trajectory given below using Bellman equation. Assume Y=1 | **10 Marks** | **L2** | **CO1** |
| **b** | Compute final value of states for the given policyFirst iteration: | **10 Marks** | **L3** | **CO2** |
| **Or** |
| **12.** | **a.** | Discuss stochastic environment and deterministic environment in RL with an example. | **10 Marks** | **L2** | **CO1** |
| **b** | Using the Model dynamics table of State A determine the optimal policy using Policy Iteration. | **10 Marks** | **L3** | **CO2** |
|  |  |  |  |  |  |
| **13.** | **a.** | For the following grid world Environment Calculate the Value function which follows a deterministic policy. | **10 Marks** | **L3** | **CO2** |
| **b.** | Discuss TD prediction algorithm in FZLE environment | **10 Marks** | **L2** | **CO3** |
| **Or** |
| **14.** | **a.** | Articulate TD Control algorithm in FZLE environment | **10 Marks** | **L3** | **CO2** |
| **b.** | Summarize Markov Decision Process in Detail | **10 Marks** | **L2** | **CO3** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **15.** | **a.** | Articulate the Algorithmic steps of the off policy TD Control-Q Learning technique .Also find the updated Q Value of state (3,2) using the following data , apply epsilon greedy technique (choose exploitation) to choose current and next actions. Assume alpha as 0.1 and gamma as 1 | **10 Marks** | **L3** | **CO3** |
| **b** | Discuss Thompson Sampling to overcome the exploration –exploitation dilemma with the algorithm | **10 Marks** | **L2** | **CO4** |
| **Or** |
| **16.** | **a.** | Determine how Upper confidence bound strategy is used to overcome the exploration – exploitation dilemma with the algorithm | **10 Marks** | **L3** | **CO3** |
| **b** | Describe TD control and its types in brief | **10 Marks** | **L2** | **CO4** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **17.** | **a.** | Interpret SARSA in FZLE | **10 Marks** | **L2** | **CO3** |
| **b.** | Articulate softmax exploration method of exploration-exploitation strategy for the following 4 arm bandit  | **10 Marks** | **L3** | **CO4** |
| **Or** |
| **18.** | **a.** | Contrast the DP, MC and TD methods | **10 Marks** | **L2** | **CO3** |
| **b.** | Articulate contextual Bandits and list out the applications of MAB | **10 Marks** | **L3** | **CO4** |

**\*\*\*\*\* BEST WISHES \*\*\*\*\***