2811N850

Candidate's Seat No :_____

IM.Sc DS Sem.-9 Examination

CC 504

Reinforcement Learning

November-2024

Time : 2-30 Hours] **[Max. Marks : 70**

Instructions: All questions are compulsory. Use of non-programmable scientific calculator is allowed.

Q.1 **(a)** What is reinforcement learning? What are some common challenges in reinforcement learning? **(07)**

(b) Explain the concept of the K-Armed Bandit Problem and how it relates to Sequential Decision Making with Evaluative Feedback in RL. **(07)**

OR

(a) What are the trade-offs between Exploration and Exploitation in Reinforcement Learning, and why is it important to address them? **(07)**

(b) Consider the Frozen Lake environment where the reward for reaching the goal is 1, and the discount factor $\gamma = 0.95$. The transition probabilities from state S_5 when moving right are: (i) To S_6 with probability 0.8 (successful move), (ii) To S_5 with probability 0.1 (no move), (iii) To S_9 with probability 0.1 (slip down). If $V(S_6) = 0.8$ and , $V(S_9) = 0.2$, calculate the value of state S_5. **(07)**

Q.2 **(a)** Discuss the efficiency of the Dynamic Programming. **(07)**

(b) Explain Markov Decision Process (MDP) with an example. **(07)**

OR

(a) Provide comparison of all Bandits algorithms. **(07)**

(b) Assume a MDP with two states S_1 and S_2 , and two actions a_1 and a_2 . The transitions probabilities and rewards are given below: **(07)**

- From S_1 with a_1 : transitions to S_2 with probability 1, reward = 5.
- From S_2 with a_2 : transitions to S_1 with probability 1, reward = 2.

Assume initial value of 0 for both states and a discount factor $\gamma = 0.9$. Use Bellman Equation to calculate the value functions for S_1 and S_2 after one iteration.

Q.3 **(a)** In the context of reinforcement learning, explain the concept of prediction, control, on-policy and off-policy. **(07)**

(b) Explain in detail Generalized Policy Iteration (GPI) framework. **(07)**

OR

(a) What is importance sampling? Provide a detailed description of the two importance sampling methods used in offline learning. **(07)**

(b) What are the advantages of TD prediction methods. **(07)**

(P.T.O)

Q.4 **(a)** Explain how Monte Carlo Policy Evaluation operates, including a detailed description of the algorithm. **(07)**

(b) How does maximization bias arise in value function estimation? What strategies, can be implemented to mitigate this bias? **(07)**

OR

(a) Consider a finite, episodic and undiscounted Markov Decision Process (MDP) with states P and Q apart from the terminal state. Let the following two samples are observed when a Monte-Carlo (MC) evaluation is being carried out. **(07)**

- $(P,+3)\rightarrow(P,+2)\rightarrow(Q,-4)\rightarrow(P,+4)\rightarrow(Q,-3)$
- $(Q,-2)\rightarrow(P,+3)\rightarrow(Q,-3)$

For example, a sample such as, $(Q,-2)\rightarrow(P,+3)\rightarrow(Q,-3)$ means that the episode starts at Q then goes to P again, then goes to Q and then terminates. On the way, the agent gets rewards of -2, $+3$ and -3, respectively.

(i) Estimate the state value of both P and Q using *first-visit* Monte-Carlo evaluation.

(ii) Estimate the state value of both P and Q using *every-visit* Monte-Carlo evaluation.

(b) Compare on-policy and off-policy methods. **(07)**

Q.5 Attempt any **SEVEN** out of **TWELVE:** **(14)**

(1) State the primary component of the reinforcement learning system.

(2) If Arm A has an average reward of 2.8 after 8 pulls, and Arm B has an average reward of 3.2 after 5 pulls, which arm would you select using the upper confidence bound (UCB) approach, assuming both have been pulled for at least one round? Take $c = 2$.

(3) State Markov property.

(4) Suppose $\gamma = 0.8$ and the sequence of rewards is received $R_1 = 2, R_2 = 3$ and $R_3 = 5$ with $T = 3$. What are the values for G_0, G_1, and G_2?

(5) What is the significance of the discount factor in reinforcement learning?

(6) What is the primary difference between value-based and policy-based methods in reinforcement learning?

(7) What is the main principle behind Monte Carlo methods in reinforcement learning?

(8) State with justification whether the following decision task has Markov Property.

Choosing a medical treatment for a patient

(9) In SARSA, if you have the following values: $Q(S,A) = 2$ and reward R = 1, and the next action value $Q(S',A') = 3$ with $\gamma = 0.9$ and $\alpha = 0.5$. Calculate the updated value $Q(S,A)$.

(10) What is the main advantage of using TD learning over dynamic programming?

(11) In TD learning, how is the value function updated using the Bellman equation?

(12) Why is Q-leaning off-policy?

**** ****