2012E1562

Candidate's Seat No:____

MSc (AIML) Sem.-3 Examination
Reinforcement Learing
December-2025

Time : 3.00 Hours] **[Max.Marks :100**

- **Write both the sections in a separate answer book.**
- **Both sections have equal weightage.**
- **Draw diagrams wherever necessary.**
- **Make assumptions wherever necessary.**

SECTION-I

Q.1 If today's weather is sunny, there is a probability of 0.3 to have a cloudy day tomorrow with no chance of rain. If today is cloudy the probabilities of rainy and sunny on tomorrow are 0.3 & 0.2 respectively. Rain will continue through tomorrow with a probability of 0.6 and there is a 0.2 probability that it will be sunny. **10**

a. Draw the state transition diagram.
b. Write the steps for transition probability matrix.
c. Find steady state.
d. It is given that yesterday was a sunny day, find probability that it will not rain tomorrow.

Q.2 Attempt any three from following:

1. Define Reinforcement learning and Explain the main elements of RL over which agent rely? **30**
2. What is the n-arm bandit problem? Name and explain different algorithms used to select action in n-arm bandit problem?
3. Explain Incremental update rule for stationary and non-stationary problems in reinforcement learning?
4. Explain the concept of policy evaluation. How is it implemented in dynamic programming?

Q.3 Calculate V(A), V(B), V(C) using both methods (Bellman equations and iterative policy evaluation). The discount factor is γ=0.6, The immediate rewards are: R(A)=2, R(B)=−1, R(C)=0 **10**

	From		
To	**A**	**B**	**C**
A	**0.3**	**0.2**	**0.5**
B	**0.4**	**0.5**	**0.3**
C	**0.3**	**0.3**	**0.2**

a. Write the Bellman expectation equations for three states and solve them.
b. Starting with V(A)=V(B)=V(C)=0, perform iterative policy evaluation updates up to 3 steps.

E1562-2

SECTION-2

Q.4 **Attempt any five from the following provide example where asked :** **25**

1. Describe and explain the process of Generalize policy iteration (GPI) and how it is used to find an optimal policy.
2. How does actor-critic method combine policy-based and value-based approaches?
3. Differentiate between MC, DP and TD. What is eligibility traces, explain the process of n-step TD prediction.
4. Explain the concepts of selection, crossover, and mutation in genetic algorithms.
5. Explain the process of Q-learning and updation rule with example.
6. Explain how a perceptron can be used as a linear function approximator. What are the limitations of using perceptron's for function approximation?

Q.5 Read the situation given for a grid world problem and do as asked in the questions: **25**
The states are grid square, it has rows and column number respectively. The agent starts its journey from state (1,1) marked with letter S. There are three terminal state, goal at (3,3) with reward as +10 and the other two terminals are obstacles at (1,3) and (3,1) with reward -5, all other non-terminal states have reward as 1. The Transition function is such that the intended agent movement (up/ down/ left/ right) happens with probability 0.6. With probability 0.2 each the agent ends up in one of the states perpendicular to the intended move. If collision with obstacle happens, agent stays in same state.
States' S= {(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)}
P(s'/(s, a)) = {(intended:0.6),(Perpendicular to intended:0.2 each)}

-5		**+10**
S		**-5**

1. Draw optimal policy for the given grid.
2. Write algorithm of first visit MC, every visit MC.
3. The agent starts with the policy that always chooses policy to go right first and executes these episodes:
   a. (1,1) → (1,2) →(1,3)
   b. (1,1) → (1,2) → (2,2) → (3,2) → (3,3)
   c. (1,1) → (1,2) → (2,2) → (2,3) → (3,3)
   d. (1,1) → (2,1) → (3,1)
   e. (1,1) → (2,1) → (2,2) → (2,3) → (3,3)
   f. (1,1) → (2,1) → ((2,2) → (3,2) → (3,3)
   What are the Monte Carlo (direct utility and every visit) estimates for (1,1), (2,2) and (1,2)
4. Using learning rate of 1, discount factor as 0.9 and initial value of 0, what updates does TD learning agent make after first three episodes(a,b,c) of above?