1312N964

Candidate's Seat No :_____

MSc AIML & AIML (DS) Sem.-3 Examination

Reinforcement Learning

December-2024

Time : 3-00 Hours] **[Max. Marks : 100**

Instructions:

- **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 Consider the Markov chain with three states, S = {A, B, C}, that has the following transition matrix: [10]

	A	B	C
A	1/2	1/4	1/4
B	1/3	0	2/3
C	1/2	1/2	0

a. Draw the state transition diagram for this chain.
b. If we know P(X□=A) = P(X□=B) = 1/4, find P(X□=C, X□=B, X□=A).
(Here in (X□=S), n represents time step and S represents the state).

Q.2 A petrol station owner is considering the effect on his business (HP) of a new petrol station (JIO-BP) which has just opened down the road. Currently HP has 80% of the market share and JIO-BP has 20% of the market share. Analysis over the last week has indicated the following probabilities for customers switching the station they stop at each week: [10]

From	**HP**	**JIO-BP**
To HP	0.75	0.55
To JIO-BP	0.25	0.45

a. What will be the expected market share of HP and JIO-BP after another two weeks have passed?
b. What would be the long-run prediction for expected market share for HP and JIO-BP?

Q.3 (Attempt any three of the following) [30]

a. Derive the incremental update rule for estimating value of an action for a stationary problem. Also write the equation to estimate the value of a non-stationary problem.
b. Explain three different algorithms used to select an action (arm) in multi-armed bandit methods.
c. Explain the policy gradient approach used to update the parameters of a policy.
d. Name and explain the elements of Reinforcement Learning.

(P.T.O)

SECTION - II

Q.4 Answer the following (Any Four): [20]

a. Which function is referred to when we discuss about function approximation in RL? Mention all the advantages of using function approximators, and finally state the parameter update rule using gradient descent for linear functions.

b. Explain how 'Genetic Algorithm' optimization method is used to find the optimal policy in a reinforcement learning problem.

c. Explain the advantages of Temporal Difference learning over Monte-Carlo method. Write the value update equation for a 2-step Temporal Difference backup.

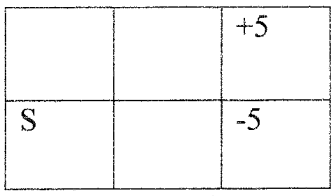
d. Explain how generalized policy iteration is used to find an optimal policy.

e. Explain how Actor-Critic method is used to find an optimal policy.

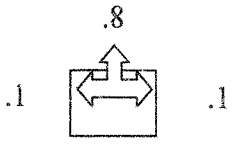
Q.5 *(Read the instructions and attempt the following grid world problem)* [30]

The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (North, South, West, or East) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.

- **S = { (1,1), (1,2), (1,3), (2,1), (2,2), (2,3) }**
- **P((s,a),s') = {(intended movement: 0.8), (perpendicular movement: 0.1 each)}**



(a)



(b)

a. Draw the optimal policy for this grid.

b. Suppose the agent knows the transition probabilities. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. (Assume V_0 is 0 everywhere and compute V_t for times t = 1, 2.)

c. Suppose the agent does not know the transition probabilities. What does it need to be able to do (or have available) in order to learn the optimal policy?

d. The agent starts with the policy that always chooses to go right and executes the following three trials:

1. (1,1)→(1,2)→(1,3)
2. (1,1)→(2,1)→(2,2)→(2,3)
3. (1,1)→(2,1)→(2,2)→(2,3)

What are the Monte Carlo (direct utility) estimates for states (1,1) and (2,2), given these traces?

e. Using a learning rate of 1 and assuming initial values of 0, what updates does the TD-learning agent make after trials 1 and 2 above?