OEOD Quiz Preparation (Questions and Answers)

MCD – Iscte, Diana Mendes – 2023

---

1. What is Reinforcement Learning (RL)?

A. An unsupervised learning approach

B. A form of unsupervised learning

C. Learning from labeled data

D. A machine learning training method based on rewarding desired behaviors and/or punishing undesired ones

2. In RL, what represents the learning agent's environment?

A. The learner and the decision-maker

B. The data used for training

C. The model architecture

D. The set of actions available

3. What is the objective of reinforcement learning?

A. To minimize rewards

B. To maximize the loss function

C. To minimize the policy

D. To train an agent to complete a task within an uncertain environment

4. What is the action-value function in RL?

A. The probability of taking an action

B. The immediate reward of an action

C. The future reward of an action

D. The probability of exploring an action

5. What does the "discount factor" in RL determine?

A. The learning rate

B. The agent's exploration rate

C. The value of the reward signal over time

D. The agent's decision-making speed

6. What is the term for the method in which an RL agent explores the environment to learn optimal actions?

A. Exploitation

B. Generalization

C. Exploration

D. Policy Optimization

7. In RL, what is a policy?

A. A set of states

B. A sequence of actions

C. A mapping of states to actions

D. A series of rewards

8. What is the exploration-exploitation trade-off in RL?

A. Balancing the model complexity

B. Balancing the learning rate

C. Balancing immediate and future rewards

D. Balancing between exploring and exploiting

9. What sets Reinforcement Learning apart from other machine learning paradigms?

A. Pre-trained models

B. Supervised labeling

C. Interaction with an environment

D. Batch processing

10. What term describes the strategy of choosing actions to maximize cumulative rewards over time?

A. Hyperparameter tuning

B. Feature extraction

C. Reinforcement learning

D. Policy optimization

11. What does the term “agent” refer to in Reinforcement Learning?

A. A person supervising the learning process

B. A software program making decisions

C. A labeled data point

D. A neural network architecture

12. What numerical values are used to evaluate the outcomes of actions taken by an agent?

A. Observations

B. Rewards

C. Policies

D. Loss functions

13. The “reward function” in RL is used for:

A. Defining the neural network architecture

B. Calculating the probability of actions

C. Evaluating the quality of an agent’s actions

D. Filtering noisy observations

14. Which algorithm is particularly well-suited for environments with continuous action spaces?

A. Q-Learning

B. Deep Q-Network (DQN)

C. Policy Gradient

D. Monte Carlo Tree Search (MCTS)

15. Which real-world applications benefit from Reinforcement Learning?

A. Image classification

B. Text generation

C. Autonomous driving

D. Data clustering

16. What is a sequential decision problem?

A: the agent has to make a sequence of decisions in order to solve a problem

17. What is the Markov property?

A: the next state depends only on the current state and the actions available in it (no influence of historical memory of previous states)

18. What is a policy pi(a|s)?

A: a conditional probability distribution for each possible state specifies the probability of each possible action.

19. What's on-policy?

A: The learning takes place by consistently backing up the value of the selected action back to the same policy function that was used to determine the action is on-policy

20. What is Grid world?

A: Grid worlds are the simplest environments; they consist of a rectangular grid of squares, with a start square and a goal square.

21. What is $Q(s, a)$?

A: Q-value function, the estimated value of taking action a in state s

22. Which model or function is meant when we say "model-free" or "model-based"?

A: When we say "model-free" we refer to the absence of the environments model, such as the transition function.

23. What type of action space and environment suit value-based methods?

A: Discrete action space and discrete/continuous state space/environment.

23 Reinforcement learning is a _____

A. Prediction-based learning technique
B. Feedback-based learning technique
C. History results-based learning technique

24. Which kind of data does reinforcement learning use?

A. Labeled data
B. Unlabelled data
C. None
D. Both

25. Reinforcement learning methods learned through ____?

A. Experience
B. Predictions
C. Analyzing the data

26. Which of the following is the practical example of reinforcement learning?

A. House pricing prediction
B. Market basket analysis
C. Text classification
D. Driverless cars

27. What is an agent in reinforcement learning?

A. An agent is a situation in which rewards are being exchanged
B. An agent is a simple value in reinforcement learning.
C. An agent is an entity that explores the environment.

28. What are actions in reinforcement learning?

A. Actions are the moves that the agent takes inside the environment.
B. Actions are the functions that the environment takes.
C. Actions are the feedback that an agent provides.

29. In which of the following reinforcement learning approaches do we find the optimal value function?

A. Value-based
B. Policy-based
C. Model-based

30. The agent's main objective is to ____the total number of rewards for good actions?

A. Minimize
B. Maximize
C. Null

31. Reinforcement learning is defined by the ____?

A. Policy
B. Reward Signal
C. Value Function
D. Model of the environment

32. Which element in reinforcement learning defines the behavior of the agent?

A. Policy
B. Reward Signal
C. Value Function
D. Model of the environment

33. Can reward signals change the policy?

A. Yes
B. No

34. Which of the following elements of reinforcement learning imitates the behavior of the environment?

A. Policy
B. Reward Signal
C. Value Function

D. Model of the environment

35. The approach in which reinforcement learning problems are solved with the help of models is known as ____?

A. Model-based approach

B. Model-free approach

C. Model known approach

36. Gamma (γ) in the Bellman equation is known as?

A. Value factor

B. Discount factor

C. Environment factor

37. How do you represent the agent state in reinforcement learning?

A. Discount state

B. Discount factor

C. Markov state

38. Consider the following condition: P[S(t+1) | S(t)] = P[S(t +1) | S(1),..., S(t)]. What is the meaning of S(t)?

A. State factor

B. Discount factor

C. Markov state

39. What do you mean by MDP in reinforcement learning?

A. Markov discount procedure

B. Markov discount process

C. Markov deciding procedure

D. Markov decision process

40. Why do we use MDP in reinforcement learning?

A. We use MDP to formalize the reinforcement learning problems.

B. We use MDP to predict reinforcement learning problems.

C. We use MDP to analyze the reinforcement learning problems.

41. What do you mean by SARSA in reinforcement learning?

A. State action reward state action

B. State achievement rewards state action

C. State act reward achievement

D. State act reward act

42. Which of the following policy types is a learning algorithm that evaluates and improves a policy dissimilar from the policy used for action selection?

A. behavior policy

B. Target policy

C. On-policy

D. Off-policy

43. Q-learning is a model-free or model-based learning algorithm?

A. Model-free

B. Model-based

44. The matrix created during the Q-learning algorithm is commonly known as ____?

A. Query-table

B. Q-table

C. Quick-matrix

D. Table

45. Does reinforcement learning provide any previous training?

A. Yes

B. No

46. Reinforcement Learning is about how computer systems can learn to _____.

A: Reinforcement Learning is about how computer systems can learn to take actions in an environment to maximize their rewards

47. The Markov Assumption assumes that to predict the future, you'll use _____.

A: The Markov Assumption assumes that in order to predict the future, you will use only the present state's information (the system is "memory-less" or that the future states of some process are only dependent upon the present state and not any of the states that preceded that).

48. What is the difference between state and history?

A: A state is generally a function of history. There could be other information that an agent would like to use, but we only consider observations seen so far, the actions taken, and the rewards gained.

49. How is a policy evaluated?

A: We can do this with the Value function, the expected discount sum of future rewards under a particular policy pi. It can be used to quantify the goodness or badness of states and actions and allows the agent to decide how to act by letting us compare policies

50. What is a transition matrix for Markov Process?

A: A transition matrix is a square matrix where the rows are usually non-negative real numbers that sum to 1. This matrix describes the transitions of a Markov Chain, where each corresponding element in the ith row and jth column represents the probability of moving from state i to state j in one-time step.

51. What is a Bandit?

A. Nickname for slot-machine

B. Nickname for Black-Jack

C. Losing reward

52. Which is the difference between Bellman (Value) Equation and Bellman (Value) Optimality Equation?

A: The Bellman equation defines the relationships between a given state (or state-action pair) and its successors. Bellman Optimality Equation is the maximum between all value functions.

53. What is policy iteration?

A: In policy iteration algorithms, you start with a random policy, then find the value function of that policy (policy evaluation step), then find a new (improved) policy based on the previous value function, and so on. In this process, each policy is guaranteed to be a strict improvement over the previous one (unless it is already optimal). Given a policy, its value function can be obtained using the Bellman equation.

54. What is value iteration?

A: In value iteration, you start with a random value function and then find a new (improved) value function in an iterative process, until reaching the optimal value function. Notice that you can easily derive the optimal policy from the optimal value function. This process is based on the optimality Bellman equation. The optimality Bellman equation contains a max-operator, which is non-linear and, therefore, has different features.

55. What does P(s' | s,a) stand for?

A: This is the Transition Probability. It indicates the probability of reaching state s' from state s given that you take the action a.

56. How do we calculate the Q(s,a) of a given policy pi?

- State-action value of a policy

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi}(s')$$

- Take action a, then follow the policy π

57. How do we calculate the Total Discounted Reward?

$$total\ discounted\ reward = \sum_{i=1}^{T} \gamma^{i-1} r_i$$

58. What is the difference between *Off-line* vs *On-line* learning?

A: These are two fundamental methods for solving MDPs. Both value-iteration and policy-iteration assume that the agent knows the MDP model of the world (i.e. the agent knows the state-transition and reward probability functions). Therefore, they can be used by the agent to (offline) plan its actions given knowledge about the environment before interacting with it. In Q-learning the agent improves its behavior (online) through learning from the history of interactions with the environment.

59. In multiarm bandits, what is regret, and why do we use it?

A: In MABs, there are multiple solutions to a problem, and usually, people measure regret to rank each solution. (Regret == simply put, the amount of penalty that we get for not pulling the optimal arm.). So, to minimize regret, we have to pull the arm with the highest probability of giving us a reward.

60. How do we calculate Reward for MABs?

A: Here, we are just counting how much reward we are getting for each arm, and dividing by the number of times we pulled each arm (hence calculating the percentage of getting a reward directly.)

61. Monte Carlo policy evaluation is a good choice when we know _____ about the dynamics and/or reward model.

A: Monte Carlo policy evaluation is a good choice when we know *nothing* about the dynamics and/or reward model.

62. Monte Carlo assumes _____ {one-shot, episodic, infinite} MDP.

A: Monte Carlo assumes *episodic* MDP.

63. An MDP given a fixed policy is a Markov chain with rewards.

A: True

64. If we know the optimal Q values, we can get the optimal V values only if we know the environment's transition function/matrix.

A: False, you can convert from Q-values to V-values without knowing the transition matrix

65. A policy that is greedy–with respect to the optimal value function–is not necessarily an optimal policy.

A: False. A policy "which is greedy with respect to the optimal value function" is by definition, optimal.

66. Why is the discount factor usually smaller than 1?

A: 1) Mathematical convenience (will converge)
2) Avoids infinite loops
3) Might be the truth for some cases, like financial rewards
4) Models uncertainty of future rewards

67. What is the idea of the Belman equation

A: Rewards can be decomposed into immediate and future rewards.

68. Define the Return

A: R(t) = r(t+1) + gamma*$r(t+2) + gamma^2$ r(t+3)+….

69. Define the state-value function

A: V(s) = E[R(t) |S(t) = s], R is the return

70. What is the direct solution of the Belman equation for MRP's using matrix/vector form?

A: V= ((I – gamma*P(ss'))^(-1)) * R

71. What happens when gamma, the discount factor, is close to 0 and 1

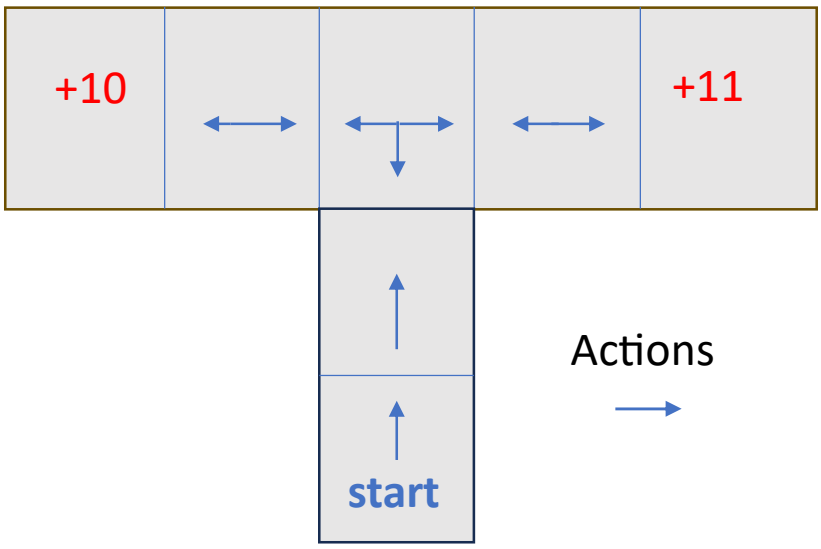
A: Close to 0 leads to a “short-sighted” evaluation (only consider current rewards), while close to 1 leads to a “far-sighted” evaluation (the return is not discounted)

72. What is a stochastic environment?

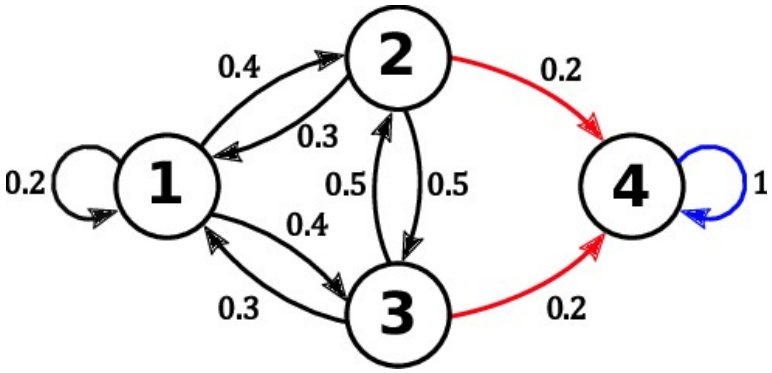
A: a non-deterministic environment, where the outcome of an action depends on elements in the environment, that are not known to the agent

73. Calculate the value function V* from Bellman equation for the maze shown in the image below.

A simple T-maze with 7 states. The start location is at the bottom, the possible action of each state are indicated as arrows. When arriving at the left or right end the agent obtains a reward of 10 or 11 reward units, respectively and gets ’transfered’ to the starting state. Consider each action as equally likely when writing down the Bellman equation and use gamma= 0:1 for your numerical solution.



74. Consider the following transition graph associated with a Markov chain. Write down the probability transition matrix P.

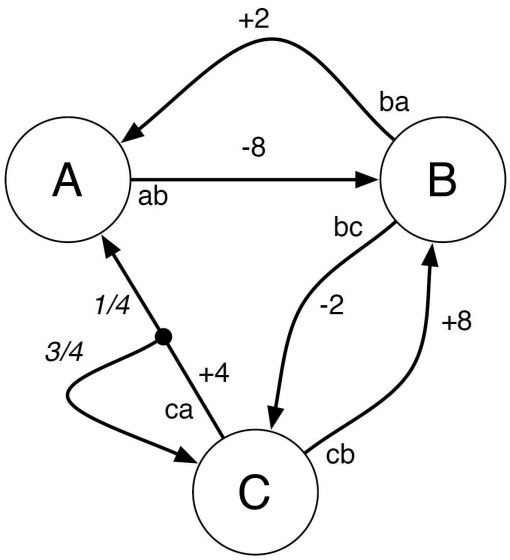


75. Consider the following Markov Decision Process (MDP) with discount factor *gamma = 0.5*. Upper case letters A, B, C represent states; arcs represent state transitions; lower case letters ab; ba; bc; ca; cb represent actions; signed integers fractions represent transition probabilities.

A. Define the state-value function V(s) for the MDP

B. Write the Bellman optimal equation for state-value functions

C. Starting with an initial value function of V1(A) = V1(B) = V1(C) = 2, apply one iteration of value iteration (i.e. one backup for each state) to compute a new value function V2(s).



76. Consider the following transition graph with Rewards:

A. Compute the Return for the sequence 1-1-2-3-Exit, for gamma=0.8

B. Compute the state-value function at state 2 for the previous example.