DMU 2024 Online Quizzes

March 10, 2025

Online Quiz 1

Question 1. Consider an MDP with integer states ($\mathcal{S} = \mathbb{Z}$) and $\mathcal{A} = \{1, 2\}$. State 3 is terminal and the discount factor is $\gamma = 0.9$. Suppose that you are performing reinforcement learning, and you observe an episode that takes the following trajectory:
$(s = 1, a = 2, r = 10, s' = 2)$
$(s = 2, a = 2, r = 5, s' = 1)$
$(s = 1, a = 1, r = 0, s' = 2)$
$(s = 2, a = 2, r = 0, s' = 3)$

1. Suppose that you are using the SARSA algorithm starting with all Q values initialized to zero before the episode. If the learning rate is $\alpha = 0.1$, what are the Q-value estimates for states 1 and 2 after the episode (including an update based on the final step)?
    - Q(1,1)
    - Q(1,2)
    - Q(2,1)
    - Q(2,2)
2. Suppose that you are using maximum likelihood tabular model-based reinforcement learning (MLMBTRL). After the trajectory above, what are the maximum likelihood reward estimates for states 1 and 2 after the episode?
    - R(1,1)
    - R(1,2)
    - R(2,1)
    - R(2,2)

Solution:

1. - Q(1,1) = 0.045

- Q(1,2) = 1.0
- Q(2,1) = 0.0
- Q(2,2) = 0.45

2.
- R(1,1) = 0.0
- R(1,2) = 10.0
- R(2,1) = 0.0
- R(2,2) = 2.5

Question 2. Consider the following neural-network-like function approximators where W's are weight matrices and b's are bias vectors and σ is the sigmoid function. Which of these structures is the most powerful in the sense that it can approximate the most general class of functions?

(A) $f(x) = W_3(W_2(W_1x^3 + b_1) + \sigma(b_2))$

(B) $f(x) = W_3(W_2(W_1x^2 + b_1) + \sigma(b_2))$

(C) $f(x) = W_3\sigma(W_2\sigma(W_1x^3 + b_1) + b_2)$

(D) $f(x) = W_3\sigma(W_2\sigma(W_1x^2 + b_1) + b_2)$

Briefly justify your answer.

Solution: (A) and (B) are only linear functions of x^2 and x^3 respectively. For (C), since x is squared, the output for any x and $-x$ must be the same. (D) is the most general and can approximate any function.

Online Quiz 2

Question 3. Suppose that I want to find an optimal policy for a Markov decision process with integer states and actions ($S = A = \mathbb{Z}$), discount factor $\gamma = 0.9$ and reward function $R(s, a) = s^2$.

I am considering training under the following modified reward functions. Which reward function would **NOT** be guaranteed to yield the same optimal policy?

(A) $R(s, a, s') = 2\,s^2$

(B) $R(s, a, s') = 2\,s^2 - \gamma s'^2$

(C) $R(s, a, s') = s^2 + \sin(a)$

(D) $R(s, a, s') = s^2 + \gamma \sin(s') - \sin(s)$

Briefly justify your answer

Solution: The optimal policy is independent of reward scaling and potential-based reward shaping. All of the answers except (C) involve only scaling and potential-based reward shaping. (C) introduces a non-linear action term, which could change the optimal policy.

Question 4. Double Q learning is meant to address which of the following problems?

(A) Insufficient exploration

(B) Maximization bias

(C) Incorrect credit assignment

(D) Poor sample efficiency

Solution: Maximization Bias

Question 5. Which of the following is **NOT** an advantage of entropy regularization, as used in the Soft Actor Critic algorithm

(A) It usually learns a combination of many near-optimal policies, making the resulting solution more robust to modeling errors

(B) Given unlimited time and computational resources, it is possible to learn better policies with entropy regularization than without

(C) The randomness incentivized by the entropy term leads to natural exploration

(D) All of these are advantages

Solution: (B) is not an advantage of entropy regularization. Given unlimited time and computational resources, it is possible to learn the optimal policy with or without entropy regularization.

Question 6. In DQN, the Q network typically uses what inputs and outputs?

(A) Input: History of states and actions; Output: The maximum Q-value

(B) Input: A state and action; Output: A scalar Q-value for the state and action

(C) Input: A state; Output: The maximum Q value and the action that maximizes it

(D) Input: A state; Output: Q values for all actions at the input state

Solution: (D)

Question 7. Which of the below statements is the best description of the natural gradient?

(A) The policy parameter adjustment expected to maximize rewards subject to a constraint on the KL divergence of trajectories under the new policy with respect to the trajectories under the old policy

(B) The gradient of the value (expected sum of discounted rewards) with respect to the policy parameters

(C) The policy parameter adjustment that minimizes the KL divergence of the updated policy with respect to the original policy

(D) The gradient that will maximize the KL divergence of the rewards attained by the new policy with respect to the rewards obtained with the original policy

Solution:‘ (A)

Online Quiz 3

Question 8. Suppose that two roommates need to decide who will take out the trash. They decide to flip a fair coin and the loser has to take out the trash.

1. What type of equilibrium best describes this arrangement?
   (A) Dominant strategy equilibrium
   (B) Pure Nash equilibrium
   (C) Mixed Nash equilibrium
   (D) Correlated Equilibrium
2. Choose one of the other types of equilibrium listed above and explain why the coin-flip arrangement is preferable to that type of equilibrium for this game.

Solution: (D) Correlated Equilibrium.
Assuming the payoffs are such that both players would rather take out the trash than leaving it, there is a pure Nash equilibrium where one player always takes out the trash. This is unfair to the person who takes out the trash. The coin-flip arrangement is preferable because it is fair.

Question 9. Which of the following is a step in the fictitious play algorithm applied to a Markov Game?

(A) Create a POMDP model where the unknown state variable includes the other players' possible strategies and solve the POMDP to find the optimal policy.

(B) Create an MDP based on the frequency of the other players' previous plays and solve this MDP to find a best response strategy.

(C) Perform a policy-gradient step to improve the policy of the current player based on the state trajectory from the previous episode.

(D) Solve a linear program to find a correlated equilibrium that is the best outcome for all players.

Solution: (B)

Question 10. In a two-player partially observable Markov game, a policy π^i can be pruned (i.e. ignored in a Nash equilibrium computation) if

(A) Another strategy outperforms π^i at some (s, π^{-i}) pair.

(B) At least one strategy outperforms π^i at each (s, π^{-i}) pair.

(C) There is no *distribution* over (s, π^{-i}) pairs, $b(s, \pi^{-i})$, where π^i outperforms all other strategies.

(D) It is impossible to prune any strategies in a POMG because all pure strategies could potentially be part of a mixed Nash equilibrium.

(In the context of this question, π^i "outperforms" $\pi^{i'}$ at an (s, π^{-i}) pair if $U^{\pi^i, \pi^{-i}, i}(s) > U^{\pi^{i'}, \pi^{-i}, i}(s)$)

Solution: (C)

Question 11. Which of the following is a valid reason that inverse reinforcement learning has more potential for generalization than behavioral cloning in some settings?

(A) Inverse reinforcement learning requires less information about the MDP compared to behavioral cloning.

(B) Behavioral cloning can only learn to imitate deterministic policies, while inverse reinforcement learning can also learn probabilistic policies.

(C) Since inverse reinforcement learning attempts to learn the expert's reward function, it may yield expert-like policies in other environments where the reward function is the same but the dynamics differ from the environment where the data was collected.

Solution: (C)