Graded Quiz

测验, 13 个问题

13/13 分 (100%)

✓ **恭喜！您通过了！**

下一项

✓ 1 / 1
分数

1.
Which approach ensures continual exploration? (Select all that apply)

□ Exploring starts

正确
Correct! Exploring starts guarantee that all state-action pairs are visited an infinite number of times in the limit of an infinite number of episodes.

□ On-policy learning with a deterministic policy

未选择的是正确的

□ On-policy learning with an ϵ-soft policy

正确
Correct! ϵ-soft policies assign non-zero probabilities to all state-action pairs.

□ Off-Policy learning with an ϵ-soft behavior policy and a deterministic target policy

正确
Correct! ϵ-soft policies have non-zero probabilities for all actions in all states. The behavior policy is used to generate samples and should be exploratory.

□ Off-Policy learning with an ϵ-soft target policy and a deterministic behavior policy

未选择的是正确的

✓ 1 / 1 分数

2.
When can Monte Carlo methods, as defined in the course, be applied? (Select all that apply)

□ When the problem is continuing and there are sequences of states, actions, and rewards

未选择的是正确的

□ When the problem is continuing and there is a model that produces samples of the next state and reward

未选择的是正确的

□ When the problem is episodic and there are sequences of states, actions, and rewards

正确
Correct! Well-defined returns are available in episodic tasks.

□ When the problem is episodic and there is a model that produces samples of the next state and reward

正确
Correct! Well-defined returns are available in episodic tasks.

✓ 1 / 1 分数

3.
Which of the following learning settings are examples of off-policy learning? (Select all that apply)

□ Learning about multiple policies simultaneously while following a single behavior policy

正确
Correct! Off-policy learning enables learning about multiple target policies simultaneously using a single behavior policy.

□ Learning the optimal policy while continuing to explore

正确
Correct! An off-policy method with an exploratory behavior policy can assure continual exploration.

☑ Learning from data generated by a human expert

正确

Correct! Applications of off-policy learning include learning from data generated by a non-learning agent or human expert. The policy that is being learned (the target policy) can be different from the human expert's policy (the behavior policy).

✓ 1 / 1

分数

4.

If a trajectory starts at time t and ends at time T, what is its relative probability under the target policy π and the behavior policy b?

● $\prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$

正确

Correct! This is the importance sampling ratio and is used to weight returns in off-policy Monte-Carlo Policy Evaluation.

○ $\sum_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$

○ $\frac{\pi(A_{T-1} \mid S_{T-1})}{b(A_{T-1} \mid S_{T-1})}$

○ $\frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$

✓ 1 / 1

分数

5.

When is it possible to determine a policy that is greedy with respect to the value functions v_π, q_π for the policy π? (Select all that apply)

☑ When state values v_π and a model are available

正确

Correct! With state values and a model, one can look ahead one step and see which action leads to the best combination of reward and next state.

□ When state values v_π are available but no model is available.

未选择的是正确的

□ When action values q_π and a model are available

正确
Correct! Action values are sufficient for choosing the best action in each state.

□ When action values q_π are available but no model is available.

正确
Correct! Action values are sufficient for choosing the best action in each state.

✓ 1 / 1
分数

6.
Monte Carlo methods in Reinforcement Learning work by...

○ Averaging sample rewards

○ Planning with a model of the environment

○ Performing sweeps through the state set

● Averaging sample returns

正确
Correct! Monte Carlo methods in Reinforcement Learning sample and average returns much like bandit methods sample and average rewards.

✓ 1 / 1
分数

7.
Which of the following is a requirement for using Monte Carlo policy evaluation with a behavior policy b for a target policy π?

○ For each state s and action a, if $b(a \mid s) > 0$ then $\pi(a \mid s) > 0$

● For each state s and action a, if $\pi(a \mid s) > 0$ then $b(a \mid s) > 0$

正确

Correct! Every action taken under π must have a non-zero probability under b.

○ All actions have non-zero probabilities under π

✓ 1 / 1
分数

8.

Suppose the state s has been visited three times, with corresponding returns 8, 4, and 3. What is the current Monte Carlo estimate for the value of s?

○ 3

○ 15

[✓ ○ 5]

正确

Correct! The Monte Carlo estimate for the state value is the average of sample returns observed from that state.

○ 3.5

✓ 1 / 1
分数

9.

When does Monte Carlo prediction perform its first update?

○ After the first time step

○ When every state is visited at least once

[✓ ○ At the end of the first episode]

正确

Correct! Monte Carlo Prediction updates value estimates at the end of an episode.

✓ 1 / 1
分数

10.
In Monte Carlo prediction of state-values, **memory** requirements depend on (select all that apply)

□ The number of states

正确
Correct! Monte Carlo Prediction needs to store the estimated value for each state.

□ The number of possible actions in each state

未选择的是正确的

□ The length of episodes

正确
Correct! Monte Carlo Prediction needs to store the sequence of states and rewards. during an episode

---

✓ 1 / 1
分数

11.
For Monte Carlo Prediction of state-values, the number of **updates** at the end of an episode depends on

○ The number of states

○ The number of possible actions in each state

○ The length of the episode

正确
Correct! Monte Carlo Prediction updates the estimated value of each state visited during the episode.

✓ 1 / 1
分数

12.
Which approach can find an optimal deterministic policy? (select all that apply)

□ Exploring Starts

正确

Correct! Exploring starts ensure that every state-action pair is visited even if the policy is deterministic.

□ ϵ-greedy exploration

未选择的是正确的

□ Off-policy learning with an ϵ-soft behavior policy and a deterministic target policy

正确

Correct! In this case, the behavior policy can maintain exploration while the target policy is deterministic.

✓ 1 / 1
分数

13.

In an ϵ-greedy policy over $\mathcal{A}$ actions, what is the probability of the highest valued action if there are no other actions with the same value?

○ $1 - \epsilon$

○ ϵ

[✓ ○ $1 - \epsilon + \frac{\epsilon}{\mathcal{A}}$]

正确

Correct! The highest valued action still has a chance of being selected as an exploratory action.

○ $\frac{\epsilon}{\mathcal{A}}$