Reinforcement Learning 시험지

과목명: Reinforcement Learning | 문항수: 20 | 시험일: 2025-09-27

1. [MDP] MDP의 구성요소로 옳지 않은 것은 무엇인가?

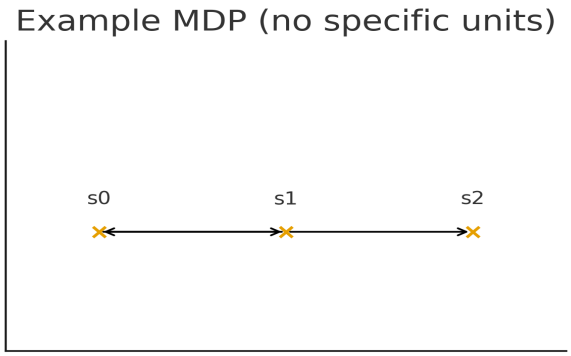

그림 1. 예시 MDP 다이어그램 (참고용)

A. 상태 집합 S

B. 행동 집합 A

C. 전이확률 P

D. 학습률 α

2. [Value/Action Value] 상태가치함수 V^π(s)와 행동가치함수 Q^π(s,a)의 차이에 대한 설명으로 옳은 것은?

A. V^π는 상태만의 장기보상을, Q^π는 상태-행동 쌍의 장기보상을 나타낸다

B. 둘 다 상태만의 보상을 나타낸다

C. 둘 다 행동만의 보상을 나타낸다

D. Q^π는 정책 π와 무관하다

3. [Bellman Equations] 벨만 최적 방정식에 대한 설명으로 옳은 것은?

A. 현재 가치가 미래 가치에 의존하지 않는다

B. 최적가치는 한 단계 보상과 다음 상태의 최적가치의 합의 기댓값으로 표현된다

C. 감가율 γ는 항상 1로 고정되어야 한다

D. 정책이 랜덤이면 최적 방정식은 정의되지 않는다

4. [DP/MC/TD] DP, MC, TD 방법의 비교로 옳은 것은?

A. MC는 모델(전이확률)을 반드시 필요로 한다

B. TD(0)는 부트스트래핑을 사용한다

C. DP는 모델이 없어도 작동한다

D. MC는 부분 에피소드로도 편향 없이 추정한다

5. [Q-learning] Q-learning의 학습 특성으로 옳은 것은?

A. 온폴리시 알고리즘이다

B. 오프폴리시 알고리즘이다

C. 정책 평가만 수행한다

D. 탐험 없이도 항상 수렴한다

6. [SARSA] SARSA에 대한 설명으로 옳은 것은?

A. 오프폴리시로 동작한다

B. 탐험-활용 정책 그대로를 평가한다

C. 다음 상태에서의 최대 Q 값을 사용한다

D. 정책과 무관하게 목표를 평가한다

7. [Exploration (ε-greedy)] ε-탐욕(ε-greedy) 정책에 대한 설명으로 옳지 않은 것은?

A. 확률 ε로 무작위 행동을 고른다

B. 확률 1-ε로 현재 최선의 행동을 고른다

C. ε가 0이면 항상 무작위 선택이다

D. ε 스케줄은 점감시킬 수 있다

8. [Policy Gradient/Baselines] 정책 경사(Policy Gradient)에서 베이스라인(baseline)을 사용하는 주된 이유는?

A. 편향을 증가시켜 수렴을 빠르게 한다

B. 분산을 감소시켜 안정적인 학습을 돕는다

C. 학습률을 자동 조절한다

D. 목표함수를 변경한다

9. [DP/MC/TD] TD 타깃은 다음 중 무엇인가?

A. 샘플 리턴 G_t

B. r_{t+1} + γ V(s_{t+1})

C. r_{t} + V(s_{t})

D. G_{t+1} - r_{t+1}

10. [Bellman Equations] 다음 중 최적 정책 π*와 Q*의 관계로 옳은 것은?

A. π*(s)=argmin_a Q*(s,a)

B. π*(s)=argmax_a Q*(s,a)

C. π*(s)은 Q*와 무관하다

D. Q*는 π*가 정해진 뒤 계산된다

11. [Value/Action Value] 상태가치 V와 행동가치 Q의 관계로 옳은 것은?

A. V^π(s)=max_a Q^π(s,a)

B. V^π(s)=E_{a∼π}[Q^π(s,a)]

C. $Q^{\pi}(s,a)=\min_s V^{\pi}(s)$

D. V와 Q는 서로 독립적이다

12. [Q-learning] Q-learning 업데이트에서 사용되는 타깃은?

A. $r + \gamma \max_{a'} Q(s', a')$

B. $r + \gamma Q(s', a)$

C. G_t 전체 리턴

D. $r - \gamma Q(s, a)$

13. [단답] [MDP] MDP의 정의와 구성요소(S, A, P, R, γ)를 간략히 설명하고, 왜 마르코프 성질이 중요한지 요약하시오. (100자 이상)

14. [단답] [DP/MC/TD] DP, MC, TD의 핵심 차이를 '모델 필요 여부'와 '부트스트래핑' 관점에서 비교 설명하시오. (예시 1개 포함)

15. [단답] [Bellman Equations] 벨만 기대 방정식과 벨만 최적 방정식의 차이를 수식/개념으로 정리하시오.

16. [단답] [Exploration (ε-greedy)] ε-탐욕 정책에서 ε 스케줄(예: 선형/지수 감소)을 사용하는 이유와 주의점을 설명하시오.

17. [단답] [Policy Gradient/Baselines] 정책 경사에서 Advantage의 역할과 baseline을 V(s)로 둘 때의 장점을 설명하시오.

18. [계산] [Returns] 감가율 $\gamma=0.9$일 때, 보상열 r0=2, r1=0, r2=3이 주어지면 $G0= r0 + \gamma\, r1 + \gamma^2\, r2$ 를 계산하시오. (소수점 1자리 반올림, 단위: pts)

반올림 규칙: 소수점 1자리. 단위를 반드시 표기(예: pts, value).

19. [계산] [TD(0)] TD(0) 업데이트: V(s) ← V(s) + α [r + γ V(s') - V(s)]. α=0.5, r=1.0, γ=0.9, V(s)=2.0, V(s')=3.0 일 때 새로운 V(s)를 구하시오. (소수점 1자리, 단위: value)

반올림 규칙: 소수점 1자리. 단위를 반드시 표기(예: pts, value).

20. [서술] [통합 비교] DP, MC, TD의 개념과 차이를 정의→비교→한계/응용 순으로 정리하고, 정책 반복과 가치 반복의 위치를 설명하시오. 또한 온폴리시/오프폴리시 관점에서 Q-learning과 SARSA를 비교하시오. 마지막으로 정책 경사(Actor-Critic 포함)의 장점과 베이스라인/어드밴티지의 역할을 서술하시오. (300-600단어)