강화학습 (Reinforcement Learning) 시험지

과목: Reinforcement Learning | 시험일: 2025-09-29 | 총 문항: 20

문항 1 [MDP] · 난이도: hard

다음 중 유한 MDP의 구성요소가 아닌 것은 무엇인가?

A. 상태 집합 S
B. 행동 집합 A
C. 전이 확률 P
D. 관찰 확률 O

문항 2 [Value/Action Value] · 난이도: easy

상태가치 함수 V와 행동가치 함수 Q의 관계로 옳은 것은?

A. V(s)=max_a Q(s,a)
B. Q(s,a)=min_s V(s)
C. V(s)=Σ_a Q(s,a)
D. Q(s,a)=γ·V(s)

문항 3 [Bellman Equations] · 난이도: medium

벨만 최적 방정식의 핵심 아이디어는?

A. 탐욕적 정책의 무작위성
B. 가치의 분해와 재귀적 정의
C. 모든 상태에서 동일 보상
D. 할인율을 0으로 고정

문항 4 [DP/MC/TD] · 난이도: easy

MC와 TD의 차이로 가장 적절한 설명은?

A. MC는 부트스트래핑, TD는 에피소드 전체 반환 사용
B. MC는 모델 필요, TD는 모델 불필요
C. MC는 에피소드 종료 후 갱신, TD는 단계별 부트스트래핑
D. 둘 다 항상 편향 없음

문항 5 [Q-learning] · 난이도: medium

오프폴리시 학습의 대표인 Q-learning에서 타깃으로 사용되는 것은?

A. 현재 정책에 따른 다음 행동의 가치
B. 무작위 행동의 평균 가치
C. max_a' Q(s',a')
D. V(s')

문항 6 [SARSA] · 난이도: medium

SARSA의 타깃은 다음 중 무엇인가?

A. r + γ max_a' Q(s',a')
B. r + γ Q(s',a') (a'는 실제 선택)
C. r + γ V(s')

D. r + Q(s,a)

문항 7 [Exploration (ε-greedy)] · 난이도: medium

ε-greedy 탐색에서 ε를 점진적으로 감소시키는 주된 이유는?

A. 탐색을 늘려 수렴을 방해
B. 초기 탐험 후 수렴을 위해 탐색 축소
C. 항상 최적 행동만 선택하기 위해
D. 보상을 0으로 만들기 위해

문항 8 [Policy Gradient/Baselines] · 난이도: hard

REINFORCE에서 baseline을 사용하는 주된 목적은?

A. 편향을 증가
B. 분산을 감소
C. 수렴 속도 저하
D. 학습률 제거

문항 9 [MDP] · 난이도: medium

그림 1을 참고하여, 간단한 3상태 MDP에서 화살표는 무엇을 나타내는가?

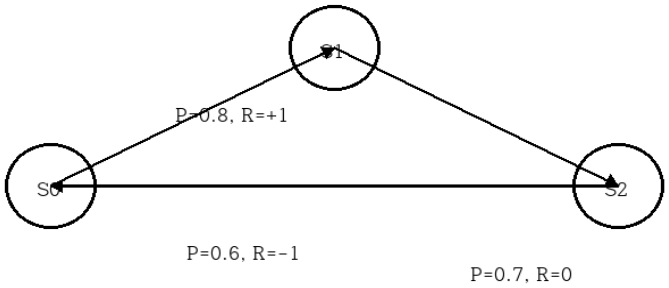

그림 1. 간단한 MDP 다이어그램 (재구성)

A. 정책 확률 π(a|s)
B. 전이 P(s'|s,a)
C. 즉시 보상 r
D. 가치함수 V(s)

문항 10 [Bellman Equations] · 난이도: medium

그림 2는 어떤 연산을 도식화한 것인가?

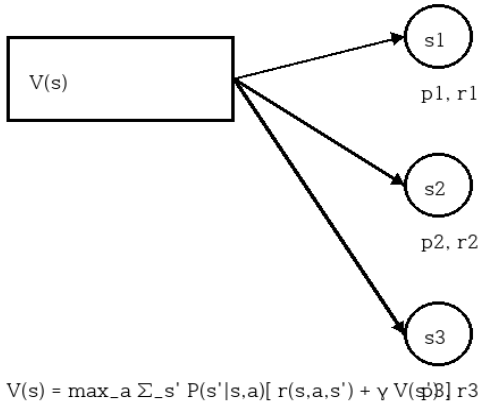

그림 2. 벨만 백업 도식 (재구성)

A. 폴리시 이벨류에이션의 잔차 계산
B. 벨만 백업(최적가치)

C. 모델 프리 샘플 평균
D. 정책 경사 추정

문항 11 [DP/MC/TD] · 난이도: hard

다이나믹 프로그래밍(DP)을 사용하려면 필요한 가정은?

A. 모델이 필요 없다
B. 전이 확률과 보상 모델을 알고 있어야 한다
C. 에피소드가 유한할 필요가 없다
D. 오프폴리시만 가능하다

문항 12 [Exploration (ε-greedy)] · 난이도: medium

ε-greedy에서 ε가 너무 크면 발생할 수 있는 문제는?

A. 탐색 부족
B. 조기 수렴
C. 과도한 무작위성으로 성능 저하
D. 학습률 폭증

문항 13 [Value/Action Value] · 난이도: easy

상태가치 함수 V(s)와 행동가치 함수 Q(s,a)의 차이를 정의하고, 왜 Q가 정책 개선에 직접적으로 유용한지 설명하시오.

단답/서술 답안 영역

문항 14 [DP/MC/TD] · 난이도: medium

Monte Carlo, Temporal-Difference, Dynamic Programming의 핵심 차이를 '모델 필요성'과 '부트스트래핑 여부' 관점에서
비교하시오.

단답/서술 답안 영역

문항 15 [Q-learning] · 난이도: easy

Q-learning이 오프폴리시인 이유를 한 문단으로 설명하고, SARSA와의 차이를 언급하시오.

단답/서술 답안 영역

문항 16 [Policy Gradient/Baselines] · 난이도: hard

Policy Gradient에서 baseline을 사용하는 이유와 대표적인 baseline 형태(상태가치, advantage)를 설명하시오.

단답/서술 답안 영역

문항 17 [Exploration (ε-greedy)] · 난이도: medium

ε-greedy에서 ε 스케줄링(예: 선형/지수 감쇠)의 목적과 주의점을 간단한 예시와 함께 설명하시오.

단답/서술 답안 영역

문항 18 [Bellman Equations] · 난이도: medium

할인율 γ=0.9인 그리디 정책 하에서, 상태 s에서 행동 a를 취했을 때 한 단계 벨만 최적 타깃은 r=2, 다음 상태 s'에서 가능한 Q값이 [5.0,
4.5, 3.0]라면 무엇인가? 반올림 규칙(소수 첫째 자리)과 단위를 확인하시오. (단위: value)

※ 반올림 규칙: 소수 첫째 자리. 단위 확인 필수.

수치 계산 답안 영역

문항 19 [Q-learning] · 난이도: easy

Q-learning 업데이트: 학습률 α=0.5, 할인율 γ=0.9, 관측 (s,a,r,s')에서 r=1.0, 현재 Q(s,a)=3.0, 그리고 max_a'
Q(s',a')=4.0일 때 새로운 Q(s,a)는 무엇인가? (소수 첫째 자리 반올림, 단위: value)

※ 반올림 규칙: 소수 첫째 자리. 단위 확인 필수.

수치 계산 답안 영역

문항 20 [Policy Gradient/Baselines] · 난이도: easy

REINFORCE와 Actor-Critic을 비교하여 정책경사 방법의 기본 원리, baseline/critic의 역할, 분산-편향 관점의 장단점, 그리고 실무
적용 시 고려해야 할 한계와 안정화 기법을 논하시오.

에세이 답안 영역 (최소 300단어)