Reinforcement Learning 해설/정답

※ 출처 표기는 자료 미제공으로 '개념설명' 모드로 대체합니다.

1. [MDP] (정답: D)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
B) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: $r + \gamma V(s')$ 또는 $r + \gamma \max_{a'} Q(s',a')$

출처: (개념설명)

2. [Value/Action Value] (정답: A)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
B) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: $r + \gamma V(s')$ 또는 $r + \gamma \max_{a'} Q(s',a')$

출처: (개념설명)

3. [Bellman Equations] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: $r + \gamma V(s')$ 또는 $r + \gamma \max_{a'} Q(s',a')$

출처: (개념설명)

4. [DP/MC/TD] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

5. [Q-learning] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

6. [SARSA] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

7. [Exploration (ε-greedy)] (정답: C)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
B) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

8. [Policy Gradient/Baselines] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

9. [DP/MC/TD] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

10. [Bellman Equations] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

11. [Value/Action Value] (정답: B)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
A) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

12. [Q-learning] (정답: A)

정답 근거: 정의와 수식에 따르면 선택지가 옳습니다. V, Q, 벨만 방정식, 온/오프폴리시 특성 등 개념을 바탕으로 판단합니다.

오답 분석:
B) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
C) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.
D) 해당 선택지는 정의/조건과 맞지 않습니다. 핵심 개념 또는 가정이 틀렸습니다.

참고식: 벨만 기대/최적 방정식, TD 타깃: r + γ V(s') 또는 r + γ max_a' Q(s',a')

출처: (개념설명)

13. [단답] [MDP]

모범답안(요약): 핵심정의, 예시/직관, 주의점을 포함하여 100자 이상으로 기술합니다. 예: MDP는 (S,A,P,R,γ)로 구성되며 마르코프 성질이 중요합니다. DP/MC/TD는 모델 필요 여부와 부트스트래핑 사용 여부에서 차이가 납니다.

출처: (개념설명)

14. [단답] [DP/MC/TD]

모범답안(요약): 핵심정의, 예시/직관, 주의점을 포함하여 100자 이상으로 기술합니다. 예: MDP는 (S,A,P,R,γ)로 구성되며 마르코프 성질이 중요합니다. DP/MC/TD는 모델 필요 여부와 부트스트래핑 사용 여부에서 차이가 납니다.

출처: (개념설명)

15. [단답] [Bellman Equations]

모범답안(요약): 핵심정의, 예시/직관, 주의점을 포함하여 100자 이상으로 기술합니다. 예: MDP는 (S,A,P,R,γ)로 구성되며 마르코프 성질이 중요합니다. DP/MC/TD는 모델 필요 여부와 부트스트래핑 사용 여부에서 차이가 납니다.

출처: (개념설명)

16. [단답] [Exploration (ε-greedy)]

모범답안(요약): 핵심정의, 예시/직관, 주의점을 포함하여 100자 이상으로 기술합니다. 예: MDP는 (S,A,P,R,γ)로 구성되며 마르코프 성질이 중요합니다. DP/MC/TD는 모델 필요 여부와 부트스트래핑 사용 여부에서 차이가 납니다.

출처: (개념설명)

17. [단답] [Policy Gradient/Baselines]

모범답안(요약): 핵심정의, 예시/직관, 주의점을 포함하여 100자 이상으로 기술합니다. 예: MDP는 (S,A,P,R,γ)로 구성되며 마르코프 성질이 중요합니다. DP/MC/TD는 모델 필요 여부와 부트스트래핑 사용 여부에서 차이가 납니다.

출처: (개념설명)

18. [계산] [Returns]

풀이 단계:
문제 해석: G0=r0+γr1+γ^2 r2.
공식/대입: 2 + 0.9*0 + 0.9^2*3.
계산: 2 + 0 + 0.81*3 = 2 + 2.43 = 4.43.
반올림/단위: 소수점 1자리 → 4.4 pts.

정답: 4.4 pts

출처: (개념설명)

19. [계산] [TD(0)]

풀이 단계:
문제 해석: TD(0) 업데이트 식 적용.
공식/대입: 2.0 + 0.5 * (1.0 + 0.9*3.0 - 2.0).
계산: 1.0 + 2.7 - 2.0 = 1.7; 0.5*1.7=0.85; 2.0+0.85=2.85.
반올림/단위: 소수점 1자리 → 2.9 value.

정답: 2.9 value

출처: (개념설명)

20. [서술] [통합 비교]

개요(Outline): 정의 → 비교 → 한계/응용 → 정책 반복·가치 반복 위치 → 온폴리시/오프폴리시 비교(Q-learning vs SARSA) → 정책 경사/Actor-Critic 장점 → 베이스라인/어드밴티지 역할. (300-600단어)

- 정의: DP(모델 필요), MC(에피소드 평균), TD(부트스트래핑).
- 비교: 데이터/계산 효율, 편향-분산 특성.
- 한계: MC의 분산, TD의 편향, DP의 모델 의존.
- Q-learning(오프폴리시) vs SARSA(온폴리시).
- 정책 경사: 연속/고차원 행동, 확률적 정책에 유리.
- Advantage와 baseline으로 분산 감소.

출처: (개념설명)