|  |  |
| --- | --- |
| 교육 제목 | **머신러닝 강화학습 MD** |
| 교육 일시 | 2021년 11월 29일 월요일 |
| 교육 장소 | 영우글로벌러닝 2층 |
| **교육 내용** | |
| 오전 | 1. Reinforcement Learning이 가지는 특징    1. There is no supervisor(감시자) : 누구도 답(행동)을 알려주지 않음 2. supervised learning과의 차이점    1. supervisor가 없으므로 optimal solution을 찾는 목적에 적합하다    2. Only a reward signal : 오직 reward(보상) 신호만이 존재    3. Reinforcement Learning은 reward를 maximize(최대화)하는 목적을 가진다    4. Feedback is delayed, not instaneous : Feedback이 즉각적이지 않고 지연될 수 있다.    5. Action(행동)이 당장의 결과가 아닌 미래의 결과에 영향을 줄 수 있다.    6. Time really matters: 시간이 중요하다!    7. Sequential data: 순서와 시간이 중요한 데이터    8. non i.i.d data: independent identically distributed(독립 동일 분포)가 아니다.    9. i.i.d: 동일한 확률분포로 부터 뽑힌 확률 변수 / 각각은 서로에게 독립적이다.    10. Agent’s actions affect the subsequent data it receives    11. Agent의 Action이 후의 데이터에 영향을 미친다. 3. Rt는 Step t에서 얼마나 Agent가 잘 행동하고있는지를 나타낸다. Agent의 목적은 Reward를 Maximize하는 것이다. 4. Reward는 반드시 Scalar값이어야 한다.(Vector인자는 학습이 불가하다) 5. 모든 Reinforcement Learning 은 Reward hypothesis에 근거한다 6. Reward hypothesis: 강화학습의 모든 목적은 축적되는 Reward를 극대화하는 것으로 표현될 수 있다. 7. Q. Do you agree with this statement?(동의하는가?)    1. Reward hypothesis는 항상 성립하는가?    2. 당연하게도 항상 성립하지 않는다! Reward hypothesis가 성립하는 문제가 강화학습을 적용하기 적합한 문제이다.    3. 복잡한 문제들은 Reward hypothesis가 적용되지 않는 경우가 더 많다. 이럴경우 작은 문제들로 나누어 강화학습을 적용해야 한다. 8. Vector Reward로 학습하고 싶다면 어떻게 할 수 있는가?    1. Vector를 분해하여 각각의 성분을 적절하게 더하여 Scalar로 표현하면 된다.    2. Ex. Reward\_Vector = (a,b) / Reward\_Scalar = 10a + 6b    3. 강화학습의 핵심   -   * 1. Agent 그림에서 뇌를 의미한다      1. Action(At)을 하는 주체      2. OtRt를 받아 At를 행동한다   2. Environment      1. 그림에서 외부의 환경(지구)을 의미한다      2. Agent에게 Observation(Ot)과 Reward(Rt)를 준다      3. At를 받아 Ot+1, Rt+1을 Agent에게 전달한다 |
| 오후 | 1. Value Function: 어떤 State로 이동하거나 Action을 할 때 앞으로 받게될 총 보상(Discounted Reward)의 기댓값 2. 좋은 State/Action 을 판단하는 지표가된다. 3. Discounted Reward: 하나의 Episode를 Sampling해보자.(Terminal State(종결)에 도달하기 까지 Policy를 따라 움직인 1개의 경로를 생각해보자) 이때 각 action마다의 보상이 존재할 것이다. 이러한 보상의 Sequence을 그대로 더할 수도 있겠지만 대부분의 RL에서는 현재의 보상이 미래의 보상보다 Valuable한 것으로 본다. 이를 위해 discount factor(γ)의 개념을 도입하였다. 4. discount factor(γ)을 적용하여 미래에 받을 보상보다 현재의 보상에 더욱 비중을 두었다. 5. step이 지날수록 받는 보상에 discount factor(γ)가 추가적으로 곱해진다(작아진다) γ∈[0,1) 6. 핵심!(오개념 주의) 7. 앞으로 받게 될 총 보상의 기댓값이라는 것에 유의하자 8. Discounted Reward는 1개의 Episode에 대하여 discount factor(γ)의 개념을 도입하여 Reward의 합을 구한것이다. 9. Value Function은 이러한 합의 기댓값이다. 10. Value Function은 policy의 확률적인 영향뿐만 아니라 Environment의 확률적인 영향도 받는다. 11. 배가 앞으로 이동할 때 바람이 이 배를 추가로 이동시켰다면 보상은 달라질 것이다. Value Function은 이러한 영향까지 모두 고려한 수치이다. 12. 중요 개념 13. Prediction: 미래(Reward)를 평가 / Policy가 주어짐 14. Prediction Problem: Value Function을 학습시키는 문제 15. Control: 미래(Reward)를 최적화 16. Control Problem: best Policy를 찾음 17. 마르코프 결정 과정(MDP, Markov Decision Process)는 의사결정 과정을 모델링하는 수학적인 틀을 제공한다. 이 때 의사결정의 결과는 의사결정자의 결정에도 좌우되지만, 어느 정도 임의적으로 주어진다. 마르코프 결정 과정은 동적 계획법과 강화 학습 등의 방법으로 푸는 넓은 범위의 최적화 문제에 유용한 도구로 활용되며, 로봇 공학, 제어 자동화, 경제학, 제조업 등의 영역에서 폭넓게 사용되고 있다. 18. 세 개의 상태(연두색 원)와 두 개의 행동(붉은색 원) 및 두 개의 보상(붉은색 화살표)가 있는 간단한 마르코프 결정 과정의 예시. 19. 마르코프 결정 과정의 예시 |