강화 학습 개론

Assignment 2: Model-Free Reinforcement Learning

1. 목표

- Open AI Gym Taxi-v3 환경에서의 Model-Free RL 방법으로 문제 해결하기

2. 개발 환경

- Python (version 3.6+)
- 조교는 리눅스 (Ubuntu) 환경에서 평가 합니다.

3. 템플릿 파일 및 제출 파일

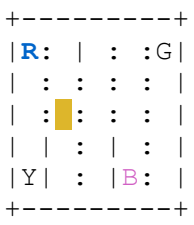
- 제공되어지는 템플릿 파일 2개
  - taxi.py: Taxi 환경 설정 및 기초 템플릿 코드 (**제출하지 않음**)
  - agent.py: Model-free RL 구현 후 제출할 코드

4. Open AI Gym Taxi-v3 환경 소개

*** 먼저 Open AI Gym 을 설치 합니다.**

"There are 4 locations (labeled by different letters), and our job is to pick up the passenger at one location and drop him off at another. We receive +20 points for a successful drop-off and lose 1 point for every time-step it takes. There is also a -10 point as the penalty for illegal pick-up and drop-off actions."

https://github.com/openai/gym/blob/master/gym/envs/toy_text/taxi.py



- 5x5 Map

  env = gym.make('Taxi-v3')

  action_size = env.action_space.n

  space_size = env.observation_space.n

- State Space = 5 X 5 X 5 X 4 = 500.

  State 는 taxi locations (5 x 5), 5 possible passenger locations (4 corresponding to RGBY with 1 additional of being in the taxi), and 4 possible destinations (corresponding to RGBY).

- The Action Space = 6

  0. down, 1. Up, 2. Right, 3. Left, 4. Pickup, 5. Dropoff

  Action 은 이미 정의되어 있으며 agent.py 에서 0 부터 5 까지 새로 정의할 필요 없습니다.

- 위와 같이 환경이 reset( ) 되면, R, G, Y, B 지점에 승객이 기다리고, 상하좌우로 움직여서 해당 위치로 이동한 후 pickup 한 후 목적지로 가서 dropoff 해야함.

- next_state, reward, done, _ = env.step(action)

  reward 는 -1, -10, 또는 20

  done 은 올바른 목적지에 승객을 내려주어 20 점 보장을 받을 때 True,

  또는 200 step 을 초과하면 True, 그 이외는 False.

5. taxi.py 설명

*** taxi.py 파일은 수정하지 않으며 제출하지도 않습니다.**

해당 파일을 실행하면 아래와 action space 크기와, state space 크기를 출력한 후 메뉴가 출력됨.

```
1. Testing without learning
2. MC-Control
3. Q-learning
4. Testing after learning
5. Exit
select:
```

그림 1. 실행 옵션 설정

1 번은 학습과 무관하게 맵을 이동하고 승객을 태우고 내리는 행동을 수동으로 해볼 수 있음.

2 번은 Monte Carlo control 방식으로 학습 후 최근 100 번 에피소드의 평균 reward 가 **6.0** 이상이 되도록 함.

3 번은 Q_learning 방식으로 학습 후 최근 100 번 에피소드의 평균 reward 가 **7.0** 이상이 되도록 함.

4 번은 학습 하기 전에는 음수 값이 출력됨(무작위). 2, 3 번을 통해 MC-control 또는 Q-learning 학습으로 업데이트된 q-table 을 활용하여 테스트 평균 reward 로 각각 **6.0** 또는 **7.0** 이상 나오는지 확인.

```
1. testing without learning
2. MC-control
3. q-learning
4. testing after learning
5. exit
select: 4
avg: -782.43
```

그림 2. 학습하지 않고 test.

```
1. testing without learning
2. MC-control
3. q-learning
4. testing after learning
5. exit
select: 4
avg: 7.63
```

그림 3. '2, MC' 로 학습 후 test.

6. agent.py 설명

*** agent.py 파일명은 "학번.py" 로 수정한 후 제출합니다.**

```
class Agent:
    def __init__(self, Q, mode="mc_control"):
```

생성자로 넘겨받는 Q 변수가 q-table 에 해당함. mode 로 넘겨받는 값이 "mc_control", "q_learning", 또는 학습 완료 후 테스트 목적인 "test_mode" 에 따라 적절히 학습 및 테스트를 수행하도록 개발함.

템플릿 코드는 자유롭게 수정할 수 있음. 단 taxi.py 파일 수정 없이 연동되어야 함.

참고 : Epsilon의 값에 따라 학습 속도가 많이 차이가 나게 됩니다.

40000 episode 이전에 양수가 나오는 게 일반적인 학습 속도 입니다.

Reward 1 부터 7까지의 학습 속도는 비교적 느린 편입니다.

```
♪Episode 100/100000 || Best average reward -810.38
♪Episode 200/100000 || Best average reward -1074.62
♪Episode 201/100000 || Best average reward -1083.62
♪Episode 1000/100000 || Best average reward -569.8
♪Episode 1001/100000 || Best average reward -549.7
♪Episode 2000/100000 || Best average reward -227.5
♪Episode 2001/100000 || Best average reward -221.84
♪Episode 5000/100000 || Best average reward -7.64
♪Episode 5001/100000 || Best average reward -12.92
♪Episode 10000/100000 || Best average reward -0.24
♪Episode 10001/100000 || Best average reward -0.14
♪Episode 50000/100000 || Best average reward 2.1
♪Episode 50001/100000 || Best average reward 2.16
♪Episode 90100/100000 || Best average reward 8.0
♪Episode 90101/100000 || Best average reward 7.96
```

그림 4. train 과정. (참고. epsilon 값에 따라 더 좋은 결과를 낼 수도 있습니다)

그림 4에서 90000회 이상에서 reward가 7 ~ 8의 값이 나오는 것을 볼 수 있습니다.

7. 제출 관련

- 제출 마감일: 4.10(일) 23:59.
  - 지연 제출은 받지 않습니다.
  - 표절은 양쪽 모두 F 입니다.
- "agent.py" 파일명을 "학번.py" 로 변경 후 iCampus 로 제출합니다.

8. 채점 기준

- Total 100 points
  - 40 points: 2. mc-control 학습 및 최종 100회 평균 reward 가 6.0 이상 출력.
  - 10 points: 3. Testing after learning 실행하여 mc-control 로 업데이트 된 Q-table 로 테스트 평균 reward 가 6.0 이상 출력.
  - 40 points: 2. q-learning 학습 및 최종 100회 평균 reward 가 7.0 이상 출력.
  - 10 points: 3. Testing after learning 실행하여 q-learning 으로 업데이트 된 Q-table 로 테스트 평균 reward 가 7.0 이상 출력.