Assignment (강화학습 - 김기응 교수님 연구실)

담당조교: 이병준, 윤든솔 (bjlee@ai.kaist.ac.kr, dsyoon@ai.kaist.ac.kr)

-과제설명

실습시간에 구현해본 SARSA를 확장한 **SARSA(λ)알고리즘을 구현**해보는 과제입니다. 매 타임스탭 마다 획득한 보상 정보를 토대로 Q 테이블을 업데이트 할 때, SARSA은 딱 한 개의 Q(s,a)값만 업데이트 시키는 점을 기억하시나요? 하지만 이 경우 maze 도메인에서처럼 Goal 상태에서 받는 보상 정보가 초기 상태의 Q값에 전달되어 반영될 때까지는 많은 시간을 필요로 하게 됩니다.

SARSA(λ) 알고리즘은 에이전트의 이동해온 경로, trace를 정보를 이용해 Q 테이블의 **여러 (s,a) 쌍들에 대한 값을 한 번에 업데이트** 시킵니다. 이러면 Goal 지점에서 받은 보상 신호가 기존 SARSA에 비해 훨씬 빠르게 다른 (s,a) 쌍들로 퍼져나가겠죠?

다음 페이지 SARSA(λ) 알고리즘의 Pseudo-code를 참고하여 구현해주시면 됩니다. (SARSA의 Pseudo-code는 참고 및 비교용으로 함께 첨부드렸습니다.)

-필요환경

numpy, gym, pygame

-제출물

완성된 코드(assignment.py)를 제출해 주시면 됩니다

-채점

채점은 코드 10점 만점으로 채점됩니다.

Algorithm 1 SARSA

$Q(\tilde{s}, \tilde{a}) \leftarrow 0$ for all $(\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$
for episode $= 0, 1, \cdots$ **do**
  $s \leftarrow$env.reset()
  $a \leftarrow$ EpsilonGreedy$(Q(s, \cdot))$
  for $t = 0, 1, \cdots$ **do**
    $(s', r, \text{done}) \leftarrow$env.step$(a)$
    $a' \leftarrow$ EpsilonGreedy$(Q(s', \cdot))$
    if done **then**
      $\delta \leftarrow r - Q(s, a)$
    else
      $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
    end if
    $Q(s, a) \leftarrow Q(s, a) + \alpha\delta$
    if done **then**
      break
    end if
    $s \leftarrow s'$ and $a \leftarrow a'$
  end for
end for

Algorithm 2 SARSA(λ)

$Q(\tilde{s}, \tilde{a}) \leftarrow 0$ for all $(\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$
for episode $= 0, 1, \cdots$ **do**
  $e(\tilde{s}, \tilde{a}) \leftarrow 0$ for all $(\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$
  $s \leftarrow$env.reset()
  $a \leftarrow$ EpsilonGreedy$(Q(s, \cdot))$
  for $t = 0, 1, \cdots$ **do**
    $(s', r, \text{done}) \leftarrow$env.step$(a)$
    $a' \leftarrow$ EpsilonGreedy$(Q(s', \cdot))$
    if done **then**
      $\delta \leftarrow r - Q(s, a)$
    else
      $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
    end if
    $e(s, a) \leftarrow e(s, a) + 1$
    for all $(\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$ **do**
      $Q(\tilde{s}, \tilde{a}) \leftarrow Q(\tilde{s}, \tilde{a}) + \alpha\delta e(\tilde{s}, \tilde{a})$
      $e(\tilde{s}, \tilde{a}) \leftarrow \gamma\lambda e(\tilde{s}, \tilde{a})$
    end for
    if done **then**
      break
    end if
    $s \leftarrow s'$ and $a \leftarrow a'$
  end for
end for

- SARSA(λ) 알고리즘에는 trace를 얼마나 빨리 감쇠 시킬지 결정하는 λ 파라미터가 하나 더 존재합니다. (λ = 0) 이면 SARSA(λ)와 SARSA 알고리즘은 동일해집니다.