Assignment (강화학습 - 김기응 교수님 연구실)

담당조교: 이종민 (jmlee@ai.kaist.ac.kr)

-과제설명

실습시간에 구현해본 Q-Learning을 확장한 **Watkin's Q(λ)알고리즘을 구현**해보는 과제입니다. 매 타임스탭마다 획득한 보상 정보를 토대로 Q 테이블을 업데이트 할 때, Q-Learning은 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.

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

다음 페이지 Watkin's Q(λ) 알고리즘의 Pseudo-code를 참고하여 구현해주시면 됩니다. (Q-Learning의 Pseudo-code는 참고 및 비교용으로 함께 첨부드렸습니다.)

-제출물

완성된 코드(assignment_rl.py)와 보고서 (docx나 pdf로 형식)를 제출해 주시면 됩니다

-채점

채점은 코드 5점, 보고서 5점, 총 10점 만점으로 채점됩니다.

보고서에는

- Maze 환경에서 (다른 도메인의 결과도 포함하셔도 좋습니다) Q-Learning과 Q(λ)을 각각 어떤 결과의 차이가 있는지
- 다양한 파라미터 (예: epsilon, alpha 등)를 바꿔가며 실행했을 때, 결과의 차이가 있는지. 있다면 어떻게 달라지는지

를 서술해주시면 됩니다.

Algorithm 1 Q-Learning

$Q(s,a) \leftarrow 0$ for all s, a
for episode $= 0, 1, \ldots$ **do**
  $s \leftarrow$ env.reset()
  $a \leftarrow \text{EpsilonGreedy}(Q(s,\cdot))$
  for $t = 0, 1, \ldots$ **do**
    $(s', r, \text{done}) \leftarrow \text{env.step}(a)$
    $a' \leftarrow \text{EpsilonGreedy}(Q(s',\cdot))$
    $a^* \leftarrow \arg\max_b Q(s', b)$
    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 Watkin's $Q(\lambda)$

$Q(s,a) \leftarrow 0$ for all s, a
for episode $= 0, 1, \ldots$ **do**
  $e(s,a) \leftarrow 0$ for all s, a
  $s \leftarrow$ env.reset()
  $a \leftarrow \text{EpsilonGreedy}(Q(s,\cdot))$
  for $t = 0, 1, \ldots$ **do**
    $(s', r, \text{done}) \leftarrow \text{env.step}(a)$
    $a' \leftarrow \text{EpsilonGreedy}(Q(s',\cdot))$
    $a^* \leftarrow \arg\max_b Q(s', b)$
    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 s, a **do**
      $Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a)$
      if $a' = a^*$ **then**
        $e(s,a) \leftarrow \gamma\lambda e(s,a)$
      else
        $e(s,a) \leftarrow 0$
      end if
    end for
    if done **then**
      break
    end if
    $s \leftarrow s'$ and $a \leftarrow a'$
  end for
end for

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