강화 학습

1주차
소개

01

1주차

강화학습의 정의

- 강화 학습

  - 환경안에서 정의된 에이전트가 현재의 상태를 인식하여 선택 가능한 행동들중 보상을 최대화 하는 행동

상태 + 행동 → 보상

02

1주차

강화학습의 정의

환경 : 대한민국

보상: 국민의 신뢰

행동: 박근혜 구속

행동: 우병우 구속

보상: 국민의 신뢰

에이전트

검찰
PROSECUTION SERVICE

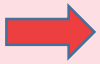
상태 : 박근혜 탄핵

03

1주차

강화학습의 적용

- 알파고
  - 바둑의 대가 이세돌로부터 승리



강화학습한 인공지능이 인간보다 잘함

강화 학습

3주차
Q-Learning(table)

01

3주차

Q란 무엇인가?

- Q
  - 현재의 상태에서 최선의 행동을 알려주는 하느님같은 존재

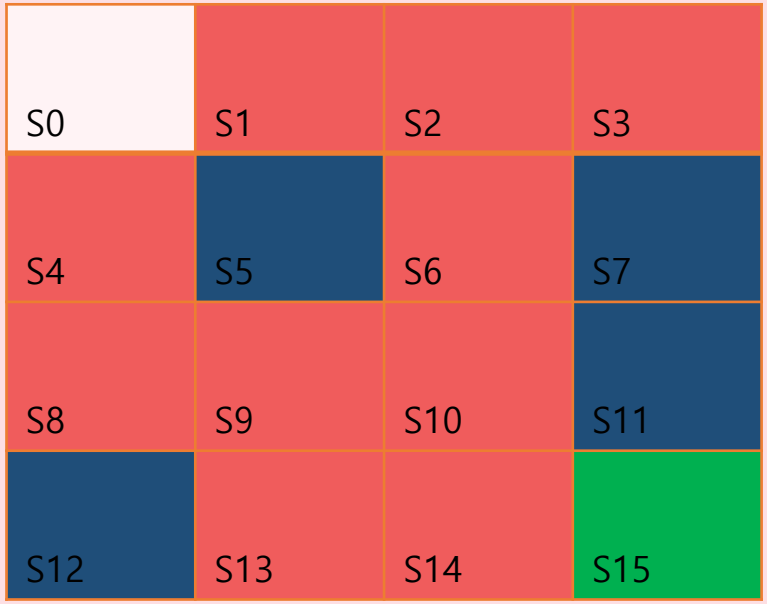
02

3주차

Q란 무엇인가?

Q

16개의 상태(S0,...,S15)
4개의 행동(위,아래,좌,우)



시작점
함정
도착점

03

3주차

Q의 알고리즘

- Q 의 알고리즘

$$\hat{Q}(s,a) = r + \max_{\grave{a}} \hat{Q}(\grave{s},\grave{a})$$

- Q 의 알고리즘 기원

$$\mathrm{R} = r_1 + r_2 + r_3 + r_4 + \cdots + r_n$$
$$R_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_n$$
$$R_{t+1} = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_n$$
$$R_t = r_t + R_{t+1}$$
$$R_t^* = r_t + maxR_{t+1}$$
$$Q_{(s,a)} = \mathrm{r} + \max_{\grave{a}} \widehat{Q}(\grave{s},\grave{a})$$

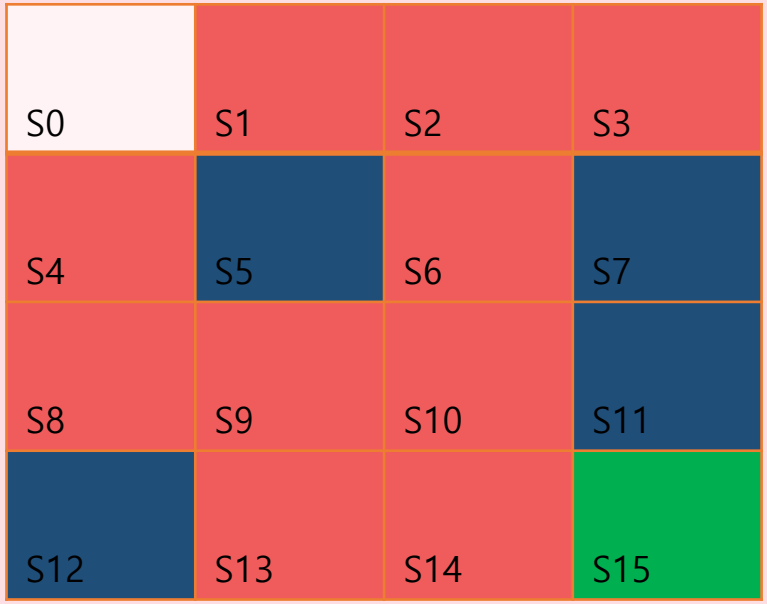
04

3주차

Q적용

Q

16개의 상태(S0,...,S15)
4개의 행동(위,아래,좌,우)



$$\hat{Q}(s,a) = r + \max_{\grave{a}} \hat{Q}(\grave{s},\grave{a})$$

05

3주차

알아야 할점

- Q는 상태 x 행동 크기의 행렬으로써 값은 0으로 초기화
- 액션을 추측하고 고를 수 있음
- 보상을 받는다
- 액션으로 인한 새로운 상태의 모든걸 볼수있음(새로운상태의 보상값)
- Q의 값 업데이트는 3장의 알고리즘을 따름

문제점

- 이런 알고리즘으론 간데만 간다.
- 이동제한이 없어 최적화를 할수없다.

강화 학습

4주차

Q-Learning(table)

exploit&exploration and discounted future reward

01

4주차

3주차 Q-Learning의 문제점

1.일관된 이동경로
2.최적화 안된 이동경로

02

4주차

1. 일관된 이동경로 해결방안

New

1. E-greedy
   - 20번중 2번은 새로운 데 가자! >>> random함수

많이 가봤는데 새로운데 그만가자...

2. decaying E-greedy
   - 처음에 많이 갔으니까 20번중에 1번만가자!

03

4주차

1. 일관된 이동경로 해결방안

Old

1. add random noise
   - A가 더 맛있긴 한데 질린다. B가자! >>> random value +

질려도 A만한데가 없더라....

2. decaying add random noise
   - 질려도 B보다 A이지!

Discounted reward
- 많이 움직이고 오면 상 많이 안줄꺼야!

어떻게?

t의 범위 $= 0<t<1$

$$R_t = r_t + t * r_{t+1} + t^2 * r^{t+2} + t^2 * r_{t+3} + \cdots + t^{n-t} * r_n$$
$$= r_t + t(r_{t+1} + (t * r_{t+2} + \cdots))$$
$$= r_t + t * R_{t+1}$$
$$Q_{(s,a)} = \mathrm{r} + t * \max_{\grave{a}} \widehat{Q}(\grave{s}, \grave{a})$$