AI_project2

Professor : 최상호 교수님

Student ID : 2020202037

Name : 염정호

Date : 2024.10.08

Introduction

본 프로젝트는 Grid word 환경에서 강화 학습과 DP 알고리즘을 시뮬레이션 하는 프로그램을 구현하고 학습 과정과 결과를 확인한다. Grid World 는 2 차원 배열 형태의 환경으로 시작점, 도착지점, 함정과 길이 존재하며 시작점과 도착점은 0, 함정은 100, 그 이외의 state 는 -1 로 초기화 된다. Grid world 7 * 7 크기로 설정되고 가능한 이동 가능한 경로는 상, 하, 좌, 우 4 가지 이다 Grid world 밖으로 나가는 행동은 금지된다. 다양한 행동과 상태의 조합을 통해 에이전트가 목표 달성하는 과정을 모델링한다.

Algorithm

- Policy iteration

  강화 학습에서 최적의 정책을 찾기 위해 사용되는 기본적인 동적 계획법 알고리즘으로 Policy Evaluation 과 policy_improvement 를 통해 최적의 정책과 가치 함수를 계산한다.

  Policy Evaluation 함수는 주어진 정책에 따라 state 의 가치 V(s)를 계산한다. 현재 진행 방향이 주어졌을 때 해당 진행 방향을 통해 기대되는 가치 V(s)를 계산한다.

$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

  delta 값을 사용하여 이전 가치와 현재 가치의 차이를 추적하면서 변화량이 특정 임계값보다 작아지는 경우 평가를 종료한다.

  Policy Improvement

  정책 평가를 통해 계산된 가치 함수를 사용하여 각 상태에서 최적 행동을 선택한다

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \sum_{s',r} P(s', r|s, a)\left[r + \gamma V(s')\right]$$

  정책 평가를 통해 얻어진 가치함수를 통해 상태별 최적의 행동이 반영된 새로운 정책을 출력한다.

각 상태에서 가능한 모든 행동을 평가하고, 가치가 최대가 되는 행동을 선택한다. 새로운 정책이 이전 정책과 모두 동일하면 수렴한 것으로 판단하고 반복을 종료한다. 위의 두 과정을 반복하여 최적 정책과 최적 가치 함수에 도달하는 것이 가능하다.

초기 정책은 무작위로 선택되며 반복된 개선을 통해 최적의 정책을 업데이트 한다.

value iteration

강화 학습의 동적 계획법, 알고리즘 중 하나로 최적의 가치함수 V 와 최적 정책을 계산하는 방법이다. 반복적으로 가치함수와 정책을 갱신하며 수렴할 때 까지 반복한다.

초기화 단계에서 상태 가치함수 V 는 0 으로 초기화 되며 policy 는 각 상태에서 최적의 행동을 정책을 저장할 배열로 초기화 된다. 우리는 장애물 위치의 value 를 -100 으로 고정함으로써 정책 개선에 사용하지 않게 하고, 다를 상태에서 도달하지 못하도록 한다. 모든 상태에 대해 가치 함수를 개선하며 최적의 행동을 결정한다. 각 상태에서 취할 수 있는 행동 중 가치가 가장 높은 것을 선택하여 가치함수를 업데이트 한다.

$$V(s) = \max_a \left[R(s,a) + \gamma \cdot V(s') \right]$$

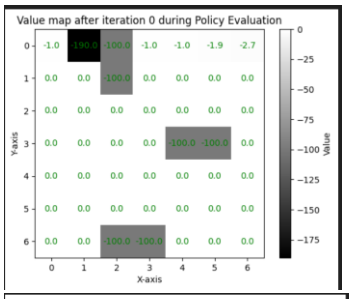
R(s,a)는 s 에서 a 행동을 선택했을 때의 보상이며 r 은 감가율로 미래 보상의 중요도를 결정하고 V(s′)은 행동 a 로 인해 도달할 상태 s′의 가치이다.

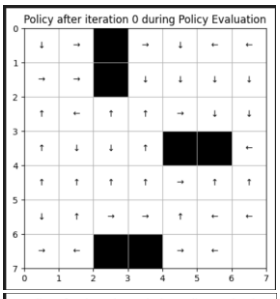
최적의 행동의 결정은 위 수식의 값을 최대로 하는 행동 a 를 선택한다.

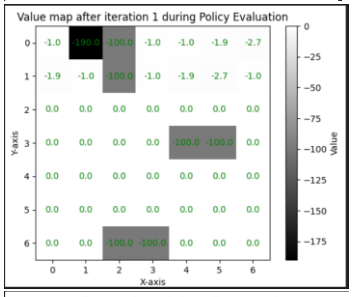
모든 상태에서 가치함수 V(s)의 변화량이 수렴 기준 보다 작아지면 알고리즘이 종료된다.

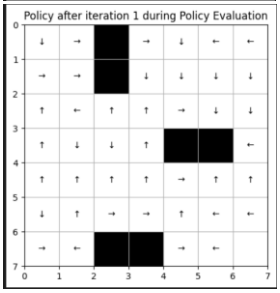
Result

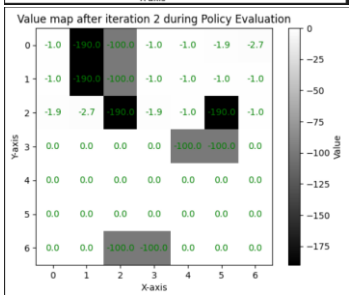
Policy iteration & value iteration

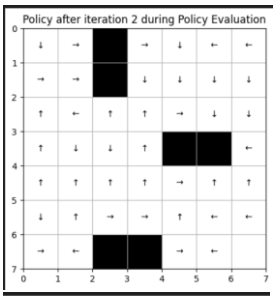


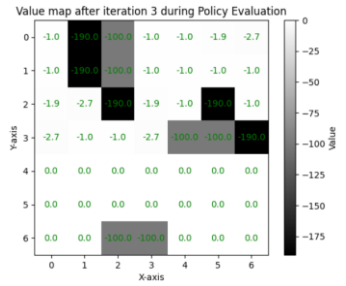


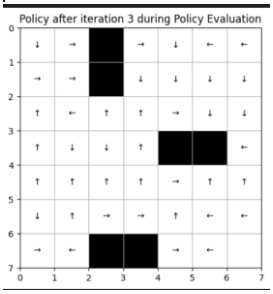




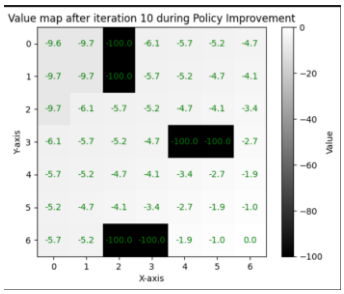


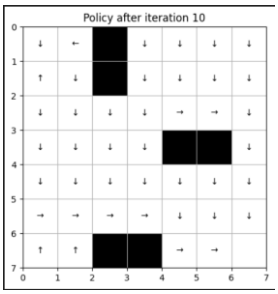


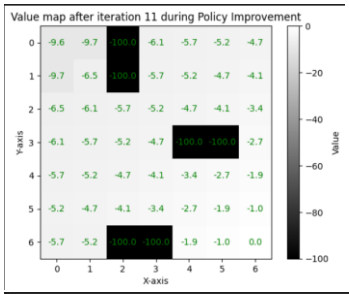


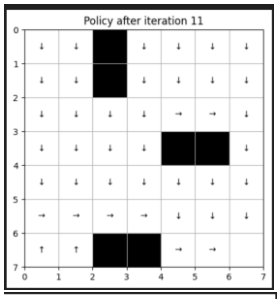


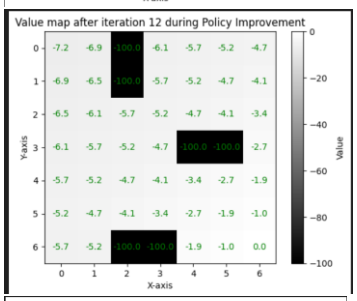
초반 1~5 회의 수행과정에서는 모든 state 의 value 가 할당되지 않았기 때문에 최적의 정책을 찾는 과정에서 어려움을 겪는 것을 볼 수 있다 4 방향 모두 value 가 0 인경우 학습과는 무관하게 특정 방향을 선택하기 때문이다.

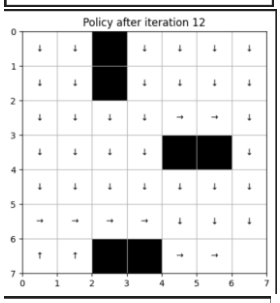


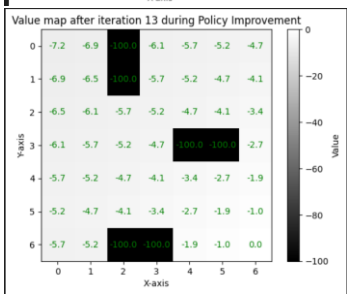


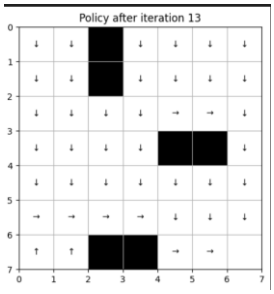




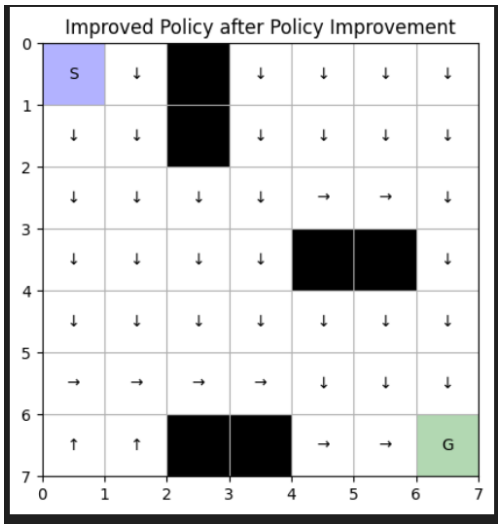


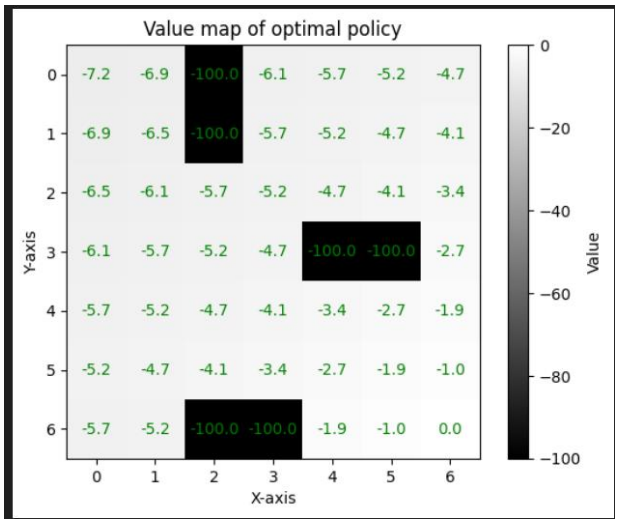






10 회 이후의 반복에서는 각 state 별로 value 가 모두 지정되었으며 이에 따라 최적의 policy 가 구해지는 것으로 보인다. 10 회차 반복에서는 부분적으로 다른 policy 가 보이는 것을 알 수 있으나 11 회 이후 부터는 value 와 policy 모두 수렴하는 것을 확인 할 수 있다. 최종 결과에서도 동일한 policy 가 적용되는 것을 볼 수 있다.





Consideration

본 프로젝트는 DP 알고리즘을 활용한 Grid World 에서의 정책 반복(Policy Iteration)과 가치 반복(Value Iteration) 알고리즘을 구현하는 과제를 수행했다. 실험에서는 7x7 크기의 그리드 환경에서 Policy Evaluation, Policy Improvement, Policy Iteration, Value Iteration 을 구현하여 최적의 경로를 학습하도록 했다. 가장 신경써야 했던 부분은 어떻게 수식을 통해 다음 경로를 지정할 것인가 였다. 지금 이 경로가 옳다고 하더라고 미래의 그 이후의 결과에 대해서는 알 수 없기 때문에 각 state 별로 value 를 구현한뒤 감마 값을 통해 학습 결과가 계속 누적되도록 구현하였고 계속된 학습을 통해 프로그램이 최종적으로 최적의 경로를 구할 수 있었다. 또한 함정과 목표 지점의 value 가 변하지 않도록 조정해야 했는데 해당 위치의 값들이 변하게 될 경우 경로에서 무한 루프를 돌 가능성이 생기기에 해당 값들은 -100 과 0 으로 고정시킨 뒤에 과제를 수행하였다.