Chapter06

벨만의 방정식

다이내믹 프로그래밍

FASTCAMPUS
ONLINE
금융공학/퀀트 I

강사. 장순용

I 키포인트

- 벨만의 방정식 (Bellman’s Equation).
- 다이내믹 프로그래밍 (Dynamic Programming).
- 가치 반복 알고리즘 (Value Iteration Algorithm).
- 정책 반복 알고리즘 (Policy Iteration Algorithm).
- 격자세상 (Grid World).

I 벨만의 방정식 (Bellman's Equation)

- 다음과 같은 두가지 방법으로 상태와 정책이 주어졌을 때의 가치 $v_\pi(s)$를 표현할 수 있다:

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma\, v_\pi(S_{t+1}) | S_t = s]$$

$$v_\pi(s) = R(s) + \gamma \times \sum_{s'} P_{s\,s'}^{\pi(S)}\, v_\pi(s')$$

I 다이내믹 프로그래밍 (Dynamic Programming) 개요

- 다이내믹 프로그래밍 (Dynamic Programming)은 복잡한 문제를 하위 문제로 나눈 후, 각 하위 문제를 해결하는 계층적 해결 방법이다.
- 벨만 방정식의 최적 정책을 구하는 목적으로 적용할 수 있다.
- 선형 대수학 연산에 기반한 수렴 알고리즘이다.

⇒ “작은” 문제에는 적용 가능하나, “큰” 문제에는 적용이 어렵다.

I 가치 반복 알고리즘 (Value Iteration Algorithm)

- 최적값을 찾을 때 까지 가치함수 $v(s)$만 반복해서 갱신한다.

  $\Rightarrow$ 최적 가치함수 $v^*(s)$가 구해졌다면 이것을 최적 정책 식에 대입해서 최적 정책 $\pi^*(s)$을 구할 수 있다.

- 다음의 순서로 진행된다:

a. 모든 상태 s에 대해서 $v(s) = 0$으로 초기화 한다.

I 가치 반복 알고리즘 (Value Iteration Algorithm)

b. 모든 상태 s에 대해서 $v(s)$를 다음 방식으로 계산하여 갱신한다.

$$v(s) = R(s) + \gamma \times \max_a \sum_{s'} P^a_{s\,s'} \cdot v(s')$$

⇒ 동기화 갱신: 등호 오른쪽을 완전히 계산하여 등호 왼쪽으로 동시에 대입한다.

⇒ 비동기화 갱신: 모든 상태값을 하나씩 계산하여 갱신해 나간다.

c. 단계 b를 반복 실행하면 모든 상태에서 최적 가치함수 $v^*(s)$로 수렴한다.

I 가치 반복 알고리즘 (Value Iteration Algorithm)

d. 최적 가치함수 $v^*(s)$를 가지고 다음과 같이 최적 정책을 구할 수 있다. 여기에서 argmax는 $\sum_{s'} P^a_{s\,s'} \cdot v^*(s')$를 최고화하는 행동 a를 추출해 내는 역할을 한다.

$$\pi^*(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P^a_{s\,s'} \cdot v^*(s')$$

I 정책 반복 알고리즘 (Policy Iteration Algorithm)

- 최적의 정책에 수렴할 때까지 반복적으로 적용한다.

a. 랜덤으로 정책 π를 초기화 한다.

b. 선형 연립 방정식을 이용해서 현재 정책에 대한 벨만 방정식을 풀어서 모든 상태 s에 대한 가치함수 $v(s)$를 구한다.

c. 새롭게 갱신된 가치함수를 최적 가치함수인 것처럼 사용하여 정책을 갱신한다.

$$\pi(s) = \underset{a}{argmax} \sum_{s'} P^{a}_{s\,s'} \cdot v(s')$$

I 정책 반복 알고리즘 (Policy Iteration Algorithm)

d. 단계 b와 c를 반복하면 가치와 정책은 최적값에 수렴한다.

$$v(s) \rightarrow v^*(s)$$

$$\pi(s) \rightarrow \pi^*(s)$$

I 격자 세장 (Grid World)

- 다음과 같은 격자세상 미로를 가정해 본다: 좌표.

시작 ⇨

(0,0)	(0,1)	(0,2)	(0,3)
(1,0)	장애물	(1,2)	(1,3) 함정
(2,0)	(2,1)	(2,2)	(2,3) 목표

I 격자 세장 (Grid World)

- 다음과 같은 격자세상 미로를 가정해 본다: 보상 구조.

시작 ⇨

-0.02	-0.02	-0.02	-0.02
-0.02	장애물	-0.02	-1
-0.02	-0.02	-0.02	+1

함정

목표

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.