Chapter07

강화학습

강화학습 개요

FASTCAMPUS

ONLINE

금융공학/퀀트 I

강사. 장순용

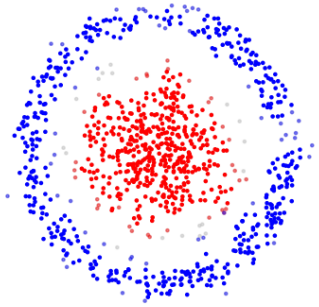
키포인트

- 강화학습 (Reinforcement Learning).

I 강화학습

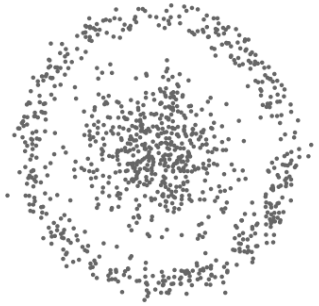
머신러닝의 유형

지도학습



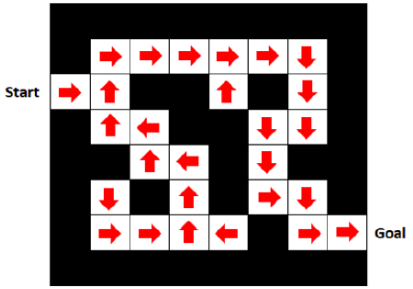
학습 목표가 분명함.

비지도학습



스스로 패턴을 밝혀냄.

강화학습



최적의 정책을 학습.

I 강화학습 개요

- 다이내믹 프로그래밍은 상태의 수가 늘어갈 수록 또는 차원이 증가할 수록 계산의 복잡도가 폭발적으로 증가한다는 단점이 있다.
- 왜냐하면 환경 모델이 그만큼 복잡해지기 때문이다. 그러므로 다이내믹 프로그래밍 방법을 그대로 적용할 수 있는 문제는 사실상 많지는 않다.
- 강화학습은 환경에 대한 모든 것을 모르더라도 환경과의 상호작용을 통해서 최적 정책을 학습할 수 있다.
- 인간이 바둑을 둘 때 모든 경우의 수를 다 알지는 못하며 게임이 진행됨과 동시에 일단 두어 보고 평가하며 본인의 전략을 갱신해 나가는 과정과도 유사하다.

I 강화학습

- 강화학습의 예: 바둑

⇒ 바둑은 주어진 상태에서 그 다음 최적 움직임을 알아야 하는 게임이다.

⇒ 지도학습의 방법으로 바둑을 두는 것은 매우 어렵다.

⇒ 주어진 판세 X에 대해서 그 다음 최적의 값 Y를 예측하는 것은 매우 어렵다.

⇒ 반면에 강화학습은 좋은 움직임에는 보상을 주고 반대의 경우에는 벌점을 줌으로서 알고리즘이 이기기 위한 전략을 스스로 배우도록 내버려 둔다.

I 강화학습

- 강화학습의 특징:

⇒ Agent는 순차적으로 의사결정을 한다. 시간의 개념이 중요한 역할을 한다.

⇒ 보상은 즉각적이지 않다. 몇 스텝 이후에 주어진다.

⇒ Agent의 행동은 후속 수신 데이터에 영향을 준다.

??

Left?? *Right??*

I 강화학습

- Agent의 유형: 가치기반 (Value based)

⇒ 모든 상태는 어떤 값을 갖는다 (할인된 미래 보상의 합).

⇒ 출발지점의 상태에 해당하는 값은 비교적 낮다.

⇒ Agent의 상태가 이 값을 키워가는 방향으로 전이되다 보면 도착지점에 다다르게 된다.

⇒ 도착지점은 상태의 값이 가장 큰 위치이다.

⇒ 그런데 Agent가 길을 잘못 들어서면 출발 지점 보다도 낮은 값의 상태에 빠질 수도 있다.

⇒ 이 유형에서는 정책이 명시되지는 않는다.

I 강화학습

- Agent의 유형: 정책기반 (Policy based)

⇒ 모든 상태에는 Agent가 따라야 할 정책 (향방)이 매핑되어 있다.

⇒ Agent가 단순히 정책을 따르면 보상값이 최대화 되는 방향으로 최적화 된다.

⇒ 이 유형에서는 상태에 따른 값이 없고 또한 가치 함수도 없다.

I 강화학습

- Agent의 유형: Actor critic

⇒ 정책과 가치가 명시되어 있다.

⇒ 단순 가치기반 또는 정책기반 보다 뛰어나다.

I 강화학습 방법

- 강화학습에서는 다음과 같은 두 가지 스텝을 반복하게 된다.

  ⇒ 평가 스텝 : Agent가 환경과의 상호작용을 통해 가치함수를 학습하는 것.

  ⇒ 제어 스텝 : 가치함수를 바탕으로 정책을 계속해서 갱신해 나가며 최적 정책을 학습하는 것.

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.