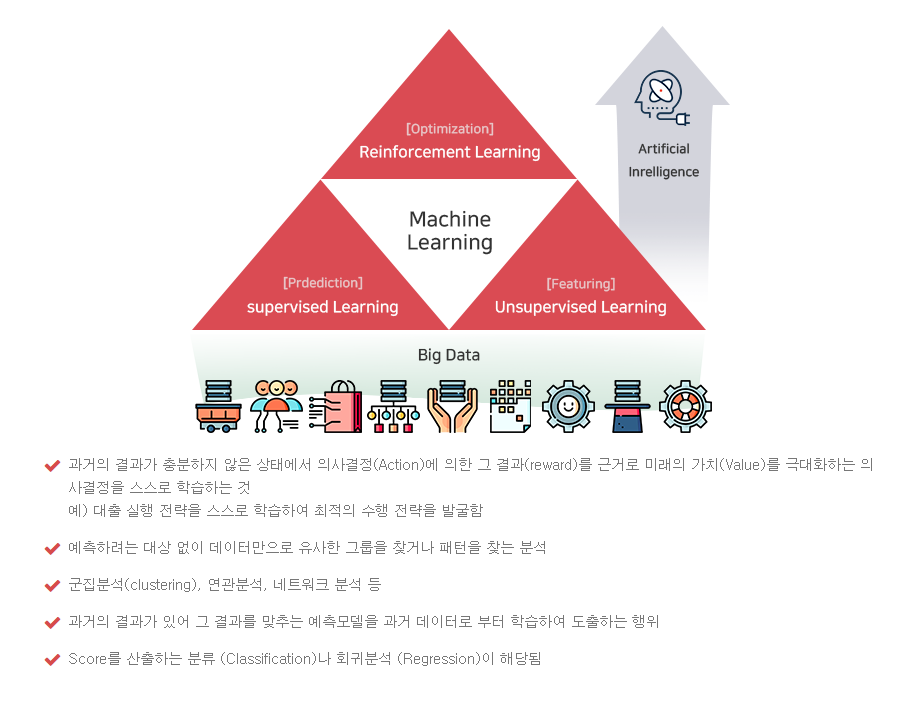
**강화 학습**

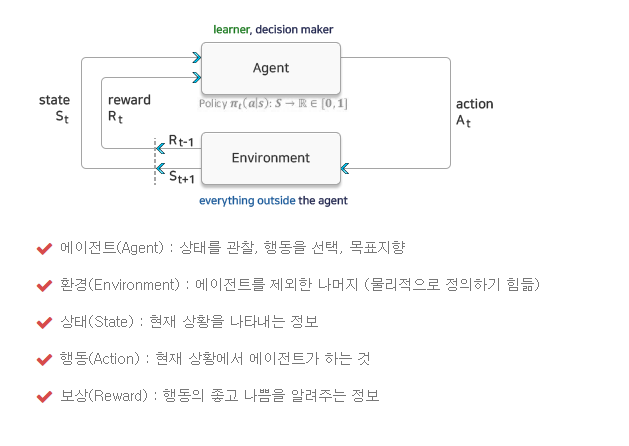
**1. 개념**

머신러닝 기법 중 최적화 분석용으로 가장 효과적인 방법 들 중에 하나이다. 기본 골자는 미래 가치를 극대화하기 위해 의사결정을 스스로 학습하는 방법으로 지도학습과는 달리 타켓은 성과이고 예측 값은 정책 혹은 수행전략이 된다. 다시 말해 Target = Reward이며(학습의 목표) Accuracy는 정책 혹은 수행 전략(Action)이 된다는 의미이다.



**2. 강화 학습(Reinforcement Learning)의 정의**

강화 학습은 기계 학습이 다루는 문제들 중 하나로 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법이다. 강화 학습에서 일어나는 기본 골자는 아래 그림과 같다.



위 그림에서 나타내는 것은 강화 학습의 가장 기초적인 부분으로 실제적으로는 좀 더 복잡하다. 아래는 좀 더 자세하게 각 부분에 대하여 설명한 것이다.

2-1. 에이전트: 에이전트는 행동을 취한다. 여기서 에이전트란 배송 서비스를 수행하는 드론이나 비디오 게임에서 슈퍼 마리오를 예로 들 수 있다.

2-2. 행동(A): 행동은 에이전트가 취할 수 있는 모든 행동을 말하며, 에이전트는 수행 가능한 행동의 리스트 중에서 앞으로 할 행동을 선택해야 한다. 비디오 게임에서의 리스트는 오른쪽이나 왼쪽으로 달리기, 높거나 낮게 점프하기, 웅크리거나 서있는 행동 등이 될 수 있다. 주식 시장에서의 리스트는 유가 증권 및 파생상품을 구매, 판매 또는 보유하는 것이 포함될 수 있으며 항공 드론의 경우 리스트에는 3D 공간에서의 여러 가지 속도와 가속도가 될 수 있다.

2-3. 할인융(Discount factor): 할인율은 보통 0과 1사이의 값으로 설정되며 이 값에 시간이 지남에 따라 에이전트가 발견하는 미래의 보상이 곱해져 에이전트가 선택한 행동에 대한 보상 효과를 약화시킨다. 그 이유는 할인율은 즉각적으로 주어지는 보상보다 상대적으로 가치가 낮은 미래의 보상을 만들기 위해서 고안되었기 때문이다. 보통 감마로 표현되며 감마가 0.8이고 3단계를 거쳐 10점을 보상으로 받는다면 보상의 현재가치는 0.8^3 x 10이 된다. 다시 말해 현재 가치는 (할인율^단계)\*보상이라 볼 수 있다.

1의 할인율은 즉각적인 보상만큼이나 가치 있는 미래의 보상을 만드는 것이다.

2-4. 환경(Environment): 에이전트가 이동하는 세계를 의미한다. 환경은 에이전트의 현재 상태 및 행동(입력)을 취하여 보상과 다음 상태(출력)을 반환한다. 어떤 사람을 에이전트라고 생각해보면 환경은 그 사람의 행동을 처리하고 그에 대한 결과를 결정하는 물리법칙과 사회 규칙과 같은 것이라 볼 수 있다.

2-5. 상태(S): 상태(State)란 에이전트가 인식하는 구체적이고 즉각적인 자신의 상황을 의미한다. 즉 도구, 장애물, 적, 또는 보상과 같은 중요한 요인들과 관련하여 에이전트가 마주하는 특정 장소와 시간이며 즉각적인 구성을 의미한다. 상태는 환경에 의해 반환되는 현재 또는 미래의 상황이라 할 수 있음

2-6. 보상(R): 보상이란 에이전트의 행동에 대한 성공이나 실패를 측정하는 피드백이라고 볼 수 있다(ex. 슈퍼마리오 게임에서 마리오가 동전에 닿으면 점수를 얻는 것). 특정 상태에서 에이전트는 선택한 행동을 출력 값으로 환경으로 보내고 환경은 그에 따른 에이전트의 새로운 상태(이전의 상태에서 취한 행동의 결과)와 보상을 반환한다. 에이전트의 행동에 의해 평가된 보상은 즉시 주어질 수도, 지연될 수도 있다.

2-7. 정책(π(파이)): 정책이란 에이전트가 현재 상태를 기준으로 다음의 행동을 결정하는 데 사용하는 전략이다. 에이전트는 특정한 상태에서 보상을 최대화할 수 있는 행동을 선택한다.

2-8. 가치(V): 단기적인 보상인 R과는 달리 Value는 장기적인 관점에서의 할인된 기대값이다. Vπ(s)란 현재의 상태에서 정책 π(파이)에 따른 기대되는 보상을 의미한다. 여기에서는 보상을 할인하거나 미래의 기대값을 더욱 낮추기도 한다.

2-9. Q- value 또는 action-value(Q): Q-value와 value는 비슷한데 Q-value는 현재 상태에서 취하는 행동 a를 고려한다는 것이 차이점이다. Qπ(S,a)은 정책 π에 따라 행동 a를 취할 경우 현재 상태 s에서 받을 장기적인 리턴을 말한다. Q는 상태-행동 쌍을 보상에 매핑한다. 여기서 Q와 정책의 차이점에 주의해야 한다.

2-10. 궤도(Trajectory): 그 상태에 영향을 미치는 연속적인 상태와 행동.

따라서 정리하자면 환경은 현재 상태에서 취한 행동을 다음 상태와 보상으로 바꾸는 함수이며, 에이전트는 새로운 상태와 보상을 다음 행동으로 변환하는 함수이다. 에이전트의 함수는 할 수 있지만 환경의 함수는 알 수 없다. 마치 블랙박스와 같아 입력값과 출력값만 볼 수 있다. 간단히 말하자면 그 내부는 어떻게 작동하는지 모르지만, 강화 학습 알고리즘은 에이전트가 앞으로 누적될 보상을 최대화하는 일련의 행동으로 정의되는 정책을 찾는 방법이다.

**3. 강화 학습의 상호작용**

3-1. 에이전트가 환경에서 자신의 상태를 관찰

3-2. 그 상태에서 어떤 기준(가치 함수: 현재 상태에서 미래에 받을 것이라 기대하는 보상의 합

3-3. 선택한 행동을 환경에서 실행

3-4. 환경으로부터 다음 상태와 보상을 받음

3-5. 보상을 통해 에이전트가 가진 정보를 수정함

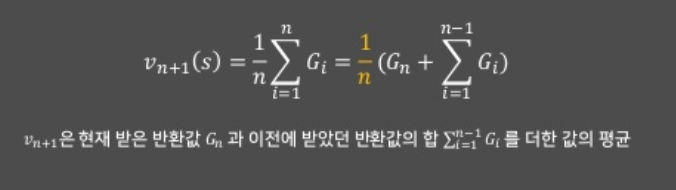
**4. 다이나믹 프로그래밍과 강화학습의 차이점**

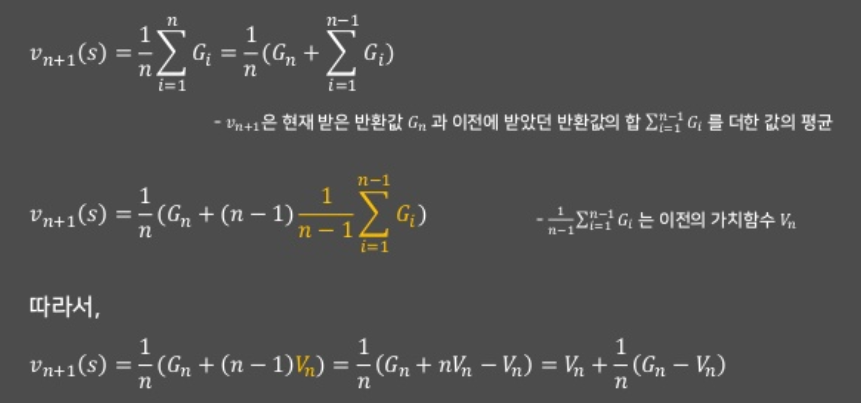
|  |  |
| --- | --- |
| **다이나믹 프로그래밍** | **강화학습** |
| 환경의 모델\*이 필요 | 환경의 모델이 불필요 |
| 정책평가 | 예측 |
| 정책 이터레이션(정책평가+정책발전) | 제어(예측과 함께 정책을 발전) |
| 환경에 대한 정확한 지식을 가지고 모든 상태에 대해 동시에 계산 | 일단 해보고, 자신을 평가하며, 평가한 대로 자신을 업데이트 (반복) |

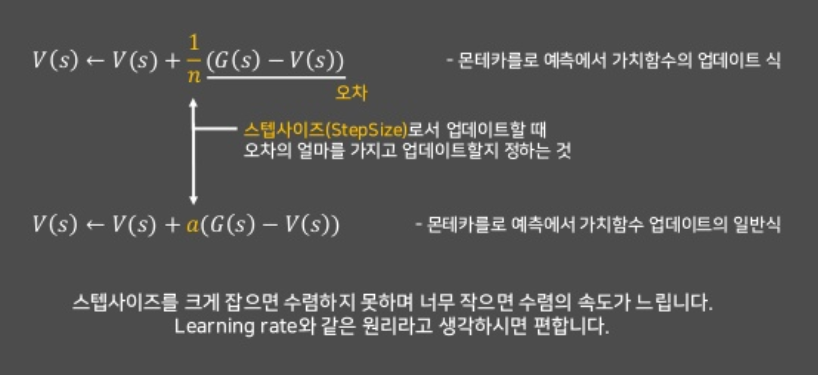
\*환경의 모델: 상태변환 확률과 보상을 가지고 만든 모델을 의미함

이때, 강화 학습은 “일단 해보고” 그 후에 가치함수의 값을 “예측”하는데 이를 위해 사용하는 방법이 몬테카를로 예측법과 시간차 예측법이 있음

\* 몬테카를로 예측법에서는 샘플링을 평균하여 가치함수의 값을 추정하는 방식을 사용하며 그 공식은 아래와 같다.







정리하자면 몬테카를로 예측은 샘플링을 평균하여 가치함수를 추정하는데



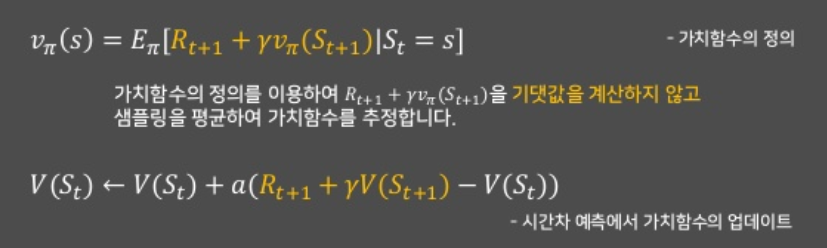
에이전트는 위의 업데이트 식을 통해 에피소드 동안 경험했던 모든 상태에 대해 가치 함수를 업데이트 하게 된다. 가치 함수를 업데이트하면 다음과 같은 두 가지 사실을 알 수가 있다.

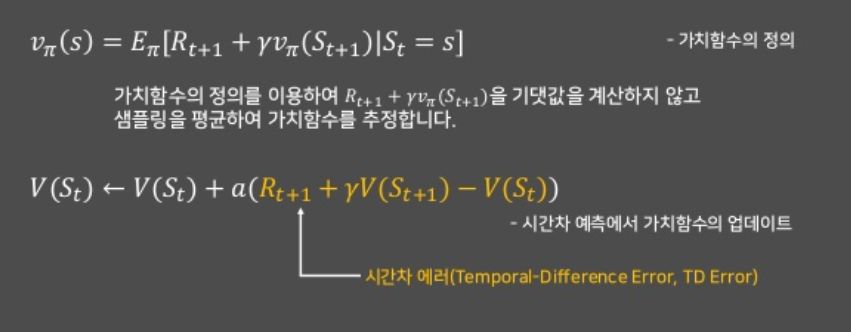
1. 업데이트 목표 = (G(s))

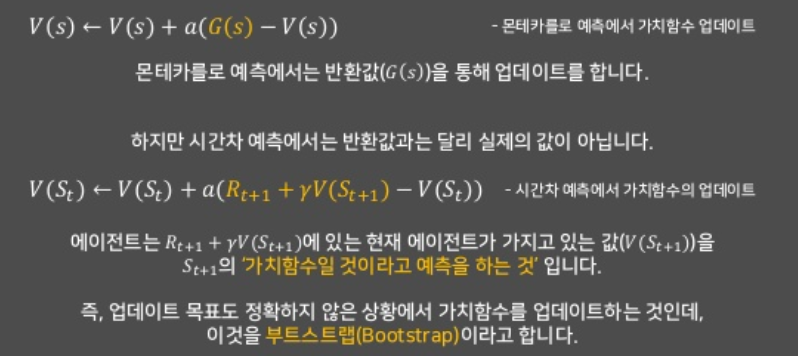
2. 업데이트의 크기 = a(G(s)-V(s))

\* 몬테카를로 예측법의 단점은 실시간 업데이트가 아니기 때문에 강화학습에서 중요하게 생각하는 시간차에 대한 것을 다루기가 매우 힘들다. 이를 위해서는 에피소드가 아니라 타임스텝마다 가치함수를 업데이트 해야 되는데 이를 보안한 방법이 시간차 예측이다.

\*시간차 예측에서의 가치함수의 업데이트는 실시간으로 이루어지며, 몬테카를로 예측과는 달리 한 번에 하나의 가치함수만을 업데이트 한다. 그 공식 및 방법은 아래와 같다.







정리하자면 시간차 예측도 마찬가지로 샘플링을 평균하여 가치함수를 추정함. 하지만 실시간 업데이트이기 때문에 시간차 예측은 에피소드가 끝날 때까지 기다릴 필요 없이 바로 가치 함수를 업데이트 할 수 있음. 가치 함수의 업데이트를 통해

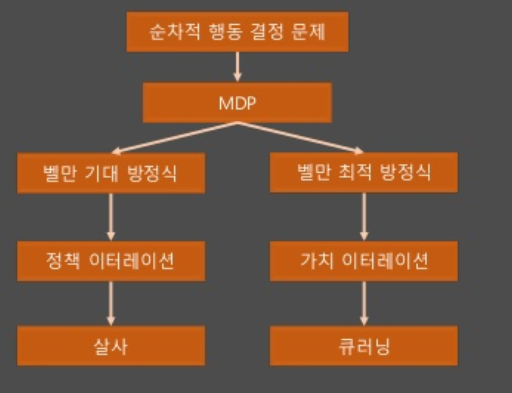
1. 업데이트 목표: 

2. 업데이트 크기: 

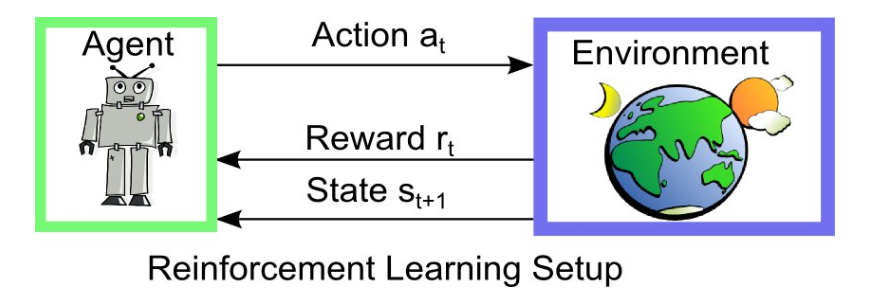
에 대하여 알 수가 있음.

**5. 강화학습 알고리즘 흐름도**

사용하는 강화학습 알고리즘에 따라서 다르지만, 보통 아래와 같은 흐름을 가지게 된다.



또한 강화 학습에서 주로 다루는 “환경”은 마르코프 결정 과정으로 주어지게 된다.



에이전트는 보상을 최대화하도록 학습이 진행되며 일종의 동적인 상태에서 데이터를 수집하는 과정까지 포함되어 있는 알고리즘이다. 강화학습의 대표적인 알고리즘은 Q-Learning이 있고, 딥 러닝과 경합하여 Deep-Q-Network(DQN) 방법으로도 사용된다.