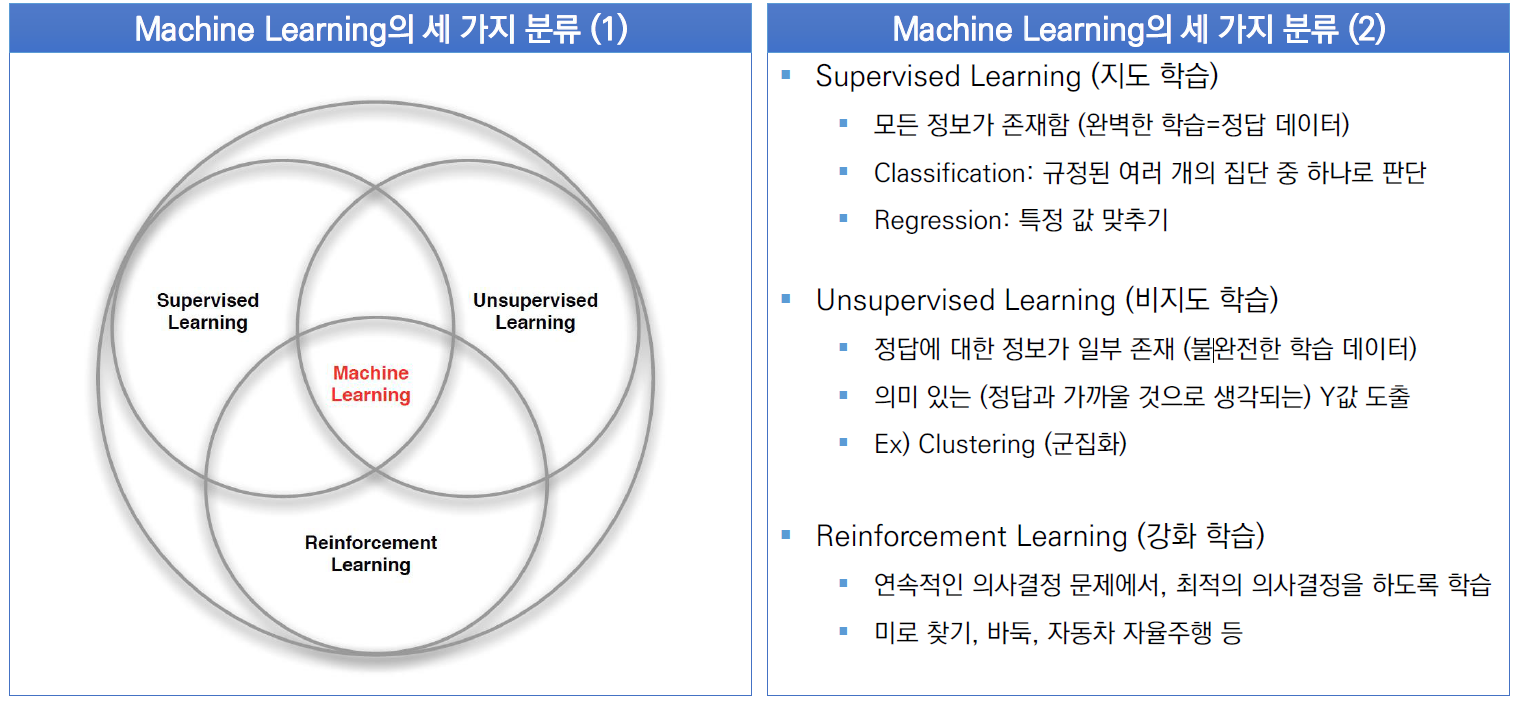
**The essence of learning from data**

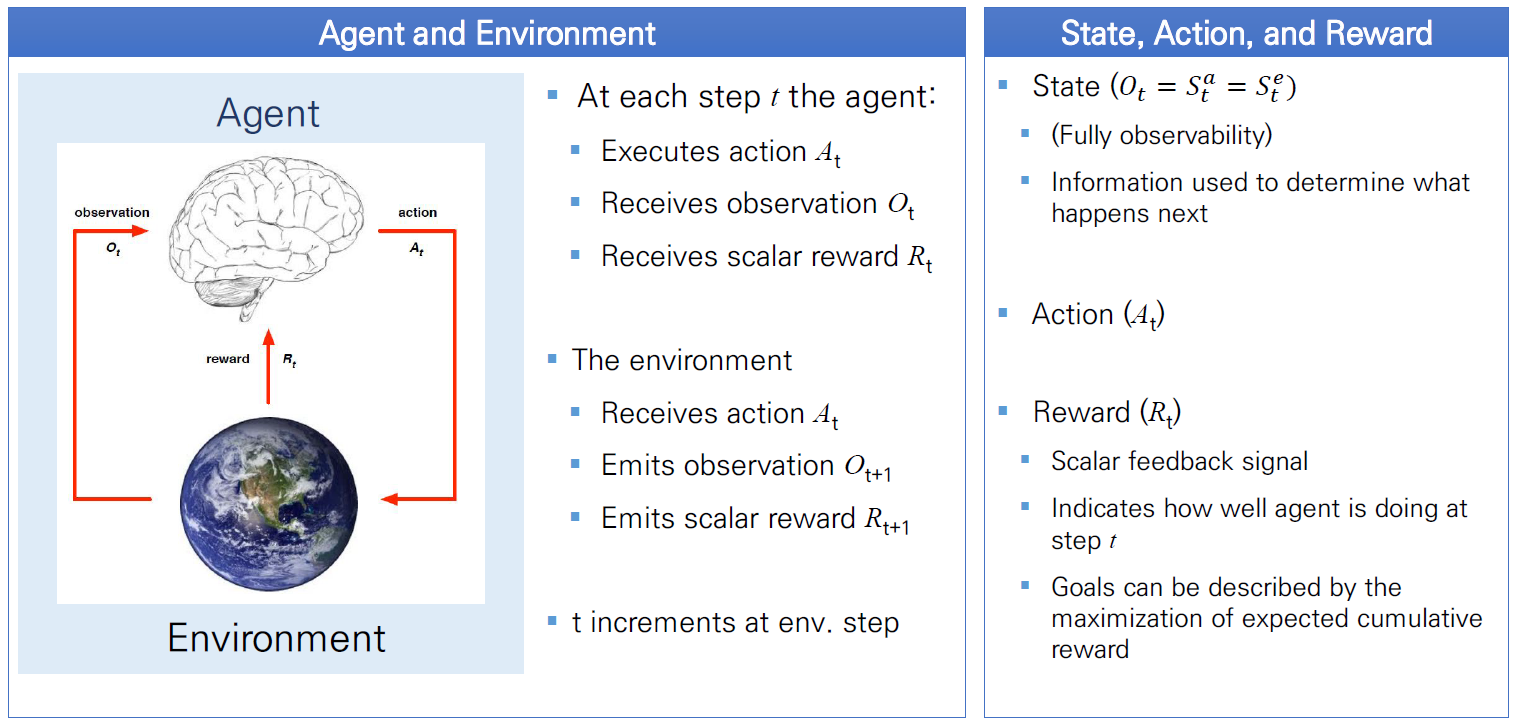
데이터

데이터에 특정 pattern이 존재함

수학적으로 해결할 수 없는 문제가 존재



* Reinforcement Learning는 한번에 끝나는 것이 아닌 연속적인 의사결정을 하도록 하는 것
* 강화 학습은 DP에서 파생된 것



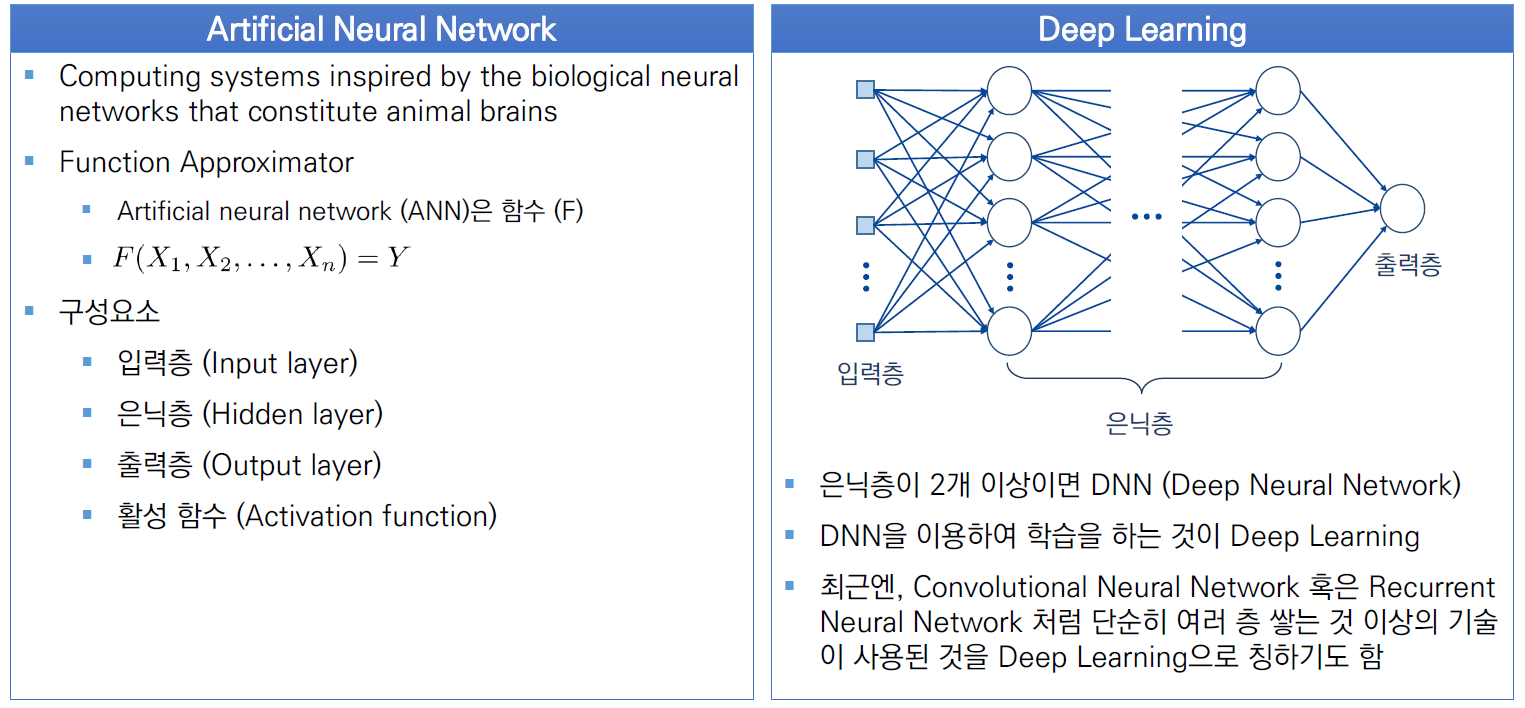
결국 Reward는 매순간 순간(sequence)이 아닌 최종적인 결과가 좋아야 높아질 수 있다.

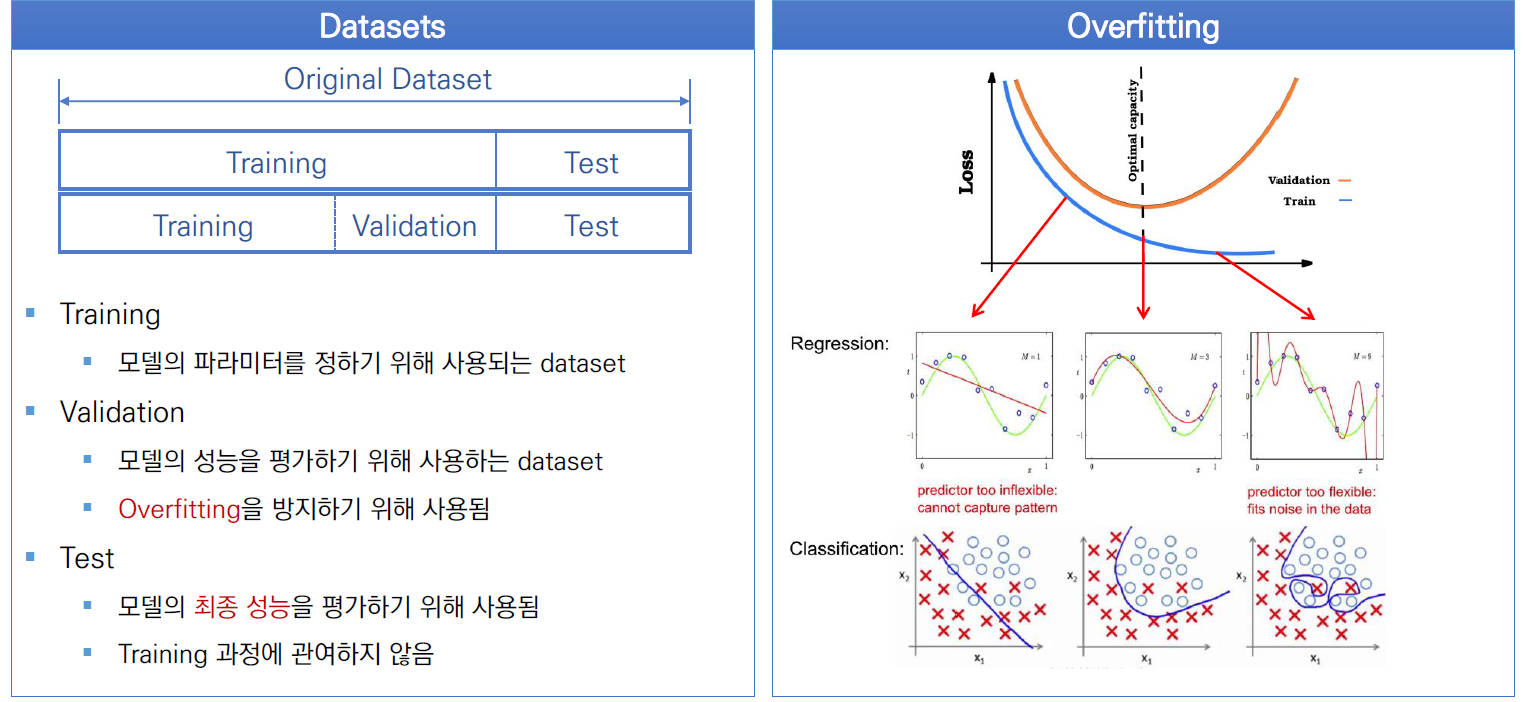
강화 학습은 supervised와 다르게 결과론적이 아닌 것

인간을 따라 하는 것은 강화 학습이 아님

강화학습은 피드백도 상당히 느림

**딥러닝**





예를 들어 100만명의 데이터를 가지고 학습을 시킬 시 모든 데이터를 학습시키는 것이 아니라 한 70만명의 데이터를 학습시키고 30만명의 데이터를 테스트한다.

일반적으로 데이터를 쪼개서 생각해야함 모든 데이터를 학습시켜야 하는 것이 아님

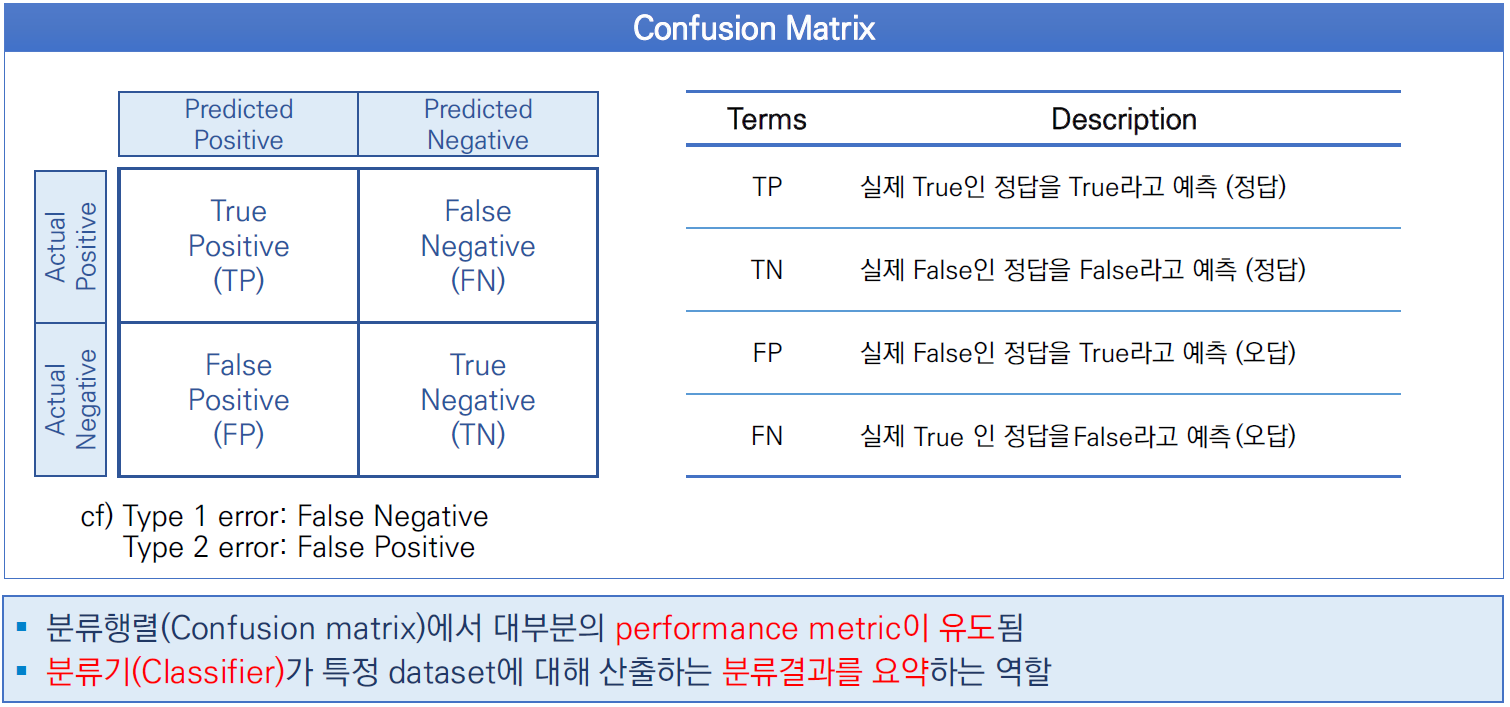
Overfitting – 학습이 너무 과하게 진행됨(Validation과 Train의 차이), Validation과 Train을 비교하여 Overfitting의 기준점을 알 수 있다.

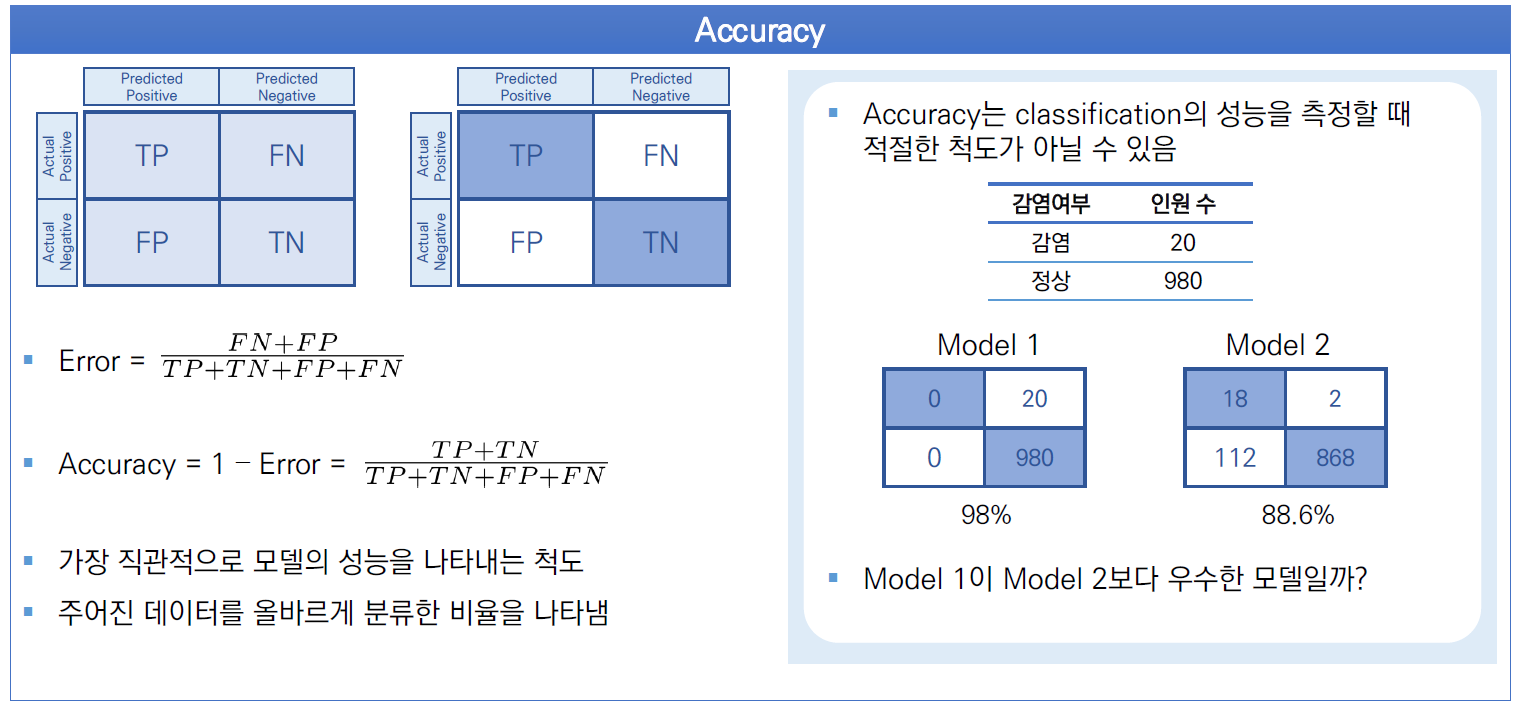
Regularization – 오버 피팅을 막기 위해 람다를 넣어 줌

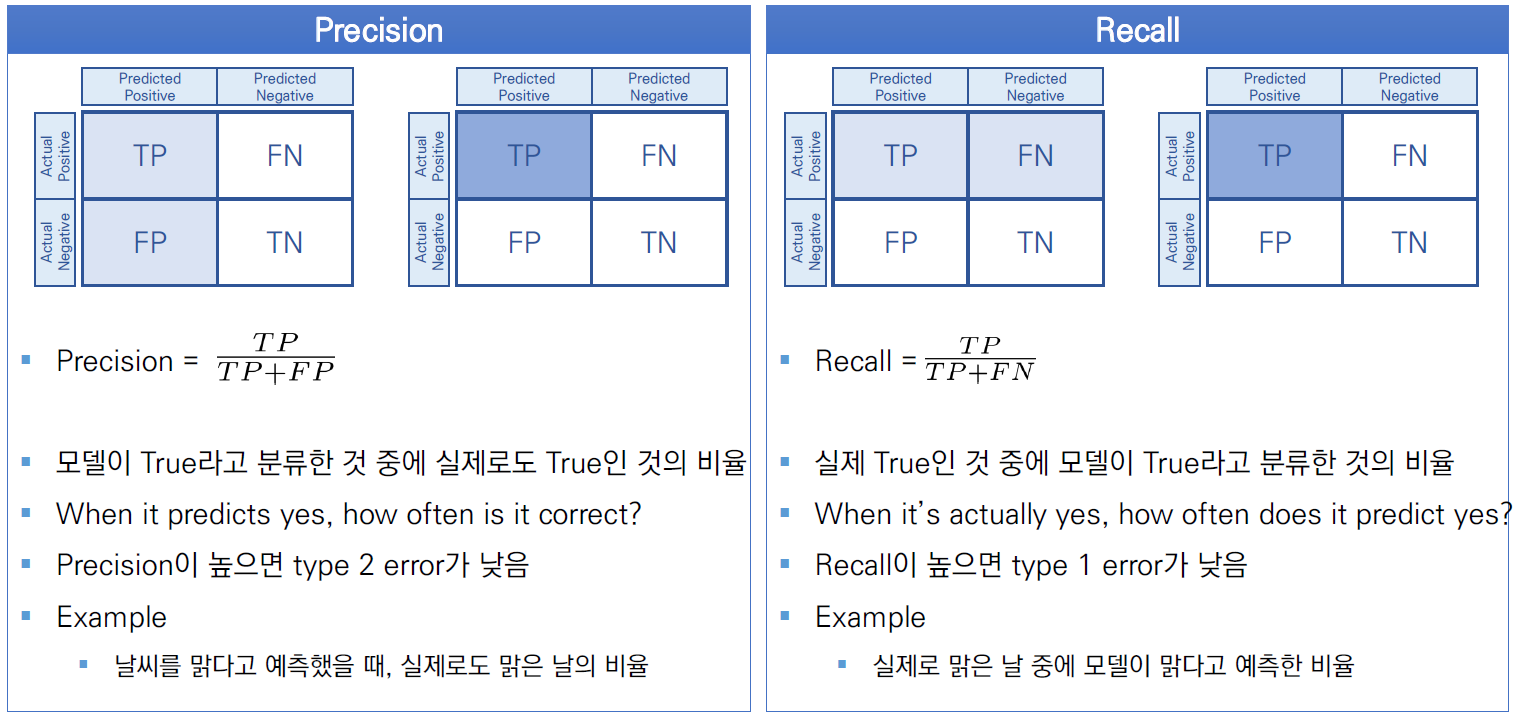
Drop out – 확률적 부분에서 오버 피팅을 막음

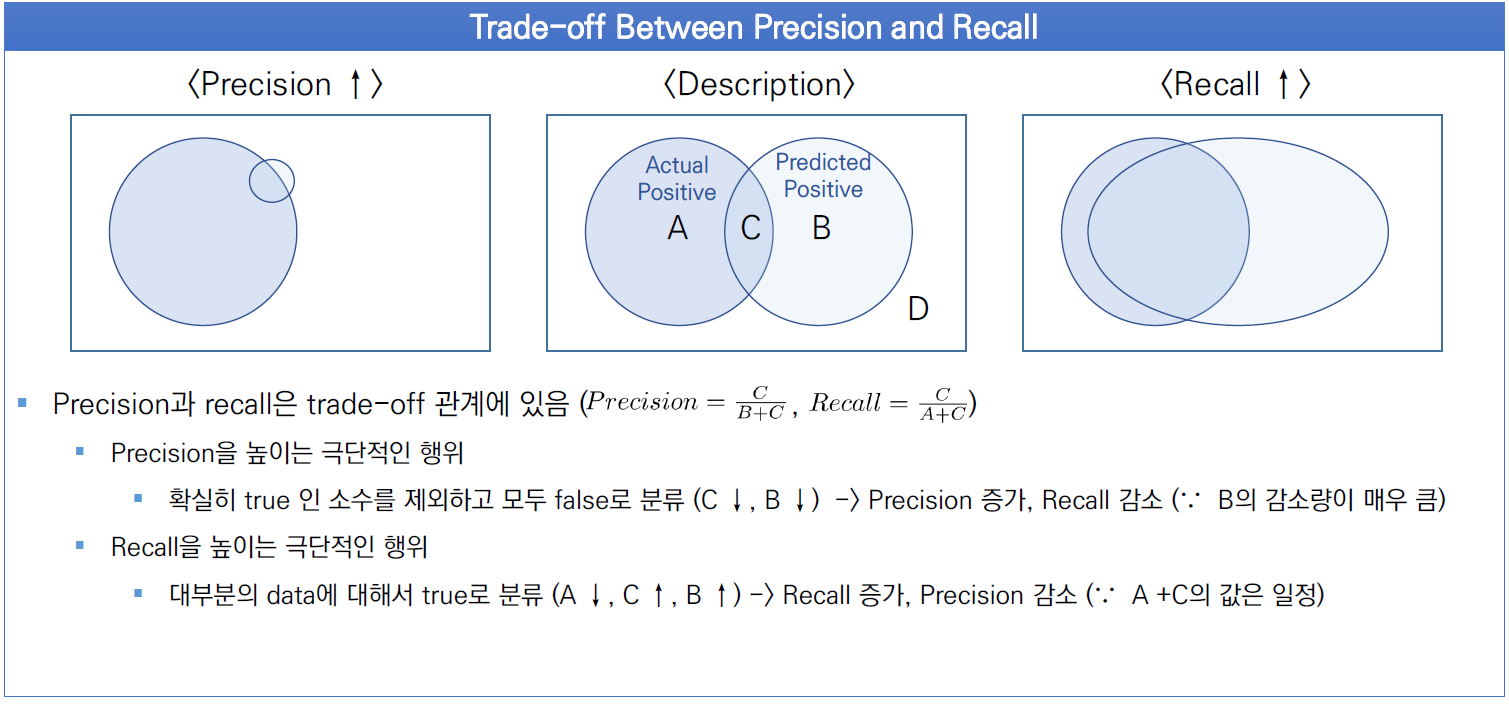
Validation set은 Training Process에서 관여 (Training 데이터에는 관여하지 않음)

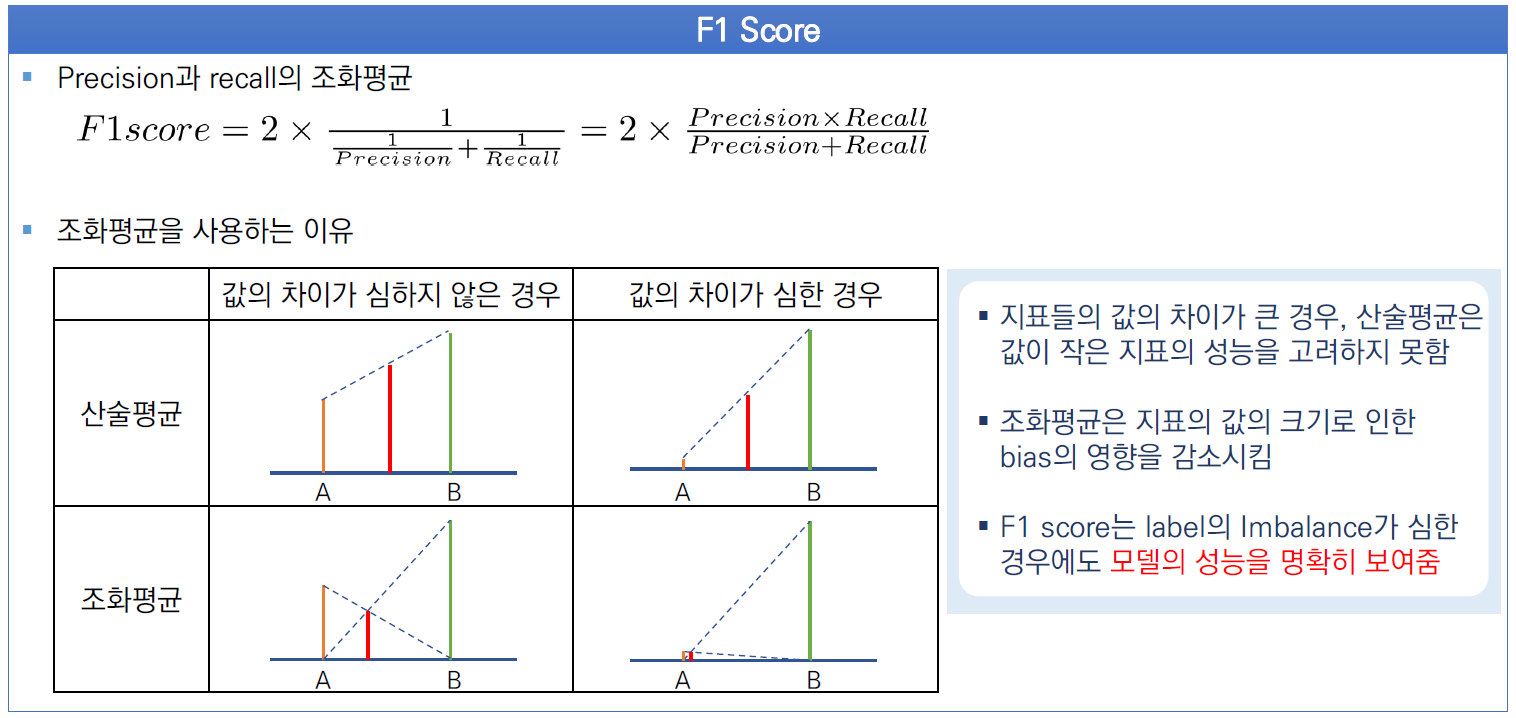
Validation data는 Training에서 사용하지만 Back tracking(개선)에서는 사용하지 않음

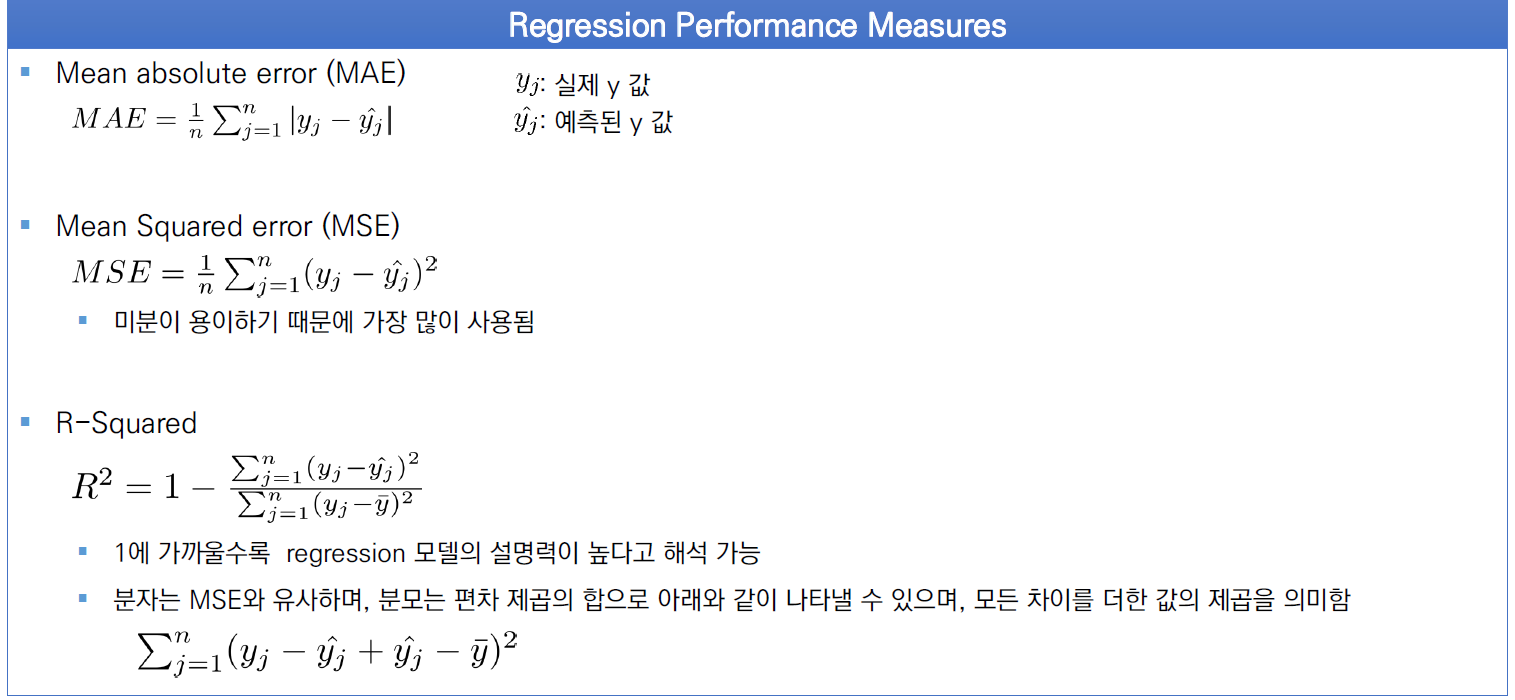












강화학습의 시초

Q-learning

Markov Decision Process

Markov Processes에서 액션(주체적인 의사)을 더함 🡪 Reward를 최대화할 수 있는 것(보상(승리, 정책)이 큰 것을 찾는 것)

S는 state A는 액션 P는 확률 R은 reward

Markov Processes – 어디로 수렴될지 확률을 구하는 것 (A에서 A로 갈 확률, A에서 B로 갈 확률 등등을 계산)

Markov property – The future is independent of the past given the present

Reward는 단순히 +1을 하는 것이 아니라 감마 상수를 reward에 곱해준다. (현재 눈 앞에 있는 것과 미래에 눈 앞에 있는 것을 고려하기 위함, 명확히 끝맺음을 위해 감마를 사용, 감마가 0에 가까울수록 현재 가치를 높이 추구, 1에 가까울수록 미래 가치를 추구)

Policy – 행동

Value Function – 어떤 상태에서 끝까지 갔을 때, 얻을 수 있는 기대 값

State Value Function – 어떤 상태에서 Policy를 따라 행동을 계속 했을 때, return의 기대 값

Action-value function – 어떤 상태에서 액션을 한 후 행동을 계속 했을 때, return의 기대 값

지금 상태에서 어떤 액션을 했을 때 과거의 것을 전부 보지 않고 바로 전 과거나 지금 상태를 보고 미래를 판단

Backup

미래의 값(next value function)들로 현재의 value function을 구함

지금 상태에서 끝까지 보는 것과 다음 상태에서 끝까지 보는 것은 다름

Bellman Optimality Equation 모든 Policy 중에서 optimal state-value function을 하는 것

모든 Policy 중에서 optimal action-value function을 하는 것