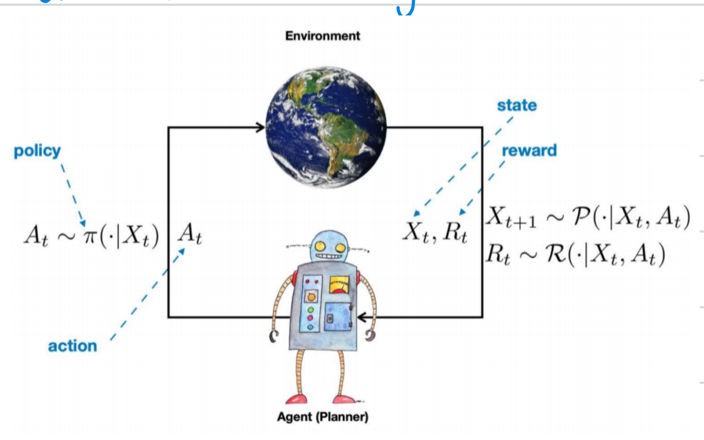
Reinforcement Learning

- taking suitable action to maximize reward in a particular situation
- Supervised Learning :-
  - labelled data
  - got target variables
- Unsupervised Learning :-
  - labelled data
  - no target variables
- Reinforcement Learning :-
  - trial & error
  - reward based learning

Reinforcement Learning :-



- Agent
  - decision maker & learner
- Environment
  - anything outside agent that it interact w/ & attempt to control
- State, X_t
  - variable that summarize whatever has happened to the agent so far
- Action, A_t
  - set of all possible action being on a state, X_t
- Policy, π
  - indicates action, A_t to be taken on a state, X_t
  - usually a mapping from state to actions
- Transition Probability Kernel
  - Describe the dynamic
  - Give action's effect in a state
- Reward, R_t
  - reward for simply being in the state, X
  - a real number specifying the immediate desirability of an action in a state

- example

	CHESS	ROBOT HAND
State, X_t	Position of pieces on board	Position of ball " hand coordinate finger
Action, A_t	movement of a piece	Δ/ part coord. of hand
Reward, R_t	+1 ⇒ checkmate +0.5 ⇒ capture a piece +0 ⇒ nothing happen	+1 closer to ball -1 further to ball
Policy, π	Prob of taking certain move train NN to output possible move	Prob of moving robot hand

- Sequence

$$\underbrace{X_1}_{\text{state } t=1}, \underbrace{A_1}_{\text{action } t=1}, \underbrace{R_1}_{\text{reward } t=1}, \overbrace{X_2, A_2, R_2}^{\text{an episode}}, \ldots$$

- terminate after certain number of t
- terminate after agent reach certain state
- never terminate

Markov Decision Process

- framework for decision making where outlines are partly random & partly under control of decision maker
- 5 state :-

$$\{\underbrace{X}_{\text{state}}, \underbrace{A}_{\text{action}}, \underbrace{P}_{\text{policy}}, \underbrace{R}_{\text{reward}}, \underbrace{\gamma}_{\text{discount}}\}$$

- Policy, P
  - state transition kernel
  - state action transition kernel
- $\bar{\pi} = \{\underbrace{\pi_1}_{\text{policy time}=1}, \pi_2, \pi_3\}$
- $\bar{\pi}_t(a_t \underset{\text{given}}{|} \underbrace{X_1, A_1, X_2, A_2, \ldots, X_t}_{\text{history } X, A})$

  conditional probability distribution given the history states selection
- $\pi_t(A | X_1, A_1, X_2, A_2, \ldots, X_t) = 1$

  policy prob. dist. over actions given history sums to 1
- Type of Policy
  1. Markov Policy (state trans. kernel)
     depend on X_t, $\pi(\cdot | X_t)$
  2. Deterministic policy (state trans. kernel)
     assign mass 1 to action for each state, X_t, $\pi_t(a_t | X_t) = 1$
  3. Stochastic Policy → can change (state trans. kernel)
     assign prob. dist over a given state X_t
  4. Stationary Policy (both)
     policy dont change over time
     $\bar{\pi} = \{\pi, \pi, \pi, \ldots, \pi\}$

Reward

- $\underbrace{R_t}_{\text{immedeate reward}} \sim R(\cdot | X_t, a_t)$
- $r(x, a) = \mathbb{E}(R | X = x, A = a)$
  - ave. reward that repeated interaction w/ environment
- receive within one episode as a measure of performance
- a way to maximize long term reward

Task

- ① Finite Horizon Task
  - interact for fixed predefined times
  - eg. solve a maze w/ fixed steps
  - agent goal is to maximize cummalative rewards
  - $G^\pi \triangleq R_1 + R_2 + \ldots + R_t$ } sum of rewards } return function
    $G^\pi \triangleq R_1 + \gamma R_2 + \ldots + \gamma^{t-1} R_t$ } discounted sum } return function
    $V^\pi = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} R_t \mid X_1 = x\right]$ } value function of π
- ② Episodic
  - well defined start & end
  - terminat when reach a certain state
  - eg. chess
  - $G^\pi \triangleq \sum_{k=1}^{T} \gamma^{k-1} R_k$ } return function
    $V^\pi(x) \triangleq \mathbb{E}(G^\pi | X_1 = t)$ } value function
- ③ Continuity
  - no endpoint
  - eg. control robot for navigation
  - $V^\pi(x) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \mid X_1 = x\right]$ } value funct.
    $Q^\pi(x) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \mid X_1 = x, A_1 = a\right]$ } action-value function