Week 3

Friday, September 6, 2019 1:05 PM

Policies

In reinforcement learning, the main goal is to maximize the return value. The agent's task is to select an action which produces a reward and the choice of this action may influce the reward both immediatly and also in the long run.

In an RL system, the actions are selected by following a policy. A policy maps a state to an action. This kind of policy is called the deterministic policy and is denoted by the symbol π.

ie $\pi(a|s)$

This gives the probability of selecting action a given the state s.

If multiple actions can be selected with ≥ 0 probability the policy is called a stochastic policy. In this case, since there can be actions with multiple probabilities

$$\sum_{\forall a \in A} \pi(a|s) = 1.$$

For policies also, the choice of action for MDP policies must only depend on the current state and not on any other state before. This shouldn't be thought must only depend on the current state and not on any other state before. This shouldn't be thought of as a limitation of an MDP policy but more of a condition to be satisfied by the current state.

Value Functions

The received rewards capture how good an action was in that state. Focussing on maximizing this immediate reward may not be ideal and its better to take all future rewards into account.

To measure future rewards, we have 2 measures:-

1. Value Function or State Value Functions
2. Action Value Function

1. State Value Function :- State value function is the expected reward the agent will receive in all future states if it follows the policy π.

It is denoted as:-

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$

2. Action value function :- This is the total expected reward the agent is expected to receive after state S_t if it takes action a_t at S_t and follows

reward the agent is expected to receive after state s_t if it takes action a_t at s_t and follows the policy π after that.

$$q_\pi(s, a) = E\left[G_t \mid S_t = s, A_t = a\right]$$

We can use both these to evaluate how good or bad a policy or action is at a state s.

For example :- we can find the action value functions for various actions at time t and following various policies after that. This way we can select the action and policy with the maximum expected return.

Note however that this is very computationally intensive.