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 immediately 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 probabilty 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.