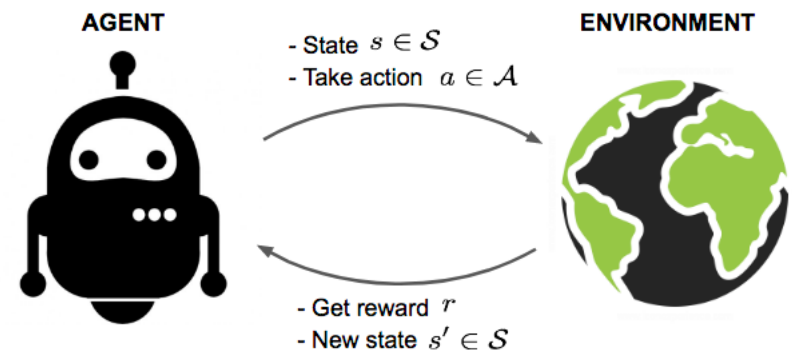
Reinforcement Learning

Cerritos Lira, Carlos

23 de junio del 2020

Markov Decision Process



Setup

$$
\begin{aligned}
R(s,a) = &\ \text{Reward of taking action } a \text{ at state } s \\
T(s,a,s') = &\ P(s' \mid s,a) \text{ Probability of getting to } s' \text{ given that we were in state } s \text{ and took aciton } a \\
\pi(s) = &\ \text{Policy, the action that we should take given that we are in state } s \\
U^\pi(s) = &\ \text{Utility, how good is a state}
\end{aligned}
$$

Bellman equation

Given a policy π, we measure how good a state is by taking the expected sum of ininite discounted rewards.

$$
\begin{aligned}
U^\pi(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi, s_0 = s] \\
U^\pi(s) &= E[R(s) + \gamma \sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi, s_0 = S'], \quad P_{S'}(s') = T(s, \pi(a), s') \\
U^\pi(s) &= R(s) + \gamma \sum_{s'} T(s, \pi(s), s') U(s') \\
\pi^*(s) &= \operatorname*{argmax}_a \sum_{s'} T(s, a, s') U(s') \\
U(s) &= R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U(s')
\end{aligned}
$$

Value interation

The straight forward way of updating U_T is:

$$U_T(s) = R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U_{T-1}(s')$$

define the Bellman operator such that:

$$U_T = BU_{T-1}$$

since the Bellman operator is a contraction and U is a fixed point this method should converge to U.

Q-learning

$$Q(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q(s', a')$$

$$Q_T(s_{t-1}, a_{t-1}) = Q_{T-1}(s_{t-1}, a_{t-1}) + \alpha_T(r_t + \gamma \max_{a'} Q_{T-1}(s_{t-1}, a') + Q_{T-1}(s_{t-1}, a_{t-1}))$$