Markov Decision Processes (MDP)

esgl Hu

November 9, 2017

1 Notation & Definition

A reinforcement learning task taht satisfies the Markov property is called a *Markov decision process*, or *MDP*. If the state and action spaces are finite, then it is called a *finite Markov decision process (finite MDP)*

2 preface

Notations & Definitions

The methods have learned the values of actions and then selected actions based on their estimated action values, their policies would not even exist without the action-value estimates.

The methods have learned a *parameterized policy* that can select actions without consulting a value function. A value function may still be used to *learn* the policy weights, but is not required for action selection.

The notation $\theta \in \mathbb{R}^n$ is used for the primary learned weigth vector, $\pi(a|s,\theta) \doteq \Pr\{A_t = a|S_t = s, \theta_t = \theta\}$ for the probability that action a is taken at time t given that the agent is in state s at time t weight vector θ.

These methods seek to maximize preformance $\eta(\theta)$, so their updates approximate gradient ascent in η:

$$\theta_{t+1} \doteq \theta_t + \alpha \widehat{\nabla(\eta(\theta_t)})) \tag{1}$$

where $\widehat{\nabla(\eta(\theta_t)}))$ is a stochastic estimate whose expectation approximates the gradients of the performance measure $\eta(\theta_t)$ with respect to its argument θ_t

Advantages & Disadvantages

Advantages

- Better convergence properties
- Effective in high-dimensional or continuous action spaces
- Can learn stochastic policies

Disadvantages

- Typically converge to a local rather than global optimum
- Evaluating a policy is typically inefficient and high variance

Definition of Policy Gradient

We let τ denote a state-action sequence $\{s_0, a_0, ..., s_H, a_H\}$. We overload nototation: $R(\tau) = \sum_{t=0}^{H} R(s_t, a_t)$.

$$\eta(\theta) = \mathbb{E}[\sum_{t=0}^{H} R(s_t, a_t), \pi_\theta] = \sum_{\tau} P(\tau;\theta)R(\tau) \tag{2}$$

In our new notation, our goal is to find θ:

$$\max_{\theta} \eta(\theta) = \max_{\theta} \sum_{\tau} P(\tau; \theta) R(\tau) \tag{3}$$