Reinforcement Learning Notes

February 18, 2018

1 Markov Reward process

A MRP(Markove reward process) does not have actions involved, that concept is MDP(markov decision process)

2 return G

The return G_t is total discounted reward for time-step t. return is defined for a given sample

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

The discount $\gamma \in [0, 1]$

3 Bellman equation for MRPs

The main idea is :

The value function can be decomposed into two parts:

- immediate reward R_{t+1}
- discounted value of successor state $\gamma v(S_{t+1})$

$$\begin{aligned} v(s) &= E[G_t | S_t = s] \\ &= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s] \\ &= E[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \dots)) | S_t = s] \\ &= E[R_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= E[R_{t+1} + \gamma v(S_{t+1}) | S_t = s] \end{aligned} \tag{2}$$

4 Formal definition of MDP

A Markov decision process is a markov reward process with decisions. It is an environment in which all states are markov.

A Markov decision Process is a tuple (S, A, P, R, γ)

- S is a finite set of states
- A is a finite set of actions
- P is a state transition probability matrix,
  $P^a_{ss'} = P[S_{t+1} = s'|S_t = s, A_t = a]$
- R is a reward function,
  $R^a_s = E[R_{t+1}|S_t = s, A_t = a]$
- γ is a discount factor

5 Policy and Stochastic policy

A policy defines behavior/decisions/actions of the agent to look for what actions to take. A policy tells what action to take given a state S i.e. $\pi : s \mapsto a$

A stochastic policy π, is a distribution over actions given state,

$$\pi(a|s) = P[A_t = a|S_t = s] \tag{3}$$

6 Value function

The **state-value function** $v_\pi(s)$ of an MDP is the expected return starting from state s, and then following policy π

$$v_\pi(s) = E_\pi[G_t|S_t = s] \tag{4}$$

The expectation above makes sense for a stochastic policy, whereas for a fixed policy value function is just return defined by policy.

The **action-value function** $q_\pi(s, a)$ is the expected return starting from state s, taking action a, and then following policy π

$$q(s, a) = E_\pi[G_t|S_t = s, A_t = a] \tag{5}$$

Relating state-value $v_\pi(s)$ and action-value $q_\pi(s, a)$ (a single step look ahead):

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s, a) \tag{6}$$

Relating action-value $q_\pi(s, a)$ and state-value $v_{pi}(s)$ (a single step look ahead):

$$q_\pi(s, a) = R^a_s + \sum_{s' \in S} \gamma P^a_{ss'} v_\pi(s') \tag{7}$$

Using the above two steps look ahead (first over all actions then over all subsequent states) we can specify state-values in terms of itself:

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \left(R^a_s + \sum_{s \in S} \gamma P^a_{ss'} v_{pi}(s') \right) \tag{8}$$

Similarly using two steps look ahead (first over all states then over all subsequent actions), we can specify action-values recursively in tersm of itself TODO