Reinforcement Learning

March 3, 2016

1 The Problem

Reinforcement learning aims at learn the environemnt in order to achieve a goal. The elements of the problem are:

Agent Entity doing the learning and the decision making

Environment Everything which is outside the agent

Action Interaction chosen vby the agent

State Representation of the environment

Policy Mapping from states to probabilities of selection the next action

The goal is to maximize the total reward in the long run. Each action on a given state has a probability to give certain rewards, and the agent's objective is try to learn the policy that will maximize the succession of rewards:

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

So, at each instant t, the agent will try to maximize G_t. Rewards at each time can depend on all the actions the agent took in the past. However to simplify the problem, rewards have the *markov property*:

$$\mathbf{Pr}\left\{R_{t+1}, S_{t+1} | S_i, A_i \forall i \leq t\right\} = \mathbf{Pr}\left\{R_{t+1}, S_{t+1} | S_t, A_t\right\} \tag{2}$$

If the environment has this property, then tasks can be more efficiently computed and, for a relatively small number of possible states, the task is already solved. It is what is called, *Markov Decision Processes* (MDP)

1.1 Markov Decision Processes

Following the previous definitions, and given that the environment has the morkov property, we can define:

$$p(s'|s,a) = \mathbf{Pr}\left\{S_{t+1}|S_t, A_t\right\} \tag{3}$$

which is the probability of each possible state s'. These are called transition probabilities. We can also define the expected reward given the current action and current and next state:

$$r(s,a,s') = E\left[R_{t+1}|S_t = s, A_t = a, S_{t+1} = s'\right] \tag{4}$$

These two quantities is what we have to remember and define most of the properties of the MDPs.

1.2 Value Functions

Estimated inmediate rewards is not the overal goal for the agent. The main task is to estimate expected rewards in the long run. So, if we know that our agent follows a policy π, mapping actions to probabilities like this: $\pi a|s$, we can define the value of a state s following a policy π:

$$v_\pi(s) = E_\pi\left[G_t|S_t = s\right] = E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s\right] \tag{5}$$

we call this function *state-value function for policy* π. We can also define the *action-value function for policy* π

$$q_\pi(s,a) = E_\pi\left[G_t|S_t = s, A_t = a\right] = E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s, A_t = a\right] \tag{6}$$

with these definitions we can relate the value of a state to the value of other states by means of the *Bellman equation*

$$\begin{align}
v_\pi &= E_\pi\left[G_t|S_t = s\right] \tag{7}\\
&= E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s\right] \tag{8}\\
&= E_\pi\left[R_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2}|S_t = s\right] \tag{9}\\
&= E_\pi\left[R_{t+1}|S_t = s\right] + E_\pi\left[\gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2}|S_t = s\right] \tag{10}
\end{align}$$

Now, taking the average according to a policy is exactly the same as:

$$E_\pi\left[f(x)\right] = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) f(x) \tag{11}$$

And also, that the second term in the previous equation has the form than the original equation. Essentially, this second term is the value of all states at time $t+1$. Therefore, the final form of the equation is:

$$v_\pi = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma v_{pi}(s')\right] \tag{12}$$

which defines a system of linear equations on the state values.