DS 102 Discussion 11

Wednesday, 18 November, 2020

In this discussion, we'll review the concepts of the value function $V(s)$ and Q-function $Q(s,a)$ introduced in Lectures, and practice going through the computations needed to solve them.

First, a brief overview of Markov Decision Process (MDP) terminology:

- $s \in S$: states
- $a \in A$: actions we can take from states
- $\mathbb{P}(s' \mid s, a)$: transition function, capturing the distribution over states we will end up in if we take action a from state s
- $R(s,a,s')$: reward function, which we receive at each iteration when we take action a from state s to end up in state s'.
- $\gamma \in [0,1]$: discount factor for rewards received after the current iteration
- $\pi : S \to A$: policy, describing a strategy of what action to take from a state

The value function $V^{\pi}(s)$ of a policy π gives the expected (discounted) reward received when starting from state s and using strategy π:

$$V^{\pi}(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \left[R(s,a,s') + \gamma V^{\pi}(s') \right].$$

This equation is also known as the **Bellman equation**.

We are often interested in the value function of a particular policy: the one that is optimal from state s. This is the **optimal value function** $V^*(s)$:

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \left[R(s,a,s') + \gamma V^*(s') \right].$$

Similarly, the **optimal Q-function** $Q^*(s,a)$ gives the expected (discounted) reward received when starting from state s, taking action a, then taking the optimal actions thereafter:

$$Q^*(s,a) = \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \left[R(s,a,s') + \gamma V^*(s') \right].$$

A typical goal in reinforcement learning is to find a policy π^* that maximizes our expected discounted reward. Building up to that goal, we first need to understand how to evaluate the optimal value function and optimal Q-function.

1. We have the following grid representation of a problem:

			1
	×	`start`	-100

where `start` represents our initial state, × is a state we can't access, and the 1 and -100 states are terminal states with corresponding rewards. The reward received when moving to any other state is zero.

(a) Assume state transitions are deterministic, meaning that an action in a particular direction always moves us in that direction (unless it's toward the × state, in which case we stay in the same state). Compute the optimal value function at each state, when $\gamma = 0.9$.

Solution:

0.9^2	0.9	1	N/A
0.9^3	×	0.9	N/A
0.9^4	0.9^3	0.9^2	0.9^3

(b) Compute the optimal Q-function at our initial state for the actions of going up, down, left, and right.

Solution: Going up gives us a Q-value of 0.9, going left gives us a Q-value of 0.9^2, going down gives us a Q-value of 0.9^3, and going right gives us a Q-value of -100.

(c) Based on the optimal Q-function you just computed, what would be the optimal move to make from `start`?

Solution: You should go up to maximize your Q-function.

(d) Now suppose the state transitions are stochastic, such that there is a 0.8 probability of going in the direction you specified, and a 0.1 probability of going in either of the directions perpendicular to what specified. For example, if you decide to go up, you go up with 0.8 probability, go left with a 0.1 probability, and go right with a 0.1 probability. What is the best action to perform from `start`?

Solution: You should go left, as you have probability of 0 of landing in the bad final state. In contrast, if you choose any other action, you might land in the bad final state (whose negative reward is of such large magnitude, it always outweighs the discounted reward of reaching the good final state).