

COMP9444 Neural Networks and Deep Learning

Session 2, 2018

Exercises 6: Reinforcement Learning

This page was last updated: 09/19/2018 11:23:39

Consider an environment with two states $S = \{S_1, S_2\}$ and two actions $A = \{a_1, a_2\}$, where the (deterministic) transitions δ and reward R for each state and action are as follows:

$$\delta(S_1, a_1) = S_1, R(S_1, a_1) = +1$$

$$\delta(S_1, a_2) = S_2, R(S_1, a_2) = -2$$

$$\delta(S_2, a_1) = S_1, R(S_2, a_1) = +7$$

$$\delta(S_2, a_2) = S_2, R(S_2, a_2) = +3$$

1. Draw a picture of this environment, using circles for the states and arrows for the transitions.
2. Assuming a discount factor of $\gamma = 0.7$, determine:
 - a. the optimal policy $\pi^* : S \rightarrow A$
 - b. the value function $V^* : S \rightarrow R$
 - c. the "Q" function $Q^* : S \times A \rightarrow R$

Write the Q values in a matrix like this:

Q	a_1	a_2
S_1		
S_2		

Trace through the first few steps of the Q-learning algorithm, with a learning rate of 1 and with all Q values initially set to zero. Explain why it is necessary to force exploration through probabilistic choice of actions, in order to ensure convergence to the true Q values.

3. Now let's consider how the Value function changes as the discount factor γ varies between 0 and 1. There are four deterministic policies for this environment, which can be written as $\pi_{11}, \pi_{12}, \pi_{21}$ and π_{22} , where $\pi_{ij}(S_1) = a_i, \pi_{ij}(S_2) = a_j$
 - a. Calculate the value function $V^{\pi_{(\gamma)}} : S \rightarrow R$ for each of these four policies (keeping γ as a variable)
 - b. Determine for which range of values of γ each of the policies $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ is optimal

Make sure you try answering the Exercises yourself, before checking the [Sample Solutions](#)