6CCS3ML1 (Machine Learning)

Tutorial 07

(Version 1.1)

1. An agent has a series of choices between taking action a_1 and action a_2. Its choices, and the resulting payoffs, are as follows:

time:	1	2	3	4	5	6	7	8
action:	a_1	a_1	a_2	a_2	a_1	a_2	a_2	a_1
payoff:	9	2	7	3	3	8	4	5

   What is the action-value estimate for each of a_1 and a_2 at each point in time?

   Note: action selection here was random, choosing between a_1 and a_2 with equal probability, and payoffs were randomly selected from values between 1 and 10 with equal probability.

2. Consider an agent that is choosing between 3 actions, a_1, a_2 and a_3, with the following average rewards: $Q(a_1) = 5$, $Q(a_2) = 7$, and $Q(a_3) = 4$.

   (a) If the agent uses ϵ-greedy action selection, and $\epsilon = 0.1$, what is the probability that each action will be selected?

   (b) If the agent uses softmax action selection, using the Gibbs distribution with $\tau = 0.1$, what is the probability that each action will be selected?

3. Consider an agent which has established, using value iteration, the utility values shown in Figure 1b. What policy should this agent adopt?

4. Consider an agent using passive reinforcement learning in the environment in Figure 1a.

   Consider the following runs:

$$\begin{aligned}
&(1,1)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (2,3)_{-0.04} \\
&\rightarrow (2,3)_{-0.04} \rightarrow (2,3)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (3,2)_{-0.04} \\
&\rightarrow (3,3)_{-0.04} \rightarrow (3,2)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (4,3)_{1} \\
&(1,1)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (2,3)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (4,3)_{1} \\
&(1,1)_{-0.04} \rightarrow (1,1)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (2,3)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (4,3)_{1}
\end{aligned}$$

   (a) Use direct utility estimation to estimate the utility of each state along the first run, after that run.

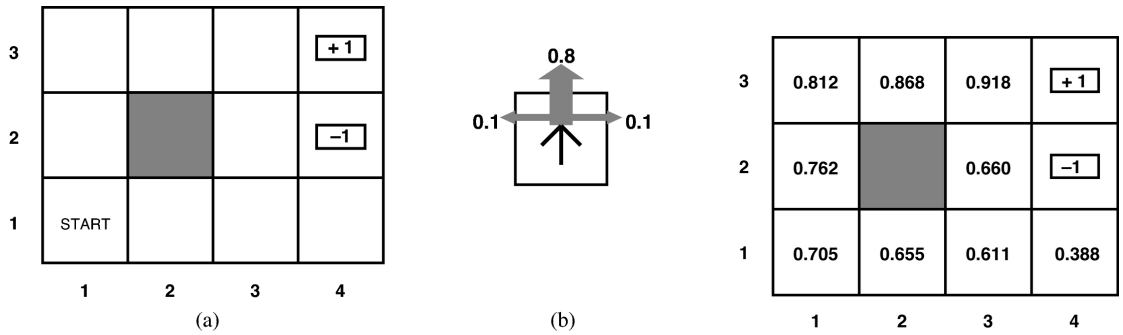
   (b) Calculate the sample estimate of $P(s'|s, \pi(s))$ for each state along the first run.

   (c) Repeat the previous calculations after the second run.
   Note that the values you should compute are the cumulative values after the first and second runs. As a result, you should include utility and probability estimates for every state visited on either run.

   (d) Now update your answer to the previous question after the third run.

   (e) What do you notice about the estimates?

   Note: these runs were randomly generated.



(a) (b)

(a) The scenario (b) Utility values

Figure 1: A familiar scenario

Version list

- Version 1.0, March 1st 2020.
- Version 1.1, February 5th 2021.