University of Illinois Spring 2018

CS 446: Machine Learning
Homework 12

Due on April 24, 2018, 11:59 a.m. Central Time

1. **[13 points]** Q-Learning

   (a) State the Bellman optimality principle as a function of the optimal Q-function $Q^*(s,a)$, the expected reward function $R(s,a,s')$ and the transition probability $P(s'|s,a)$, where s is the current state, s' is the next state and a is the action taken in state s.

   Your answer:

   $$Q^*(s,a) = \sum_{s\prime \in S} P(s\prime|s,a)\left[R(s,a,s\prime) + \max_{a\prime \in \mathcal{A}_{s\prime}} Q^*(s\prime,a\prime)\right]$$

   (b) In case the transition probability $P(s'|s,a)$ and the expected reward $R(s,a,s')$ are unknown, a stochastic approach is used to approximate the optimal Q-function. After observing a transition of the form (s,a,r,s'), write down the update of the Q-function at the observed state-action pair (s,a) as a function of the learning rate α, the discount factor γ, $Q(s,a)$ and $Q(s',a')$.

   Your answer:

   $$Q(s,a) \leftarrow (1-\alpha)\cdot Q(s,a) + \alpha \cdot \left(R(s,a,s\prime) + \gamma \cdot \max_{a\prime \in \mathcal{A}_{s\prime}} Q(s\prime,a\prime)\right)$$

   (c) What is the advantage of an epsilon-greedy strategy?

   Your answer: Since we are assuming the environment is stochastic given a **known** state and action pair, it may be the case that there is another action that yields a better reward from the same state. ϵ-greedy allows the agent to *explore* the environment in hope of finding better actions that the ones currently seen.
   You could see ϵ as a control of the tradeoff between exploitation (always performing the best action) and exploration (potentially resulting in trying new actions).

   (d) What is the advantage of using a replay-memory?

   Your answer: The main advantage is that it allows us to approximate the probability $P(s\prime|s,a)$ which is, more often than not, unknown. Another reason is that by keeping track of the past experiences ($(s,a,s\prime,r)$ tuples) and sampling randomly it removes the correlation in the observed sequence of states, as explained in `https://datascience.stackexchange.com/questions/20535/understanding-experience-replay-in-reinforcement-learning`

(e) Consider a system with two states S_1 and S_2 and two actions a_1 and a_2. You perform actions and observe the rewards and transitions listed below. Each step lists the current state, reward, action and resulting transition as: $S_i; R = r; a_k : S_i \rightarrow S_j$. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step by applying the formula from part (b). The Q-table entries are initialized to zero. Fill in the tables below corresponding to the following four transitions. What is the optimal policy after having observed the four transitions?

i. S_1; $R = -10$; $a_1 : S_1 \rightarrow S_1$
ii. S_1; $R = -10$; $a_2 : S_1 \rightarrow S_2$
iii. S_2; $R = 18.5$; $a_1 : S_2 \rightarrow S_1$
iv. S_1; $R = -10$; $a_2 : S_1 \rightarrow S_2$

Q	S_1	S_2
a_1	-5	0
a_2	0	0

Q	S_1	S_2
a_1	-5	0
a_2	-5	0

Q	S_1	S_2
a_1	-5	8
a_2	-5	0

Q	S_1	S_2
a_1	-5	8
a_2	-5.5	0

Your answer: Given the updates performed, and shown in the tables above, we can use

$$\pi(s) = \arg\max_{a \in \mathcal{A}_s} Q^*(s, a)$$

where $A_s = \{a_1, a_2\}$ for $s \in \{S_1, S_2\}$. So

$$\pi(S_1) = \arg\max_{a \in \mathcal{A}_{S_1}} Q^*(S_1, a) = a_1$$

$$\pi(S_2) = \arg\max_{a \in \mathcal{A}_{S_2}} Q^*(S_2, a) = a_1$$