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:

   (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:

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

   Your answer:

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

   Your answer:

(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: $S_i; R = r; a_k : S_i \to S_j$. 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; $R = -10$; $a_1 : S_1 \to S_1$
ii. S_1; $R = -10$; $a_2 : S_1 \to S_2$
iii. S_2; $R = 18.5$; $a_1 : S_2 \to S_1$
iv. S_1; $R = -10$; $a_2 : S_1 \to S_2$

Q	S_1	S_2
a_1	.	.
a_2	.	.

Q	S_1	S_2
a_1	.	.
a_2	.	.

Q	S_1	S_2
a_1	.	.
a_2	.	.

Q	S_1	S_2
a_1	.	.
a_2	.	.

Your answer: