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:

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

(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:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \left(r + \gamma * \max_{a' \in \mathcal{A}} Q(s',a') \right)$$

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

Your answer:
The advantage of the epsilon-greedy strategy is that it prevents the model from falling into a suboptimal strategy by selecting a random action with some probability ϵ on every step.

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

Your answer:
Replay-memory or experience-replay is a technique that uses a random sample of prior actions instead of the most recent action on the update step. This process removes correlations in the exploration/observation sequence and smooths changes in the data distribution.

(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	.	.
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:

$$S_1; R = -10; a_1 : S_1 \to S_1$$

$$Q(s_1, a_1) \leftarrow \frac{1}{2} * 0 + \frac{1}{2}\left(-10 + \frac{1}{2} * \max\{0, 0\}\right)$$

$$\leftarrow -5$$

Q	S_1	S_2
a_1	-5	0
a_2	0	0

$$S_1; R = -10; a_2 : S_1 \to S_2$$

$$Q(s_1, a_2) \leftarrow \frac{1}{2} * 0 + \frac{1}{2}\left(-10 + \frac{1}{2} * \max\{0, 0\}\right)$$

$$\leftarrow -5$$

Q	S_1	S_2
a_1	-5	0
a_2	-5	0

$$S_2; R = 18.5; a_1 : S_2 \to S_1$$

$$Q(s_2, a_1) \leftarrow \frac{1}{2} * 0 + \frac{1}{2}\left(18.5 + \frac{1}{2} * \max\{-5, -5\}\right)$$

$$\leftarrow 8$$

Q	S_1	S_2
a_1	-5	8
a_2	-5	0

$$S_1; R = -10; a_2 : S_1 \to S_2$$

$$Q(s_1, a_2) \leftarrow \frac{1}{2} * -5 + \frac{1}{2}\left(-10 + \frac{1}{2} * \max\{8, 0\}\right)$$

$$\leftarrow -5.5$$

Q	S_1	S_2
a_1	-5	8
a_2	-5.5	0

Optimal Policy

$S_1 : a_1; \quad S_2 : a_1$