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:

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

   (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:

   - Obtain a sample mini-batch $\mathcal{B} \subseteq \mathcal{D}$, which $\mathcal{D} = \{(s,a,r,s')\}$
   - Compute target

     $$y_j = R(s,a,s') + \gamma \max_{a' \in \mathcal{A}'_s} Q(s',a') \qquad \forall j \in \mathcal{B}$$

   - Use stochastic (semi-)gradient descent to optimize:

     $$\begin{aligned} Q(s,a) &= (1-\alpha)Q(s,a) + \alpha y_j \\ &= (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a' \in \mathcal{A}'_s} Q(s',a')) \end{aligned}$$

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

   Your answer:
   Advantage of Epsilon greedy strategy is to explore (generate) more random samples rather than directly use the state space. (so it is more stochastic)

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

Your answer:

- Q-learning updates are incremental and do not converge quickly, so multiple passes with the same data is beneficial, especially when there is low variance in immediate outcomes (r, s') given the same state, action pair (s, a).
- Better convergence behavior will be given when training function approximation. This is because the data is more like i.i.d. data assumed in most supervised learning convergence proofs.

(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:
According to the solution in (b), the equation will become to:

$$\begin{aligned} Q(s,a) &= (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a' \in \mathcal{A}'_s} Q(s',a')) \\ &= 0.5 \times Q(s,a) + 0.5 \times (R(s,a,s') + 0.5 \times \max_{a' \in \mathcal{A}'_s} Q(s',a')) \end{aligned}$$

- Step 1. $Q(S_1, a_1) = 0.5 \times 0 + 0.5 \times (-10 + 0.5 \times 0) = -5$
- Step 2. $Q(S_1, a_2) = 0.5 \times 0 + 0.5 \times (-10 + 0.5 \times 0) = -5$
- Step 3. $Q(S_2, a_1) = 0.5 \times 0 + 0.5 \times (18.5 + 0.5 \times -5) = 8$
- Step 4. $Q(S_1, a_2) = 0.5 \times -5 + 0.5 \times (-10 + 0.5 \times 8) = -5.5$

Q	S_1	S_2
a_1	-5	.
a_2	.	.

Q	S_1	S_2
a_1	-5	.
a_2	-5	.

Q	S_1	S_2
a_1	-5	8
a_2	-5	.

Q	S_1	S_2
a_1	-5	8
a_2	-5.5	.