**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 $s$.

   **Solution:**
   (1 point)

   Bellman Equation:
   $Q^*(s, a) = \sum_{s' \in S} P(s'|s, a)[R(s, a, s') + \max_{a' \in A_{s'}} Q^*(s', a')]$

   (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 $\alpha$, the discount factor $\gamma$, $Q(s, a)$ and $Q(s', a')$.

   **Solution:**
   (2 points)

   $Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma \max_{a' \in A_{s'}} Q(s', a'))$

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

   **Solution:**
   (1 point)

   Trade-off between exploration and exploitation. The best long-term strategy may involve short-term sacrifices resulting in not taking the best action at the beginning of the train to better explore the environment.

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

   **Solution:**
   (1 point)

   Learning from batches of consecutive samples is problematic, as the samples are correlated. This can lead to inefficient learning. For example, if maximizing action is to move left, training samples will be dominated by samples from left-hand size. Instead, a replay memory is used to store the transitions $(s_t, a_t, r_t, s_{t+1})$ as game episodes are played. The Q-network is then trained on random minibatches of transitions from the replay memory, instead of consecutive samples.

   (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. $S_1$; $R = -10$; $a_2 : S_1 \rightarrow S_2$

| $Q$ | $S_1$ | $S_2$ |
|---|---|---|
| $a_1$ | . | . |
| $a_2$ | . | . |

| $Q$ | $S_1$ | $S_2$ |
|---|---|---|
| $a_1$ | . | . |
| $a_2$ | . | . |

| $Q$ | $S_1$ | $S_2$ |
|---|---|---|
| $a_1$ | . | . |
| $a_2$ | . | . |

| $Q$ | $S_1$ | $S_2$ |
|---|---|---|
| $a_1$ | . | . |
| $a_2$ | . | . |

**Solution:**
(8 points/ 2 for each)

i. $Q(a_1, S_1) = -5$, $Q(a_2, S_1) = 0$, $Q(a_1, S_2) = 0$, $Q(a_2, S_2) = 0$

ii. $Q(a_1, S_1) = -5$, $Q(a_2, S_1) = -5$, $Q(a_1, S_2) = 0$, $Q(a_2, S_2) = 0$

iii. $Q(a_1, S_1) = -5$, $Q(a_2, S_1) = -5$, $Q(a_1, S_2) = 8$, $Q(a_2, S_2) = 0$

iv. $Q(a_1, S_1) = -5$, $Q(a_2, S_1) = -5.5$, $Q(a_1, S_2) = 8$, $Q(a_2, S_2) = 0$

The optimal policy at this point is: $\pi(S_1) = a_1$ and $\pi(S_2) = a_1$.