CS234 Problem Session Solutions

Week 3: Jan 27

1) [Breakout Rooms] Q-learning Practice

Consider an unknown MDP with three states (A, B, C) and two actions $(\leftarrow, \rightarrow)$. Suppose the agent chooses actions according to some policy π in the unknown MDP, collecting a dataset consisting of samples (s, a, s', r) representing taking action a in state s resulting in a transition to state s' and a reward of r.

s	a	s'	r
A	$\rightarrow$	B	2
C	$\leftarrow$	B	2
B	$\rightarrow$	C	-2
A	$\rightarrow$	B	4

You may assume a discount factor of $\gamma = 1$.

Recall the update function of Q-learning is:

$$Q(s_t, a_t) = (1 - \alpha) \cdot Q(s_t, a_t) + \alpha \cdot (r_t + \gamma max_{a'} Q(s_{t+1}, a')) \tag{1}$$

Assume that all Q-values are initialized to 0, and use a learning rate of $\alpha = \frac{1}{2}$.

(a) Run Q-learning on the above experience table and fill in the following Q-values:

$Q(A, \rightarrow) =?$

$Q(B, \rightarrow) =?$

Solution This question was borrowed from UC Berkeley's CS188. [1]

$Q_1(A, \rightarrow) = \frac{1}{2} \cdot Q_0(A, \rightarrow) + \frac{1}{2}(2 + \gamma max_{a'} Q(B, a')) = 1$

$Q_1(C, \leftarrow) = 1$

$Q_1(B, \rightarrow) = \frac{1}{2}(-2 + 1) = -\frac{1}{2}$

$Q_2(A, \rightarrow) = \frac{1}{2} \cdot 1 + \frac{1}{2}(4 + max_{a'} Q_1(B, a'))$

$= \frac{1}{2} + \frac{1}{2}(4 + 0) = \frac{5}{2}$.

(b) After running Q-learning and producing the above Q-values, you construct a policy π_Q that maximizes the Q-value in a given state: $\pi_Q(s) = argmax_a Q(s, a)$.

What are the actions chosen by the policy in states A and B?

[1] https://inst.eecs.berkeley.edu/ cs188/fa20/assets/section/midterm_review_rl_solutions.pdf

Solution $\pi_Q(A) = \rightarrow$
$\pi_Q(B) = \leftarrow$
Note that $Q(B, \leftarrow) = 0 > -\frac{1}{2} = Q(B, \rightarrow)$.

(c) Compute the MLE MDP model estimates of the transition function $\hat{P}(s, a, s')$ and reward function $\hat{R}(s, a, s')$.

Write down the following quantities. You may write N/A for undefined quantities.

$\hat{P}(A, \rightarrow, B) = ?$

$\hat{P}(B, \rightarrow, A) = ?$

$\hat{P}(B, \leftarrow, A) = ?$

$\hat{R}(A, \rightarrow, B) = ?$

$\hat{R}(B, \rightarrow, A) = ?$

$\hat{R}(B, \leftarrow, A) = ?$

Solution $\hat{P}(A, \rightarrow, B) = 1$
$\hat{P}(B, \rightarrow, A) = 0$
$\hat{P}(B, \leftarrow, A) = N/A$
$\hat{R}(A, \rightarrow, B) = 3$
$\hat{R}(B, \rightarrow, A) = N/A$
$\hat{R}(B, \leftarrow, A) = N/A$

2) [Breakout Rooms] Value Functions

Prove that the following two definitions of the state-value function are equivalent:

$$V^\pi(s) = \mathbf{E}[G_t|S_t = s, \pi] \tag{2}$$
$$V^\pi(s) = \mathbf{E}[G|S_0 = s, \pi] \tag{3}$$

Solution Let us denote the first definition as V_t^π and the second as $V_0^\pi(s)$.
$V_t^\pi = \mathbf{E}[G_t|S_t = s, \pi]$
$= \sum_{k=0}^{\infty} \gamma^k \mathbf{E}[R_{t+k}|S_t = s, \pi]$
$= \sum_{a \in A} \pi(s,a)(R(s,a) + \sum_{k=1}^{\infty} \gamma^k \mathbf{E}[R_{t+k}|S_t = s, \pi]$
$= \sum_{a \in A} \pi(s,a)[R(s,a) + \sum_{s' \in S} P(s,a,s') \sum_{a' \in A} \pi(s',a')(\gamma^1 R(s',a') + \sum_{k=2}^{\infty} \gamma^k \mathbf{E}[R_{t+k}|S_t = s, \pi])]$
$= \gamma^0 \sum_{a \in A} \pi(s,a)R(s,a) + \gamma^1 \sum_{a \in A} \pi(s,a) \sum_{s' \in S} P(s,a,s') \sum_{a' \in A} \pi(s',a')R(s',a') +$
$\gamma^2 \sum_{a \in A} \pi(s,a) \sum_{s' \in S} P(s,a,s') \sum_{a' \in A} \pi(s',a') \sum_{s'' \in S} P(s',a',s'') \sum_{a'' \in A} \pi(s'',a'')R(s'',a'') +$
...
$= \gamma^0 \sum_{a \in A} Pr(A_0 = a|S_0 = s)R(s,a) +$
$\gamma^1 \sum_{a \in A} Pr(A_0 = a|S_0 = s) \sum_{s' \in S} Pr(S_1 = s'|A_0 = a, S_0 = s) \sum_{a' \in A} Pr(A_1 = a'|S_1 = s')R(s',a') + ...$
$= \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R_t|S_0 = s, \pi]$
$= \mathbf{E}[G|S_0 = s, \pi] = V_0^\pi(s)$
. .

3) [Breakout Rooms] Negative Reward MDP

Consider a finite MDP with bounded rewards, where all rewards are negative. That is, $R_t < 0$ always. Let $\gamma = 1$. The MDP is finite horizon, with horizon L, and also has a deterministic transition function and initial state distribution (rewards may be stochastic). Let $H_\infty = (S_0, A_0, R_0, S_1, A_1, R_1, ...S_{L-1}, A_{L-1}, R_{L-1})$ be any history that can be generated by a deterministic policy *pi*. Prove that the sequence $V^\pi(S_0), V^\pi(S_1), ...V^\pi(S_{L-1})$ is strictly increasing.

Solution

$$\begin{aligned}
V^\pi(S_t) &= && (4)\\
&= V^\pi(s_t) && (5)\\
&= \mathbf{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s_t, \pi] && (6)\\
&= \sum_{k=0}^{\infty} \mathbf{E}[R_{t+k} | S_t = s_t, \pi] && (7)\\
&= \sum_{k=0}^{\infty} \mathbf{E}[R_{t+k} | \pi] && (8)\\
&= \mathbf{E}[R_t|\pi] + \sum_{k=0}^{\infty} \mathbf{E}[R_{t+k+1} | \pi] && (9)\\
&= \mathbf{E}[R_t|\pi] + \sum_{k=0}^{\infty} \mathbf{E}[R_{t+k+1} | S_{t+1} = s_{t+1}, \pi] && (10)\\
&= \mathbf{E}[R_t|\pi] + V^\pi(S_{t+1}) && (11)\\
&\leq V^\pi(S_{t+1}). && (12)
\end{aligned}$$

Notice that the sequence of states are deterministic, so conditioning on $S_t = s_t$ or $S_{t+1} = s_{t+1}$ is conditioning on an event that occurs with probability 1. The final inequality holds since we are given that $R_t < 0$.

Questions 2 and 3 are borrowed from Phil Thomas. [2]

[2] https://people.cs.umass.edu/ pthomas/courses/CMPSCI_687_Fall2018/687_F18_main.pdf