CS 221 Homework 4 – Blackjack

Name: Vinod Kumar Senthil Kumar

SUNET ID: vinodkum

Contributors: Xun Fang, Joe Fan

1a) states = {−2, −1, 0, +1, +2}
start state = {0}
end state = {−2, +2}
action = {−1, +1}
$T(s, +1, s+1) = 30\%$
$T(s, +1, s-1) = 70\%$
$T(s, -1, s-1) = 80\%$
$T(s, -1, s+1) = 20\%$

Initially, let V_{opt} be

s	−2	−1	0	+1	+2
$V_{opt}^0(s)$	0	0	0	0	0

Iteration 1 :-

$V_{opt}^1(-2) = V_{opt}^{+2}\ V_{opt}^1(+2) = 0$, as ±2 are goal states

$$V_{opt}^1(s) = \max_{a \in actions(s)} \sum_{s'} T(s,a,s')\left[Reward(s,a,s') + \gamma V_{opt}^0(s')\right]$$

$$V_{opt}^1(-1) = \max\Big((0.8)(20+0) + (0.2)(-5+0),\ (0.7)(20+0) + (0.3)(-5+0)\Big)$$

$$= \max(15, 12.5) = 15$$

$$V_{opt}^{1}(0) = \max\big((0.8)(-5+0) + (0.2)(-5+0),\ (0.7)(-5+0) + (0.3)(-5+0)\big)$$

$$= \max(-5, -5) = -5$$

$$V_{opt}^{1}(+1) = \max\big((0.8)(-5+0) + (0.2)(100+0),\ (0.7)(-5+0) + (0.3)(100+0)\big)$$

$$= \max(16, 26.5) = 26.5$$

At the end of iteration 1,

s	-2	-1	0	$+1$	$+2$
$V_{opt}(s)$	0	15	-5	26.5	0
$\pi_i(s)$	NA	-1	~~-1~~ $+1$	$+1$	NA

Iteration 2:-

$$V_{opt}^{2}(s) = \max_{a \in \text{actions}(s)} \sum_{s'} T(s,a,s')\left[\text{Reward}(s,a,s') + \gamma V_{opt}^{1}(s')\right]$$

$$V_{opt}^{2}(-2) = V_{opt}^{2}(+2) = 0$$

$$V_{opt}^{2}(-1) = \max\big((0.8)(20+0) + (0.2)(-5-5),\ (0.7)(20+0) + (0.3)(-5-5)\big)$$

$$= \max(14, 11) = 14$$

$$V_{opt}^{2}(0) = \max\Big((0.8)(-5+15) + (0.2)(-5+26.5),\ (0.7)(-5+15) + (0.3)(-5+26.5)\Big)$$

$$= \max(12.3, 13.45)$$

$$= 13.45$$

$$V_{opt}^{2}(1) = \max\Big((0.8)(-5-5) + (0.2)(100+0),\ (0.7)(-5-5) + (0.3)(100+0)\Big)$$

$$= \max(12, 23)$$

$$= 23$$

At the end of iteration 2,

s	-2	-1	0	$+1$	$+2$
$V_{opt}(s)$	0	14	13.45	23	0
$P_i(s)$	NA	-1	$+1$	$+1$	NA

4b)

The differences in the policy between value iteration and Q Learning out of 2745 states for the smallMDP & largeMDP are:

MDP	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Average
smallMDP	2/27	2/27	2/27	1/27	1/27	1/27	1.5/27
largeMDP	863/2745	867/2745	852/2745	882/2745	855/2745	858/2745	862.83/2745

On error analysis, I found that a lack of features that better capture the blackjack game creates a high difference between the two algorithms. Some of the examples are:

Mismatch state	Comments
Not a match at state: (31, 2, (1, 2, 3, 1, 2)) QLearnPolicy: Quit VIPolicy: Take	A better performance can be achieved if we have a feature that give a sense of what value the card index 2 stands for
Not a match at state: (34, 0, (0, 2, 1, 2, 2)) QLearnPolicy: Quit VIPolicy: Take Not a match at state: (34, 0, (3, 3, 1, 0, 3)) QLearnPolicy: Quit VIPolicy: Tak	It might make sense to 'Quit' in the first example but it is a better judgment to 'Take' in the second example as the probability to bust is low. This can be rectified by a feature indicating the presence/absence of each card type
Not a match at state: (26, None, (1, 3, 3, 0, 3)) QLearnPolicy: Quit VIPolicy: Take	There is absolutely no reason why one must 'Quit' at his state. This can be better dealt with if we have a feature that represents the total value at hand.

4c)

The differences in the policy between value iteration and Q Learning out of 2745 states for the largeMDP for different feature extractors are:

Feature Extractor	Trial 1	Trial 2	Trial 3	Trial 4	Averge
identityFeatureExtractor	863/2745	867/2745	852/2745	882/2745	866/2745
blackjackFeatureExtractor	594/2745	581/2745	599/2745	593/2745	592/2745

4d)

The rewards obtained by Value Iteration & Q Learning for 30000 trials are:

Algorithm	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Average
Value Iteration	6.84623	6.82976	6.82456	6.83166	6.83356	6.83315
Q Learning	9.54553	9.60720	9.55170	9.60343	9.58733	9.57904

Value learning fails to adapt when the conditions change as it learns only for the previous threshold and not the new one. But Q Learning, being an online algorithm, easily adapts the new conditions as and when they come. This can be observed in the average reward earned by both the algorithms in the table above where Q Learning earns roughly 40% more reward than Value Iteration.