CS 6300 HW06: Q Learning with Functional Approximation Due March 13, 2023

Please use the LaTeX template to produce your writeups. See the Homework Assignments page on the class website for details. Hand in through gradescope.

1 Functional Approximation

We revisit the simplied version of blackjack from Homework 5. The deck is infinite and the dealer always has a fixed count of 15. The deck contains cards 2 through 10, J, Q, K, and A, each of which is equally likely to appear when a card is drawn. Each number card is worth the number of points shown on it, the cards J, Q, and K are worth 10 points, and A is worth 11. At each turn, you have two possible actions: either *hit* or *stay*.

Unhappy with your experience with basic Q-learning, you decide to featurize your Q-values. Consider the two feature functions:

$$f_1(s,a) = \begin{cases} 0 & a = stay \\ +1 & a = hit, s \geq 15 \\ -1 & a = hit, s < 15 \end{cases} \quad \text{and} \quad f_2(s,a) = \begin{cases} 0 & a = stay \\ +1 & a = hit, s \geq 18 \\ -1 & a = hit, s < 18 \end{cases}$$

Which of the following partial policy tables may be represented by the featurized Q-values unambiguously (without ties)? Derive your answers for each policy table.

s	$\pi(s)$
14	hit
15	hit
16	hit
17	hit
18	hit
19	hit

(a)

s	$\pi(s)$
14	stay
15	hit
16	hit
17	hit
18	stay
19	stay

(b)

s	$\pi(s)$
14	hit
15	hit
16	hit
17	hit
18	stay
19	stay

(c)

s	$\pi(s)$
14	hit
15	hit
16	hit
17	hit
18	hit
19	stay

(d)

s	$\pi(s)$
14	hit
15	hit
16	hit
17	stay
18	hit
19	stay

(e)

Answer
Given the features above, we have the following equation for this functional approximation, $Q(s,a) = w_1 f_1 + w_2 f_2$.
Since f_1 and f_2 are 0 when $a = stay$, we know $Q(s, stay) = 0$.
There are now 3 other possible states when $a = hit$:

$$Q(s, hit) = \begin{cases} -w_1 - w_2 & s < 15 \\ w_1 - w_2 & 15 \leq s < 18 \\ w_1 + w_2 & 18 \leq s \end{cases}$$

Given this piecewise equation, we can deduce what the possible policies might be. We don't know what w_1 and w_2 are, but we can investigate their relationship to find which patterns are possible.

The equation that I've derived above tells us that the policy for $s < 15$ (14) will all be the same, $15 \leq s < 18$ (15, 16, 17) will all be the same, and $18 \leq s$ (18, 19) will all be the same. This means that policy d and policy e are not possible since the action for 18 and 19 are different.

Now we need to investigate policies a, b, and c. Examining the piecewise equation, we can see that the Q values for $s < 15$ should be opposite to $18 \leq s$. That's because $-w_1 - w_2 = -1(w_1 + w_2)$, meaning the Q values will have opposite signs and the policy action will be different. That rules out policies a and b, since in those policies, the action for 14 is the same as the action for 18 and 19.

This leaves c as the only possible policy. Here's an example of weights that would make policy c possible: $w_1 = 1, w_2 = 2$. In this case:

$$Q(s, hit) = \begin{cases} -3 & s < 15 \\ -1 & 15 \leq s < 18 \\ 3 & 18 \leq s \end{cases}$$

This works if a negative Q value maps to hit and a positive to stay. If not, we can just flip the signs on the both weights to make it work.

Thus policy c is the only possible partial policy table that may be represented by the featurized Q-values.