Reinforcement Learning - Theoretical Part

January 15, 2023

Question 1

Consider the initial values as follow:

$$v_0^{(0)} = 0, \quad v_i^{(0)} = \frac{i}{6} \forall i \in \{1,2,3,4\}, v_5^{(0)} = \frac{4}{6}, v_6^{(0)} = 0$$

For the second iteration, all terminal states will continue to yield a value of 0:

$$v_0^{(0)} = v_6^{(0)} = 0$$

Following bellman-ford with equal probabilities, and Assuming a deterministic transition model, we get that:

$$v_s^{(n+1)} = \sum_a \pi(a \mid s) \left[r + \gamma v_{s'}^{(n)} \right]$$

Where $\pi(a, s)$ is the probability of doing action a given state s according to the policy, γ is the discount factor, s' is the resulting state for action a and r is it's reward. In our case for each state, we can only go right or left with equal probability. Namely:

$$v_s^{(n+1)} = 0.5 \cdot \left[r_{s+1} + \gamma v_{s+1}^{(n)} \right] + 0.5 \left[r_{s-1} + \gamma v_{s-1}^{(n)} \right] = \frac{r_{s-1} + r_{s+1} + \gamma \left(v_{s-1}^{(n)} + v_{s+1}^{(n)} \right)}{2}$$

Using the above equation, and assuming $\gamma = 1$, we proceed to calculate the values of the next iteration:

$$v_1^{(1)} = \frac{r_0 + r_2 + v_0^{(0)} + v_2^{(0)}}{2} = \frac{0 + 0 + 0 + \frac{2}{6}}{2} = \frac{1}{6}$$

$$v_2^{(1)} = \frac{r_1 + r_3 + v_1^{(0)} + v_3^{(0)}}{2} = \frac{0 + 0 + \frac{1}{6} + \frac{3}{6}}{2} = \frac{2}{6}$$

$$v_3^{(1)} = \frac{r_2 + r_4 + v_2^{(0)} + v_4^{(0)}}{2} = \frac{0 + 0 + \frac{2}{6} + \frac{4}{6}}{2} = \frac{3}{6}$$

$$v_4^{(1)} = \frac{r_3 + r_5 + v_3^{(0)} + v_5^{(0)}}{2} = \frac{0 + 0 + \frac{3}{6} + \frac{4}{6}}{2} = \frac{7}{12}$$

$$v_5^{(1)} = \frac{r_4 + r_6 + v_4^{(0)} + v_6^{(0)}}{2} = \frac{0 + 1 + \frac{4}{6} + 0}{2} = \frac{5}{6}$$

Question 2

For a general state s_i we calculate its next iteration's value by using the following calculation:

$$v^{t+1}(s_i) = \sum_a \pi(a \mid s_i) \sum_{s',r} p(s', r \mid s_i, a) \left[r + \gamma v_{s'}^t\right]$$

Where $\pi(a, s_s)$ is the probability of doing action a given state s_i according to the policy, s' is the next state reached by following action a, γ is the discount factor, r is the reward and $v_{s'}^t$ is the value of s' in the previous iteration. In our case:

- According to π- for each state, we go right, left up and down with equal probability.
- $\gamma = 1$
- $r = -1$ for every movement.
- The env is deterministic, namely $p(s', r \mid s_i, a) = 1$

Following the explanation above, we proceed to calculate:

$$v^{t+1}(s_2) = \frac{2(-20-1) + 2(-14-1)}{4} = \frac{-21-15}{2} = -18$$

$$v^{t+1}(s_1) = \frac{(-22-1) + (-20-1) + (-14-1) + (v^t(s_1) - 1)}{4} =$$

$$\frac{-60 + v^t(s_1)}{4} = \frac{v^t(s_1)}{4} - 15$$